International Workshop on

Using Linguistic Information for Hybrid Machine Translation LIHMT

Shared Task on

Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation

Barcelona
November 2011

Published by

Universitat Politècnica de Catalunya
Jordi Girona, 31
08034 Barcelona
Spain

ISBN: 978-84-615-2995-7

Printed by

SERVIS GRÀFICS COPISTERIA IMATGE, S.L.
Cinca, 8
08030 Barcelona
Spain

Additional copies of paper proceedings and CDs can be purchased from:

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
Universitat Politècnica de Catalunya
Campus Nord, Edif. Omega 140
Jordi Girona 1-3
08034 Barcelona (Spain)

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

TALP

ixa

OPEN MT 2

eman ta zabal zazu

universidad
del país vasco

euskal herriko
unibertsitatea

LIHMT 2011

International Workshop on **Using Linguistic Information for Hybrid Machine Translation**

18th November 2011

Universitat Politècnica de Catalunya
BARCELONA

Sponsors

EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

EUROPEAN
ASSOCIATION
FOR MACHINE
TRANSLATION

Message from the Programme Committee Chairs

We are delighted to welcome you to the International Workshop on Using Linguistic Information for Hybrid Machine Translation in Barcelona.

To begin with, a few statistics: we received 15 submissions, of which 10 were accepted, providing an overall acceptance rate of 67%. Submissions came from 9 different countries, including 2 beyond Europe's borders. The countries providing most papers were France, Germany and Spain with 3 each.

As programme chairs, we are of course indebted to the panel of reviewers, whose names are listed elsewhere. The 22 reviewers looked at 1 or 2 papers each, thus ensuring that each submission had three separate reviews. We asked reviewers to work to a tight schedule, and with few exceptions they got their reviews in on time, which in turn meant that we could notify authors of acceptance in a timely fashion. We hope that authors have appreciated and benefited from the reviewers' comments, which were often quite extensive. Equally we are grateful to authors who were asked to prepare their final copy for these Proceedings within a fairly short deadline. Again, the deadline was generally met, and we were able to avoid the usual panic and scramble associated with getting the document off to the printers on time.

As Programme chairs, our job sort of ends once we have chosen which papers to accept, and arranged them into the programme you will experience and, we hope, enjoy. At this point we hand matters over to the Local Organisation Committee, but of course we have been working closely together with them since day one, a task obviously much facilitated by the fact that one of us is a "local". Nevertheless, the local organisers have been with us every step of the way, and we would like here to thank them for their support, advice and, when necessary, gentle prodding.

Finally, thanks to all authors, presenters and attendees for making this a successful workshop.

Gorka Labaka and David Farwell
LIHMT 2011 Programme Committee co-chairs

Message from the Local Organising Committee

It gives us great pleasure to welcome you to the LIHMT 2011, the International Workshop on Using Linguistic Information for Hybrid Machine Translation, here on the Campus Nord of the Universitat Politècnica de Catalunya in Barcelona, Spain. We have tried to make all the necessary arrangements to ensure that your participation in the workshop events is as productive and enjoyable as possible. While in Barcelona be sure to experience the special atmosphere the city has to offer: the Roman, medieval, and modernist architecture of the old city, Passeig de Gràcia and the Eixample and the wide array of excellent restaurants, theatre, music, galleries, shops and museums. It would be unfortunate not to take in all you can while here.

Of course, we also hope that you will benefit from a strong programme of presentations that are at the forefront of Machine Translation research and development.

In organising the workshop we have received significant financial support from the Universitat Politècnica de Catalunya, the Generalitat de Catalunya and the Spanish Ministry for Science and Innovation. We also would like to thank the Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP), the European Association for Machine Translation (EAMT) and European Language Resources Association (ELRA) for their generous sponsorship. Finally, we have also had the unselfish assistance of local staff and students. In particular we wish to thank Gorka Labaka and Meritxell Gonzàlez for their many hours of effort, especially in maintaining the workshop web site.

So without further ado, welcome and enjoy the conference.

Local Organising Committee:
David Farwell
Lluís Màrquez
Meritxell Gonzàlez
Cristina España-Bonet
Daniele Piguin
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
Universitat Politècnica de Catalunya

OpenMT-2 Workshop on Using Linguistic Information for Hybrid Machine Translation

Scientific Committee

Co-Chair: David Farwell (Technical University of Catalonia, TALP, Barcelona)
Co-Chair: Gorka Labaka (University of the Basque Country, Donostia)

Iñaki Alegría (University of the Basque Country, Donostia)
Ondřej Bojar (Charles University, Czech Republic)
Josep M. Crego (LIMSI/CNRS, France)
Arantza Díaz de Ilarraza (University of the Basque Country, Donostia)
Chris Dyer (Carnegie Mellon University, US)
Cristina España-Bonet (Technical University of Catalonia, TALP, Barcelona)
Marcello Federico (Fondazione Bruno Kessler, Italy)
Mikel Forcada (University of Alacant, Alicante)
Adrià de Gispert (University of Cambridge, UK)
Meritxell Gonzàlez (Technical University of Catalonia, TALP, Barcelona)
Kevin Knight (Information Sciences Institute, US)
Philipp Koehn (University of Edinburgh, UK)
Patrik Lambert (Universiteé du Maine, France)
José Mariño (Technical University of Catalonia, TALP, Barcelona)
Lluís Màrquez (Technical University of Catalonia, TALP, Barcelona)
Daniele Pighin (Technical University of Catalonia, TALP, Barcelona)
Aarne Ranta (Chalmers University of Technology, Gothenburg, Sweden)
Marta R. Costa-jussà (Barcelona Media, Spain)
Felipe Sánchez-Martínez (University of Alacant, Alicante)
Kepa Sarasola (University of the Basque Country, Donostia)

About the OpenMT-2 Project

(http://ixa.si.ehu.es/openmt2)

The main goal of the OpenMT-2 project is the development of Open Source Machine Translation Architectures based on hybrid models and advanced semantic processors. These architectures will be open-source systems combining the three main Machine Translation frameworks – Rule-Based MT (RBMT), Statistical MT (SMT) and Example-Based MT (EBMT) – into hybrid systems. Implemented architectures and systems will be Open Source, so it will allow rapid system adaptation or development of new advanced Machine Translations systems for other languages. We will test system functionality for different languages: English, Spanish, Catalan and Basque; thus evaluating such architectures in different contexts. While there are many corpus resources for English and Spanish, there are not so many for Catalan and Basque. While the structure of some of those languages is very similar (Catalan and Spanish), others are very different (English and Basque). Basque is an agglutinative and highly inflecting language, unlike English, Catalan and Spanish.

In parallel there has been extensive work on developing an automatic Evaluation platform that supports the introduction of linguistically motivated morphological, syntactic and semantic metrics into the design of MT Evaluation methodologies. It also supports the development and testing of concrete, linguistically-based evaluation techniques.

The main innovative points of the OpenMT-2 project are:

- The design of hybrid systems combining traditional linguistic rules, example-based methods and statistical methods.
- The development of MT evaluation methods based on linguistically motivated metrics.
- The implementation of Open Source Systems.
- The use of advanced syntactic and semantic processing in MT.

Tentative Programme

Friday, November 18, 2011.

8:00: Registration

9:00: Opening

9:10: *Improved Statistical Machine Translation Using MultiWord Expressions*, Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum
9:35: *Using Apertium linguistic data for tokenization to improve Moses SMT performance*, Santiago Cortés Vaíllo and Sergio Ortiz Rojas

10:00: Plenary Session – Alon Lavie: *Statistical MT with Syntax and Morphology: Challenges and Some Solutions*

10:45: Coffee

11:15: *A Radically Simple, Effective Annotation and Alignment Methodology for Semantic Frame Based SMT and MT Evaluation*, Chi-Kiu Lo and Dekai Wu
11:40: *Reordering by Parsing*, Jakob Elming and Martin Haulrich
12:25: *Comparing CBMT Approaches Using Restricted Resources*, Monica Gavrila and Natalia Elita

12:45: Plenary Session – Ondřej Bojar*: Rich Morphology and What Can We Expect from Hybrid Approaches to MT*

13:30: Lunch

15:00: *A New Hybrid Machine Translation Approach Using Cross-Language Information Retrieval and Only Target Text Corpora*, Nasredine Semmar and Dhouha Bouamor
15:25: *Word Translation Disambiguation without Parallel Texts*, Erwin Marsi, André Lynum, Lars Bungum and Björn Gambäck

15:50: Plenary Session – Lucia Specia: *Linguistic Indicators for Quality Estimation of Machine Translation*

16:35: Coffee

17:05: *Deep evaluation of hybrid architectures: simple metrics correlated with human judgements*, Gorka Labaka, Arantza Diaz De Ilarraza, Cristina España-Bonet, Lluís Màrquez and Kepa Sarasola
17:30: *VERTa: Exploring a Multidimensional Linguistically-Motivated Metric*, Elisabet Comelles, Jordi Atserias, Victoria Arranz, Irene Castellon and Olivier Hamon

17:55: Round Table – Recapping the Central Issues

18:20: Closing

Table of Contents

Rich Morphology and What Can We Expect from Hybrid Approaches to MT

Ondřej Bojar
Charles University
bojar@ufal.mff.cuni.cz

The talk will consist of two parts: a summary of problems caused by rich morphology and a speculation as to which of these problems can be mitigated by hybrid approaches to MT.

In the first part, I will give an overview of the most important steps in MT pipeline (training, tuning, evaluation) and the counterplaying effects of rich source and/or target side morphology on achievable MT quality. In the second part, I will relate the problems to some hybrid MT techniques (ROVER system combination, two-step translation and grammatical post-processing) including their inherent limitations in solving the problems.

Statistical MT with Syntax and Morphology: Challenges and Some Solutions

Alon Lavie
Carnegie Mellon University
alavie@cs.cmu.edu

Phrase-based Statistical Machine Translation is the most dominant approach to MT in recent years. Its linguistic shallowness, however, limits its capabilities when applied to morphologically-rich languages and to language-pairs with highly divergent syntax. Integration of morphological analysis and syntactic modeling within statistical MT are currently at the forefront of MT research. This talk will overview recent work within my research group on hybrid MT frameworks that incorporate syntax and morphology into statistical translation.

The talk will focus on three main lines of work: (1) Morphological segmentation of Arabic and its impact on English-to-Arabic phrase-base SMT; (2) Learning of syntax-based synchronous context-free grammars from large volumes of parsed parallel corpora; and (3) Automatic Category Label Coarsening for Syntax-based MT.

Linguistic Indicators for Quality Estimation of Machine Translation

Lucia Specia
University of Wolverhampton
l.specia@wlv.ac.uk

Although significant progress has been observed in the field of Machine Translation (MT) in recent years, the quality of a given MT system can vary across translated segments. As MT becomes more popular among several types of users, an increasingly relevant problem is that of automatically assessing the quality of translations at the segment level to inform such users. In this talk I will present work on modelling the problem of quality estimation for different applications, focusing on the use of linguistic indicators contrasting the input and translation segments in order to complement shallow, language-independent and confidence-based indicators.

Improved Statistical Machine Translation Using MultiWord Expressions

Dhouha Bouamor
CEA-LIST, Vision and
Content Engineering Laboratory
F-92265 Fontenay aux roses,
France
dhouha.bouamor@cea.fr

Nasredine Semmar
CEA-LIST, Vision and Content
Engineering Laboratory
F-92265 Fontenay aux roses,
France
nasredine.semmar@cea.fr

Pierre Zweigenbaum
LIMSI-CNRS,
F-91403 Orsay,
France
pz@limsi.fr

Abstract

Identifying and translating a MultiWord Expression (MWE) in a text represents an issue for numerous applications in Natural Language Processing (NLP) as MWEs appear in all text genres and pose significant problems for every kind of NLP tasks. In this paper, we describe a hybrid approach for extracting contiguous MWEs and their translations in a French-English parallel corpus. We evaluate both the alignment and the translation quality. Next, we implement a method that integrates these units to Moses, the state of the art Machine Translation (MT) system. Conducted experiments show that MWEs improve translation performance.

1 Introduction

Statistical Machine Translation (SMT) initially focused on word to word translations (Brown et al., 1993). Various improvements of SMT systems quality used phrase-based translation (Koehn et al., 2003), defined simply as n-grams consistently translated in a parallel corpora. To compensate the lack of semantic information in phrase based approaches, we study bilingual MultiWord Expressions (MWEs) and integrate them in an existing phrase-based SMT system.

(Sag et al., 2002) define MWEs very roughly as *"idiosyncratic interpretations that cross word boundaries (or spaces)"*. Theses lexical units are numerous and constitute a significant portion of the lexicon of any natural language. (Jackendoff, 1997:156) estimates that the frequency of MWEs in a speaker's lexicon is almost equivalent to the frequency of single words. While easily mastered by native speakers, their interpretation poses a major challenge for computational systems, due to their flexible and heterogeneous nature. SMT does not model MWEs explicitly. In phrase based MT systems, these units are indirectly captured but they are not distinguished from any other n-gram.

In recent years, a number of techniques have been applied to the problem of MWEs extraction (Kupiec, 1993; Okita et al., 2010; Dagan and Church, 1994). Most of them based on identifiying these units within a corpus, with the goal of including them in bilingual lexicons (Smadja, 1993). Having such type of terms is useful for a variety of NLP application such as information retrieval (Vechtomova, 2005) , word sense disambiguation (Finlayson and Kulkarni, 2011) and others.

Some researches exploited MWEs in MT systems. (Tanaka and Baldwin, 2003) described an approach of noun-noun compound machine translation, but not significant comparison was presented. In (Lambert and Banchs, 2005), authors introduce a method in which a bilingual MWEs corpus was used to modify the word alignment in order to improve the translation quality. In their work, bilingual MWEs were grouped as one unique token before training alignment models. They showed on a small corpus, that both alignment quality and translation accuracy were improved. However, in their further study, they reported even lower BLEU scores after grouping MWEs by part-of-speech on a large corpus (Lambert and Banchs, 2006). Recently, (Ren et al., 2009) implemented a method integrating a domain bilingual MWE to Moses. The method yielded an improvement of 0.61 BLEU score compared with the baseline system.

In this paper, we describe a hybrid approach combining linguistic and statistical information to extract and align MWEs from a French-English parallel corpus. Extracted MWEs are then integrated into Moses. The conducted experiments show that MWEs identification improve the translation quality. The remainder of this paper is organized as follows. In section 2, we describe the proposed method for identifying and extracting bilingual MWEs. Experiments and results are discussed in section 3. We conclude and present our future work, in section 4.

2 Bilingual MultiWord Expressions Extraction

2.1 Related Work

A number of techniques have already been applied to the problem of MWEs extraction. Starting from a sentence aligned parallel corpus, most works rely on statistical, linguistic or hybrid approaches. The work of (Kupiec, 1993) is considered as one of the early work concerned with this task. The author focused essentially on noun groups. These units are identified on the basis of their part-of-speech tag. Then, based on the Expectation Maximization (EM) algorithm, bilingual correspondences are identified. It obtained a precision rate of 90% referred to the 100 first correspondences. An extension of this method is proposed by (Okita et al., 2010). To detect MWEs, a bidirectional version of Kupiec (1993) is applied. Then, in order to add prior information, they replace the maximum likelihood estimate in the M-Step of the EM algorithm with the Maximum-A-Posteriori (MAP) estimate. (Dagan and Church, 1994) describe a semi-automatic tool, Termight, which extracts technical noun groups using a syntactic pattern filter. They use a word alignment program to align MWEs. For each source term, the tool identifies a candidate translation by selecting a sequence of target words whose first and last word are aligned with any of the words in the source term. The accuracy obtained for 192 English-German correspondences is about 40%.

Other recent related work attempt to extend the linguistic based methods used in identifying MWEs. They use additional association measures such as Mutual Information (Daille, 2001) and the Log Likelihood Ratio (Wu and Chang, 2004; Seretan and Wehrli, 2007) to capture the degree of cohesion between the constituents of a MWE. However, these measures present two main shortcomings. They are designed for bigrams and require a definition of a threshold above which an extracted phrase is considered as a MWE. Afterwards, some heuristics, are applied for the alignment task. (Tufis and Ion, 2007) and (Seretan and Wehrli, 2007) assume that MWEs keep in most cases the same morphosyntactic structure in the source and target language, which is not universal such as the English MWE *small island developing* which is aligned with the French *insulaire en développement*. The Champollion system of (Smadja et al., 1996) can produce translations of a source MWEs in the target language. It is based on a multi-word unit extraction system, Xtract, developed by (Smadja, 1993). They first extracted source MWEs. After that, for each source term, they extracted its translations in the target language by testing Dice-score. Champollion was tested on the Hansard corpus and an accuracy of 73% was reported, taking into account only MWEs appearing at least 10 times in the corpus.

2.2 MWEs Identification

In this section, we describe the MWEs extraction method from a French-English parallel corpus. The process of extraction involves full morphosyntactic analysis of source and target texts. For this, we used the CEA LIST Multilingual Analysis platform (LIMA) (Besançon et al., 2010). The linguistic analyzer produces a set of part-of-speech tagged normalized lemmas. It is needed to only permit specific strings for extraction and filter out undesirable ones such as *of the, is a*. Since most MWEs consist of noun, adjectives and sometimes prepositions, we adopted a linguistic filter that accepts n-gram units ($2 \leq n \leq 4$) matching the morphosyntactic configurations presented in Table 1.

English Pattern	French Pattern
Adj-Noun	Noun-Adj
Noun-Noun	Adj-Noun
Past_Participle -Noun	Noun-Past_Participle
Adj-Adj-Noun	Noun-Noun-Adj
Adj-Noun-Adj	Noun-Adj-Adj
Adj-Noun-Noun	Adj-Noun-Adj
Noun-Prep-Noun	Noun-Prep-Noun
Noun-Prep-Adj-Noun	Noun-Prep-Noun-Adj
Adj-Noun-Prep-Noun	Noun-Adj-Prep-Noun

Table 1: French and English MWE's morphosyntactic structure

To this list are added some prepositional idiomatic expressions (*in particular, in the light of, as regards...*) and proper noun (*Midle East, South Africa, El-Salvador...*) recognized by the morphosyntactic analyzer. Then, we scored them with their total frequency of occurrence in the corpus.

To avoid an over-generation of MWEs and remove irrelevant candidates from the process, a redundancy cleaning approach is introduced. In this approach, if a MWE is nested in another, and they both have the same frequency, we discard the smaller one. Otherwise we keep both of them. We consider also the alternative of having a MWE that appears nested in a high number of terms. We followed (Frantzie et al., 2000) by discarding all longer MWEs. An example of extracted MWEs is in Table 2.

The presented approach does not use additional correlations statistics such as Mutual Infomation or Log Likelohood Ratio since these measures require a definition of a threshold above which an extracted phrase is considered as a MWE or not. Our method consider that all extracted units are effective and valid and include all of them in the translation process. To our knowledge, none of other approaches can make this claim.

Freq	French MWEs
144	Parlement européen
25	Prestation de service
29	Industrie automobile allemand
36	Chemin de fer
65	En particulier
32	Source d'énergie renouvelable
11	Mise en place
Freq	**English MWEs**
19	Court of first instance
316	Member state
19	Point of view
65	In particular
29	Plenary meeting
32	Rural development
21	European public prosecutor

Table 2: A sample of extracted French and English MWEs

2.3 MWEs Alignment

We present a method in which we try to find for each MWE in a source language, a translation to which is adequate in the target one. We focus only on many-to-many correspondences and do not use any dictionary nor simple-word alignment tools. Our algorithm is quite simple and based on the Vector Space Model (VSM). VSM (Salton et al., 1975) is a well-known algebraic model used in information retrieval, indexing and relevance ranking. Each MWE is represented by a binary vector of size n[1] indicating for each sentence of the corpus whether it occurs in that sentence or not. Then, translation pairs of MWEs are extracted by means of the following iterative process:

1. Find the most frequent MWE in the source sentence.
2. Extract all translation candidates from the target parallel sentence.
3. Compute a confidence value for the translation relation.
4. Consider that the target MWE that maximize the confidence value is the best translation.
5. Discard the translation pair from the process and go back to 1.

To compute the confidence value, we adopted the Jaccard Index, a frequently used measure in information retrieval. It is defined as

$$IJ = \frac{NS_i}{NS_s + NS_t - NS_i} \tag{1}$$

and based on the number NS_i of sentences shared by each target and a source MWE. This is normalized by the sum of the number of sentences where the source and target MWEs appear independently of each other (NS_s and NS_t) decreased by NS_i.

2.4 Extraction Method Evaluation

To evaluate the alignment quality, we followed the evaluation framework defined in the shared task on word alignment organized as part of the HLT/NAACL 2003 Workshop on building and using parallel corpora (Mihalcea and Pedersen, 2003). Within this framework, participating teams were provided with data and asked to provide automatically derived word alignments for all the words in the test set, following a specific format. This framework is defined to evaluate simple-word alignment algorithms, but we adapted it to evaluate our MWEs alignment system. The alignment results are compared to a manually aligned reference corpus scored with respect to precision, recall and F-measure, where A is the alignment proposed by the system and G is a gold standard alignment. Because the manual construction of the alignment reference is a difficult and time-consuming task, we conducted a small-scale evaluation based on a small set of 100 French-English aligned sentences derived from the Europarl corpus.

$$P = \frac{|A \cap G|}{|A|} \tag{2}$$

$$R = \frac{|A \cap G|}{|G|} \tag{3}$$

$$F = \frac{2P * R}{P + R} \tag{4}$$

Our method yeilds a precision of 63,93% , a recall of 62,46% and an F-measure of 63.19%. We consider that obtained results are satisfactory and encouraging. In table 3 we give an example of MWEs aligned by our technique.

French → English MWEs
european parliament /parlement européen
military coup / coup d'état
in favour of /en faveur de
no smoking area/ zone non fumeur
small island developing / insulaire en développement
good faith / de bonne foi
competition policy / politique de concurrence
process of consultation / processus de consultation
railway sector / chemin de fer
with regard to / en ce qui concerne
cut in forestation / coupe forestier

Table 3: Sample of aligned MWEs

[1] n=number of the aligned sentences of parallel corpora

From observing some couples of MWEs, we have identified a class of error caused by the choice of n-gram's size. Since our system does not capture one-to-many correspondences, some MWEs were not aligned correctly. For example, the French MWE *chemin de fer* corresponding normally to the simpe word *railway* was aligned here by the MWE *railway sector*.

3 Experiments

3.1 Application of MWEs

In the previous section, we described the approach we followed to extract translation pairs of MWEs, and evaluated it by comparing the list of extracted MWEs to a hand-created reference list. As it lacks a common benchmark data sets for evaluation in MWE extraction and alignment researches, we decided to study in what respect these units are useful to improve the performance of phrase based SMT systems. We present a method that integrates extracted MWEs into the baseline system's phrase table being considered as very important element according to the following two ways. In the first way, we simply add MWEs and keep translation probabilities proposed by the aligner. We call this method "Baseline+MWE". In the second one, "Baseline+NPMWE", we assign 1 to the two translation probabilities (in both directions) for simplicity.

3.2 Baseline

We use the factored translation model of the Moses[2] SMT system as our baseline system (Koehn, 2005). It is an extention of the phrase based models which are limited to the mappings of phrases without any explicit use of linguistic information. The factored model enables the use of additional annotations at the word level. We present a model that operates on lemmas instead of surface forms, in which the translation process is broken up into a sequence of mapping steps that either :

- Translate source lemmas into target's ones.
- Generate surface forms given the lemma.

The features used in baseline system are: two translation probability features, two languages models, one generation model and word penalty. For the "Baseline+MWE" and "Baseline+NPMWE" methods, translation pairs of MWEs were extracted from the training corpus and added to the phrase table. Consequently, a new phrase table is obtained. During the translation process, the decoder would search for each phrase in input sentence, all candidates translations in both original phrases and new MWEs.

3.3 Data

Training and Test data (Table 4) come from the French-English Europarl Corpus (Koehn, 2005). It groups a set of parallel sentences extracted from the Proceedings of the European Parliament. In this work, we focus on sentences consisting of at most 50 words.

	French	English
Training sentences	9002	
Words	213489	206562
Test sentences	500	
Words	13816	12736

Table 4: Caracteristics of Training and Test data

Since we use the factored translation model, training data are annotated with lemmas. Next, word-alignment for all the sentences in the parallel training corpus is established. Here, we use the same methodology as in phrase-based models (symmetrized GIZA++ alignments). The word alignment methods operates on lemmas. We also specified two language models using the IRST Language Modeling Toolkit [3] to train two tri-gram models. Besides the regular language model based on surface forms, we have a second language model which is trained on lemmas.

3.4 Results and discussion

We test translation quality on the test set described in the previous section and calculate the BLEU score. We also consider only one reference for each test sentence. Obtained BLEU results are reported in Table 5. The first notable observation is that using bilingual MWEs improves translation in the two cases. The "Baseline+MWE" method achieves the most improvement of 0.24 BLEU score compared to the baseline system. This method performs slightly higher than the "Baseline+NPMWE" method which in turn comes with 0.23 BLEU score improvement.

Method	BLEU
Baseline	0,1758
Baseline+MWE	**0,1782**
Baseline+NPMWE	0.1781

Table 5: Translation results using extracted MWEs

In order to know in what respects our method improves performance of translations, we manually analyzed the test sentence presented in Table 6. The french MWE "chemins de fer" is not correctly aligned in baseline system. It was translated to the english phrase "way of the

[2] http://www.statmt.org/moses

[3] http://hlt.fbk.eu/en/irstlm

Source Sentence	Ce n'est que ces dernières années que la plupart des **états membres** ont investi dans l'amélioration des **chemins de fer** et parfois également dans la navigation intérieure.
Reference	Only in the last few years have most **member states** invested in improving the **railways** and sometimes inland shipping too.
Baseline	They will be that this last year that most **member states** have invested in improving the **way to go to fer** and sometimes also in the navigation internal.
Baseline+MWE	They will be that this last year that most **member states** have invested in improving the **railways sector** and sometimes also in the internal navigation.

Table 6: Translation example

fer". We can notice that in this case,a word-to-word alignment strategy is performed. It provides the following alignments:

- ”chemin”=”way to go to”
- ”de”= Not Translated
- ”fer”=Not translated

Here, the French word ”chemin” was translated into the English phrase ”way to go to” and the word ”fer” was not translated since there is no entry in the baseline system's phrase table to which we can associate it. While it is aligned to the target MWE ”railways sector” in baseline+MWE. We can consider that this is a correctly translated phrase as much as it keeps the same meaning.

4 Conclusion and Future Work

We described a method for extracting and aligning MWEs in a parallel corpus. The alignment algorithm we proposed checks only on many to many correspondences and can address both frequent and infrequent MWEs in a text. To evaluate the alignment quality, we used a small test set of 100 parallel sentences and reported an F-Measure value of 63,19%.

We also proposed a method for using extracted bilingual MWEs in Statistical Machine Translation. This method incorporates extracted MWEs in a baseline system's phrase table. Conducted experiments show that including such type of units in the translation process improves translation quality and yeilds an improvement of 0.24 BLEU score compared to a baseline system.

Although our initial experiments are positive, we believe that they can be improved in a number of ways. We fisrt intend to extend the morphosyntactic patterns to handle other forms of MWEs, e.g. starting with a verb. We will also try to develop and evaluate other statistical based methods to align MWEs.

Moreover, in the presented work the use of MWEs is actually restricted to the decoding step. We will also attempt to include these units in the training step using a larger set a parallel sentences and the two sides of MWEs as independent monolingual units.

Acknowledgments

This research work is supported by the FINANCIAL-WATCH (QNRF NPRP: 08-583-1-101) project.

References

Besançon R., De Chalendar G., Ferret O., Gara F., Laib M., Mesnard O., and Semmar N. (2010). *LIMA :A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation*. Proceedings of LREC, Valetta, Malta.

Brown P., Della Pietra S., Della Pietra V. and Mercer R. (1993). *The mathematics of statistical machine translation: Parameter estimation.*. Computational linguistics.

Daille B. (2001). *Extraction de collocation à partir de textes*. Proceedings of TALN, Tours, France.

Dagan I. and Church K. (1994). *Termight: Identifying and translating technical terminology*. Proceedings of the 4th Conference on ANLP, Stuttgart, Germany, p. 34-40.

Finlayson M. and Kulkarni N. (2011). *Detecting Multi-Word Expressions Improves Word Sense Disambiguation*. Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World. Portland, Oregon, USA. p.20-24

Frantzie C., Ananiadou S., and Mima H. (2000). *Automatic recognition of multi-word terms: the C-Value/NC-Value method.* Int. J. on Digital Libraries 3(2): 115-130.

Jackendoff R. (1997). *The Architecture of the Language Faculty*. Cambridge (Mass.), MIT Press.

Kupiec J. (1993). *An algorithm for finding noun phrases correspondences in bilingual corpora*. Proceeding of the 31st annual Meeting of the Association for Computational Linguistics. Columbus, Ohio, USA. p. 17-22.

Koehn P. (2005). *Europarl: A parallel Corpus for Statistical Machine Translation.* Proceeding of MT-SUMMIT

Koehn P and Hoang H. (2005). *Factored Translation Model.* Proceeding of MT-SUMMIT

Koehn P., Och F. and Marcu D. (2003). *Statistical Phrase-Based Translation*. Proceeding of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Edmonton, Canada. p 115-124.

Lambert P. and Banchs R. 2005. *Data Inferred Multi-word Expressions for Statistical Machine Translation*. Proceeding of MT SUMMIT.

Lambert P. and Banchs R. 2006. *Grouping Multi-word Expressions According to Part-Of-Speech in statistical Machine Translation.*. Proceeding of the Workshop on Multi-word Expressions in a multilingual context.

Mihalcea R. and Pedersen T. (2003). *An evaluation exercise for Word Alignment*. Proceedings of the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond. Edmonton, Canada. p. 1-10.

Okita T., Guerra M. Alfredo, Graham Y., and Way A. (2010). *Multi-Word Expression Sensitive Word Alignment*. Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010, Beijing, p. 26–34.

Ren Z., Lu Y., Liu Q, and Huang Y. (2009). *Improving statistical machine translation using domain bilingual multiword expressions*. In Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications, p 47–54.

Sag I , Baldwin T., Francis Bond F, Copestake A, and Flickinger D. (2002). *Multiword Expressions:A Pain in the Neck for NLP*. CICLing 2002 Mexico City, Mexico.

Salton G. , Wong A. , and Yang C. S. (1975). *A Vector Space Model for Automatic Indexing*. Communications of the ACM, vol.18 p. 613–620.

Seretan V. and Wehrli E. (2007). *Collocation translation based on sentence alignment and parsing*. Proceedings of TALN. Toulouse, France.

Smadja F. (1993). *Retrieving collocations from text:Xtract*. Computational Linguistics. vol.19 p.143-177.

Smadja F., McKeown K., and Hatzivassiloglou V. (1996). *Translating collocations for bilingual lexicons: A Statistical Approach*. Computational Linguistics. p.1-38.

Tanaka T and Baldwin T. (2003). *Noun-noun compound machine translation: A feasibility study on shallow processing*. In Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.

Tufis I. and Ion R. (2007). *Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure*. Proceedings of the 4th International Conference on Speech and Dialogue Systems, Iaşi, Romania, p.183-195.

Vechtomova O. (2005). *The role of multi-word units in interactive information retrieval*. In D.E. Losada and J.M. Fernández-Luna, editors, ECIR 2005, LNCS 3408, p 403–420. Springer-Verlag, Berlin. alia Kordoni, Carlos Ramisch, Aline Villavicencio

Wu C. and Chang S. Jason. (2004). *Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses*. Computational Linguistics. p.1-20.

VERTa: Exploring a Multidimensional Linguistically-Motivated Metric

Elisabet Comelles and
Irene Castellón
Grial Research Group
Universitat de Barcelona
Barcelona, Spain
{elicomelles, icastellon}@ub.edu

Jordi Atserias
Fundació Barcelona Media
Barcelona, Spain
jordi@yahoo-inc.com

Victoria Arranz and
Olivier Hamon
ELDA/ELRA
Paris, France
{arranz, hamon}@elda.org

Abstract

This paper describes the first steps in the design and implementation of VERTa, a metric which aims at using and combining a wide variety of linguistic features at lexical, morphological, syntactic and semantic level. A description of the modules developed up to now is provided, as well as the results of some preliminary experiments conducted in order to modify and improve the metric. No formal evaluation has been performed so far because we are in our first stages, but for the sake of comparison we report some results obtained when comparing our current metric performance with IBM's BLEU.

1 Introduction

Evaluation of MT systems is crucial in their development and improvement. However, human evaluation is expensive and complex. As a consequence, in the last decades several automatic metrics have been developed in order to assess MT output in a simple and less expensive way. From these automatic metrics, the string-based IBM's BLEU (Papineni et al. 2002) is one of the most popular and widely-spread because it is fast and easy to use. However, researchers such as Callison-Burch et al. (2006) and Lavie and Dekowski (2009) have criticized its performance and highlighted its weaknesses in relation to translation quality and its tendency to favour statistically-based MT systems. As a consequence, in response to BLEU weaknesses several linguistically-motivated metrics have arisen. Some of them are based on lexical information, such as METEOR (Banerjee and Lavie 2005); others rely on the use of syntax, either using constituent (Liu and Hildea 2005) or dependency analysis (Owczarzack et al. 2007a and 2007b; He et al. 2010); and others use semantic information, such as Named Entities and semantic roles (Giménez and Márquez 2007 and 2008a). All these metrics work at a certain linguistic level, but little research (Giménez 2008b; Specia and Giménez 2010) has been focused on the use and combination of a wide variety of linguistic information. Therefore, our proposal is a linguistically-motivated metric which aims at using and combining varied linguistic knowledge at different levels in order to cover the key features that must be considered when dealing with MT evaluation from a linguistic point of view. Our hypothesis is that the use and combination of linguistic features at different levels will help us to provide a wider and more accurate coverage than those metrics working at a specific linguistic level.

This paper describes the first stages in the design and the on-going development of the VERTa metric. We provide a description of the modules developed up to now and we report the results obtained in some preliminary experiments which will help us to see whether we are in the right direction and to discuss the use of certain linguistic knowledge used for the time being. Finally, we draw some conclusions and point out some items which must be under consideration for further development and improvement.

2 Methodology

When approaching MT evaluation from a linguistic point of view, there are different linguistic phenomena which should be taken into account. This can help in the design of our metric and can play an important role when evaluating MT output. In order to define such phenomena, we have considered linguistic issues that we had come across during some work on language data analysis carried out. After such study, we concluded that these phenomena could be classified into lexical, phrase and clause level and that they affected both syntax and semantics. Therefore, the linguistic knowledge that we intend to use is organised in different layers:

- **Lexical information:** We use word-forms and lemmas in order to check lexical units similarity and we also take into account lexical semantic relations such as synonymy, hyperonymy and hyponymy, in other words, semantically-related lexical items.
- **Morphological information:** The information at this level is basically based on lemmas, semantically-related units and the use of Part of Speech tags as the main features in order to cover issues related to inflectional morphology and morphosyntax.
- **Dependency information:** We take into account the dependency relations between the constituents of a sentence. By means of this information we try to solve issues on different word order between the hypothesis and the reference translation. In order to allow a broad coverage, the dependency module is based on the lexical information obtained in the lexical level (see section 2.3)
- **Sentence semantics:** We intend to deal with semantics at sentence level, focusing on semantic arguments.

The use of this varied range of linguistic information allows us to evaluate both adequacy and fluency, thus trying to get closer to human evaluation scores. Given the stage of our work, in this paper we only focus on adequacy for the time being.

In order to combine the above described linguistic features, we have decided to develop one similarity metric per each type of information: lexical similarity metric, morphological similarity metric, dependency similarity and semantic similarity metric respectively. Moreover, we have also added an n-gram similarity module so as to account for similarity between chunks. Each metric works first individually and the final score is the Fmean of the weighted combination of the Precision and Recall of each metric in order to get the results which best correlate with human assessment.

All metrics use a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triples, n-grams, etc) as shown below.

$$P = \frac{\sum_{\partial \in D} W_{\partial} * nmatch_{\partial}(\nabla(h))}{|\nabla(h)|}$$

$$R = \frac{\sum_{\partial \in D} W_{\partial} * nmatch_{\partial}(\nabla(r))}{|\nabla(r)|}$$

Where r is the reference, h is the hypothesis and ∇ is a function that given a segment will return the elements of each level (e.g. words at lexical level and triples at dependency level). D is the set of different functions to project the level element into the features associated to each level, such as word-form, lemma or partial-lemma at lexical level. $nmatch_{\partial}()$ is a function that returns the number of matches according to the feature ∂ (i.e. the number of lexical matches at the lexical level or the number of dependency triples that match at the dependency level). Finally, W is the set of weights]0 1] associated to each of the different features in a particular level in order to combine the different kinds of matches considered in that level.

Thus far, the metrics implemented are the lexical and morphological similarity metrics, the n-gram similarity metric and part of the dependency metric. As regards the semantic similarity metric, it has not been explored so far, but we intend to do it in the future. The metric is based on precision and recall and the traditional F-measure is applied in order to get the final score for each pair of segments. In the case of using multiple reference translations, the VERTa metric compares each hypothesis string with the corresponding string of each reference translation and the metric chooses the best score as the final score for that segment.

VERTa works at segment level, comparing the different items of the hypothesis and reference segments from left to right. It must be highlighted that a segment can be composed of one or more sentences. Thus, it could be the case that one segment of the hypothesis contains just one sentence whereas the same segment in the reference has been translated by means of two different sentences, which still belong to the same segment. In order to deal with this issue, segments are split into sentences and the linguistic tools (see sections 2.2 and 2.3 for further details) used in each stage are applied to each sentence separately. Afterwards the metric is applied at segment level; that is to say, we look for the similarity of all items inside the hypothesis segment in relation to all items in the reference segment, regardless of the number of sentences in each segment.

We describe each module in detail in the following sections.

2.1 Lexical Similarity Module

The lexical similarity metric compares lexical items from the hypothesis segment with those in the reference segment. In order to identify these matches we use the following linguistic features: word-forms, lemmas, synonyms, hyperonyms, hyponyms and partial lemmas (lemmas that share the first 4 letters). The approach followed in this module is inspired by METEOR in the sense that the metric relies on lexical items and lexical semantic relations. However, while the most recent version of METEOR (Denkowsi & Lavie, 2011) deals with semantics by means of synonymy and paraphrase tables, our metric uses not only synonymy but tries to exploit other lexical semantic relations such as hyperonymy and hyponymy and avoids the use of paraphrase tables which have to be built up for each language and domain. Moreover, we also use the information provided by lemmas, whereas METEOR relies on stemming. In addition, we also apply a system of weights (W) on the different matches established depending on their importance in terms of semantics, whereas METEOR considers all matches equal, regardless of their difference in terms of meaning.

From the linguistic features that we use, lemmas are obtained by means of WordNet (Feullbaum, 1998). Also the metric relies on some lexical semantic relations such as synonymy, direct hiperonymy and direct hyponymy. These semantic relations are also identified using Wordnet 3.0; however, in order to establish semantic relations we do not use any disambiguation tool, we rely directly on lemmas. As mentioned later in the Experiments section, we thought the use of hyperonymy and hyponymy was a useful strategy to gain more lexical coverage. First we tried to use different levels of hyperonymy and hyponymy but we realised that they introduced noise in the metric, so we decided to restrict their use at immediate levels. However, as shown later, the use of such semantic relations must be reconsidered as they do not always help.

Once established the different linguistic features used by the lexical similarity metric we focus, now, on its mechanism. The metric finds matches between the hypothesis and the reference segment by using the linguistic features explained above in the order established in Table 1.

	W	Match	Examples	
			HYP	**REF**
1	1	Word-forms	*east*	*east*
2	1	Synonyms	*believed*	*considered*
3	.9	Direct-hypern.	*barrel*	*keg*
4	.9	Direct-hypon.	*keg*	*barrel*
5	.8	Lemma	*is_BE*	*are_BE*
6	.7	Partial-lemma	*danger*	*dangerous*

Table 1. Lexical matches and examples

2.2 Morphological Similarity Module

The morphological similarity metric combines lexical and morphological information. This metric is based on the matches set in the lexical similarity metric, except for the partial-match, in combination with the Part of Speech (POS) tags from the annotated corpus[1]. By means of this combination, we apply a restriction in terms of fluency because we avoid issues such as stating that *invites* and *invite* are positive matches regarding morphology, and somehow we compensate the broader coverage that we have in the lexical module. Therefore, when assessing MT output in terms of fluency this metric will receive a higher weight, whereas when evaluating adequacy, the weight given to this module will be reduced. This module will be particularly useful when evaluating MT output of languages with a rich inflectional morphology, such as Spanish or Catalan.

Following the approach used in the lexical similarity metric, the morphological similarity metric establishes matches between items in the hypothesis and the reference sentence and a set of weights (W) is applied. However, instead of comparing single lexical items as in the previous module, in this module we compare pairs of features in the order established in Table 2.

	W	**Match**	**Examples**	
			HYP	**REF**
1	1	(Word-form, POS)	(he, PRP)	(he, PRP)
2	1	(Synonym, POS)	(VIEW, NNS)	(OPINON, NNS)
3	.9	(Hypern., POS)	(PUBLICATION, NN)	(MAGAZINE, NN)
4	.9	(Hypon., POS)	(MAGAZINE, NN)	(PUBLICATION, NN)
5	.8	(LEMMA, POS)	can_(CAN, MD)	Could_(CAN, MD)

Table 2. Morphological pairs of matches and examples.

2.3 Dependency Similarity Module – Work in progress

Once covered the lexical and morphological sections, we are now working on the dependency similarity metric which will help us to deal with syntactic structures at a deeper level. By means of this module we will be able to capture the relations between sentence constituents regardless of their position inside the sentence, which will be really helpful when comparing a hypothesis and a reference segment with a different word order of their constituents, as illustrated in the following example:

<u>Example 1</u>:
HYP: ***After a meeting Monday night with the head of Egyptian intelligence chief Omar Suleiman*** *Haniya said....*
REF: *Haniya said,* ***after a meeting on Monday evening with the head of Egyptian Intelligence General Omar Suleiman****...*

In this example, the adjunct realised by the PP *After a meeting Monday night with the head of Egyptian intelligence chief Omar Suleiman* occupies different positions in the hypothesis and reference strings. In the hypothesis it is located at the beginning of the sentence, preceding the subject *Haniya*, whereas in the reference, it is placed after the verb. By means of dependencies, we can state that although located differently inside the sentence both subject and adjunct depend on the verb as shown in Table 3.

HYPOTHESIS	**REFERENCE**
nsubj(Haniya, said)	nsubj(Haniya, said)
prep_after(meeting, said)	prep_after(meeting, said)

Table 3. Matching of triples

Therefore, the use of dependencies helps us to establish similarities between equivalent sentences which contain the same constituents but in different positions.

This dependency similarity metric works at sentence level and follows the approach used by Owczarzack et al. (2007a and 2007b) and He et al. (2010) with some linguistic additions in order to adapt it to our metric combination.

Both hypothesis and reference strings are annotated with dependency relations by means of the Stanford parser (de Marneffe et al. 2006). The reason why this parser is used is because after conducting an evaluation (Comelles et al. 2010) where the performance of several dependency parsers was assessed (Stanford, DeSR, MALT, Minipar, RASP) this proved to be the best in terms of linguistic quality. Moreover, the output file provided by this parser contains dependency relations by means of flat triples with the form **Label(Head, Mod)**. These triples are ideal in order

[1] The corpus has been annotated with POS tags using the Stanford Parser (de Marneffe et al. 2006).

to compare the dependency relations in the hypothesis and reference segments.

The dependency similarity metric also relies first on the matches established at lexical level − word-form, synonymy, hyperonymy, hyponymy and lemma − in order to capture lexical variation across dependencies and avoid relying only on surface word-form. Then, and inspired by He et al. (2010) and Owczarzak et al. (2007a and 2007b), four different types of dependency matches have been designed. Next, we describe the matches and provide examples for each of them:

- Complete (MC): Type of match used when the triples are identical, this means that the label, the head and the modifier match.

  Label1(Head1,Mod1) = Label1(Head2,Mod2)
  Example 2:
  HYP: advmod(difficult, more)
  REF: advmod (difficult, more)

- Partial (MP): Three different types of partial matches are established:
  - Partial_no_mod (MP_no_mod): The label and the head match but the modifier does not match
    - Label1 = Label2
    - Head1 = Head2

    Example 3:
    HYP:**conj_and(difficult**, dangerous)
    REF: **conj_and(difficult**, serious)

  - Partial_no_head (MP_no_head): The label and the modifier match but the head does not match.
    - Label1 = Label2
    - Mod1 = Mod2

    Example 4:
    HYP: prep_between(mentioned, Lebanon)
    REF: prep_between(crisis, Lebanon)

  - Partial_no_label (MP_no_label): The head and the modifier match but the label does not match.
    - Head1 = Head2
    - Mod1 = Mod2

    Example 5:
    HYP: predet(**parties**, **all**)
    REF: det(**parties**,**all**)

Each type of match is given a weight which ranges from the highest to the lowest weight in the following order:

- Complete (1)
- Partial_no_mod (.8)
- Partial_no_head (.7)
- Partial_no_label (.7)

In addition, we have also planned to add some extra-rules in order to capture the similarity between certain structures which are semantically equal but syntactically different. These extra-rules will be applied at phrase and sentence level. An example of these rules at phrase level affects modifiers inside the noun phrase and the latter the passive-active voice alternation. We plan to cover the similarity between an adjective premodifiying a noun and an of-prepositional phrase postmodifying it, as exemplified below.

Example 6:
HYP: *...between the* ***ministries of interior****...*
REF: *...between the two* ***interior ministries****...*

HYP_prep_of(ministries, interior) = REF_amod(ministries, interior)

Although their labels differ, this couple of triples must be considered as an exact match due to their semantic similarity. Otherwise we would penalise a couple of structures which are equal from a semantic point of view. At a clause level, an example of these rules could be the treatment of the active-passive alternation. As shown below, although syntactically different, both structures share the same meaning.

Example 7:
HYP: *After meeting* ***the Moroccan news agency published*** *a joint statement...*
REF: *A joint statement* ***published*** *(...)* ***by the Moroccan news agency****...*
HYP_nsubj(published, agency) = REF_agent(published, agency)

Similar to the pair of dependencies dealing with modifiers, *nsubj* and *agent* labels must be considered identical and thus, the previous couple of triples must be scored as an exact match.

Unfortunately, this set of rules has not been implemented yet in the dependency metric. Therefore, results shown in the Experiments section only refer to the use of the different matches.

2.4 N-gram Similarity Module

The n-gram similarity module is aimed at matching chunks[2] in the hypothesis and reference segments. Chunks length goes from bigrams to sentence length. The use of this module allows us to combine both linguistic and statistical approaches and enables us to deal with word order inside the sentence by means of a more simple approach than the parsing of constituents. The n-gram similarity module uses the matches obtained at lexical level in order to align chunks. Thus, we do not only match n-grams relying on the word-form but also taking into account synonymy, hyponymy/hyperonymy and lemmas, as shown in example 8, where the chunks [*the situation in the area*] and [*the situation in the region*] match, although *area* and *region* do not share the same word-form but a relation of synonymy.

Example 8:
HYP: … the situation in the *area*…
REF: … the situation in the *region*…

2.5 Metrics Combination

As mentioned at the beginning of the section, the modules implemented so far are combined in order to cover linguistic features at all levels depending on the type of evaluation. Therefore, if the evaluation is focused on adequacy, those modules more related to semantics will have a higher weight, whereas if evaluating fluency those related to morphology, morphosyntax and constituent word order will be more important. Moreover, metrics should also be combined depending on the type of language evaluated. If a language with a rich inflectional morphology such as Spanish is assessed, the morphology module should be given a higher weight; whereas if the language evaluated does not show such a rich inflectional morphology (i.e. English) the weight of the morphology module should be lower. As a consequence, a set of weights has been established which can be changed manually regarding the type of evaluation. So far weights have been set according to the linguistic characteristics of the language under analysis and the type of evaluation. In a near future we intend to work on the tuning of weights in order to improve the metric performance. The experiments described in the next section are all focused on evaluating adequacy, as a consequence, the lexical and dependency metrics receive higher weights than the morphology and n-gram similarity metrics. For these experiments weights have been set as follows:

- Lexical Module: 0.444
- Morphology Module: 0.111
- N-gram Module: 0.111
- Dependency Module: 0.333

3. Experiments

In this section we report a couple of preliminary experiments at segment and system level to check whether we were in the right direction. These experiments should not be regarded as a formal evaluation, but just as a set of preliminary tests which should give us information on the adequacy of the linguistic features used. They must provide us with material to discuss, reconsider and improve the on-going development of the metric. The experiments were aimed at checking (i) the influence of adding the dependency module and (ii) the influence of hyperonyms and hyponyms. For these experiments we used data provided in the MetricsMaTr 2010 shared-task[3]. From the data provided by the organization we used 100 segments of the NIST Open-MT06 data, the MT output from 8 different MT systems (a total of 28,000 words approximately) and 4 reference translations. The human judgments used were based on adequacy. In order to calculate correlations at segment level we used Pearson correlation and we took into account all segments regardless of the system providing them in order to have a more precise correlation. Table 4 shows the results obtained.

[2] By chunks we understand a group of words that go together, one next to the other, not necessarily working as a constituent

[3] http://www.nist.gov/itl/iad/mig/metricsmatr10.cfm

	NO DEP + HYP	DEP + HYP	DEP + NO HYP.
Pearson Correlation	0.734	0.755	0.759

Table 4. Pearson correlations at segment level

On the one hand, the use of the partially-implemented dependency module improved the performance of the metric. Thus, adding linguistic knowledge which deals with deep structure at clause and phrase level helped to account for certain relationships which would not be considered by means of the n-gram matching module, such as different word order of the constituents inside the sentence. On the other hand, and opposed to our hypothesis, at segment level, the metric correlates better with human judgments when lexical semantic relations are more restricted. It seems therefore that the use of direct hyperonyms and hyponyms does not help to improve the metric performance; on the contrary, it slightly degrades the correlation with human judgments. There might be a couple of reasons for this result: first, a low percentage of hyponyms and hyperonyms in the reference translations; secondly, the fact of not using any process of disambiguation might make the metric match certain words which, although being hyponyms or hyperonyms, do not share such a relationship in the domain under analysis.

For the sake of comparison and just to check that our first steps were in the right direction, we were also interested in comparing our metric with the widely-used metric BLEU. As shown in Table 5 the results obtained by our metric at system level, although being yet in its first stages, outperforms the results obtained by IBM's BLEU at both system and segment level, due to the use of more lexical semantic information by our metric and the calculation of recall.

Metric	Pearson Correlation	
	Segment	System
VERTa	**0.759**	**0.970**
BLEU	0.683	0.931

Table 5. Metric comparison at segment and system level

4. Conclusions and Future Work

In this paper we have describe the work in progress of the metric we are developing. We have described the modules of the metric which have been designed and implemented so far and we reported the results obtained in some preliminary experiments. The scores obtained in the correlations with human judgments show that the use of linguistic information dealing with different types of linguistic phenomena and at different levels helps in improving the metric performance. Although they are preliminary results, they will be extremely helpful to continue with our on-going research. Moreover, the figures obtained by our primary metric implementation when compared to BLEU show promising results for the combination and use of a wide variety of linguistic features.

In a near future, we plan to keep working on the development of the metric by exploring the use of other linguistic information (i.e. multi-words treatment, the importance of function and content words and the use of semantic information at sentence level). In addition, we also expect to improve the metric performance by finishing the implementation of the dependency module (i.e. refining the type of dependency labels and matches to take into account, and implementing the set of similarity rules) and continue working on the tuning of the weights used both inside the modules and in metrics combination. Regarding the meta-evaluation of the metric, we will analyze the coverage of each level separately and we will evaluate our metric not only in terms of adequacy but also in terms of fluency. Finally, we would also like to test the robustness of VERTa with other languages with richer inflectional morphology such as Spanish.

Acknowledgments

We are very grateful to LDC for kindly providing the development data used in the MetricsMaTr 2010 shared-task.

This work has been partially funded by the Spanish Government (projects KNOW2, TIN-2009-14715-C04-03, and Holopedia, TIN2010-21128-C02-02).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005),* pages 65-72, Ann Arbor, Michigan.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research in *Proceedings of the EACL 2006*, pages 249–256.

Elisabet Comelles, Victoria Arranz and Irene Castellon. 2010. Constituency and Dependency Parsers Evaluation. SEPLN (ed.), *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. 45, pages 59-66. SEPLN. Valencia. ISSN: 1135-5948

Michael J. Denkowski and Alon Lavie. 2011. METEOR 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems in *Proceedings of the 6th Workshop on Statistical Machine Translation (ACL-2011)*, pages 85–91, Edinburgh, Scotland, UK.

Alon Lavie and Michael J. Denkowski. 2009. The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation, 23.*

Christian Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogeneous MT systems in *Proceedings of the 2nd Workshop on Statistical Machine Translation (ACL),* pages 256-264, Prague, Czech Repubilc.

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation in *Proceedings of the 3rd Workshop on Statistical Machine Translation (ACL),* pages 195-198, Columbus. OH.

Jesús Gimenez. 2008. Empirical Machine Translation and its Evaluation. Doctoral Dissertation. UPC.

Yifan He, Jinhua Du, Andy Way and Josef van Genabith. 2010. The DCU Dependency-based Metric in WMT-Metrics MATR 2010. *In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010),* pages 349-353, Uppsala, Sweden.

Ding Liu and Daniel Hildea. 2005. Syntactic Features for Evaluation of Machine Translation in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses in *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation (LREC-2006).* Genoa, Italy.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. in *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, Manchester, UK.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation in *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation,* pages 80–87, Rochester, New York.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Labelled Dependencies in Machine Translation Evaluation in *Proceedings of the ACL Workshop on Statistical Machine Translation,* pages 104– 111, Prague, Czech Republic.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL02), pages 311-318. Philadelphia. PA.

Lucia Specia and Jesús Giménez. 2010. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. In the *Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.

Using Apertium linguistic data for tokenization to improve Moses SMT performance

Sergio Ortiz Rojas, Santiago Cortés Vaíllo
Prompsit Language Engineering
Campus UMH, Edficio Quorum III
Avda Universitat s/n, E-03202 Elx, Spain
{sergio,santiago}@prompsit.com

Abstract

This paper describes a new method to tokenize texts, both to train a Moses SMT system and to be used during the translation process. The new method involves reusing the morphological analyser and part-of-speech tagger of the Apertium rule-based machine translation system to enrich the default tokenization used in Moses with part-of-speech-based truecasing, multi-word-unit chunking, number preprocessing and fixed translation patterns. Figures of the experimental results show an improvement of the final quality similar to the improvement attained by using minimum-error-rate training (MERT) as well as an increase of the overall consistency of the output.

1 Introduction

Apertium (Tyers et al., 2009) is a free/open-source machine translation (MT) platform that provides rule-based MT (RBMT) systems for an increasing number of languages. Apertium uses human-built linguistic data, consisting of dictionaries (both monolingual for morphological analysis and generation, and bilingual for translation purposes) and transfer rules intended to perform transformations involving more than a translation unit. These data are available for more than 30 different languages, in different degrees of development.

Considering that Apertium is evolving as an independent MT solution, one open question is how other systems, in particular statistical machine translation (SMT) systems, could possibly benefit from Apertium freely available linguistic data. Some researchers have explored hybrid approaches to leverage these Apertium data. Some use the whole Apertium system as a wrapper, managing translations from other engines, like Sánchez-Martínez et al. (2009). Others like Sánchez-Cartagena et al. (2011) use transfer rules and dictionary information to generate translation hypotheses.

In this work we explore a strategy to exploit some of the information stored in the Apertium dictionaries and the part-of-speech tagger of Apertium. On the one hand, the morphological tags deliverd by the part-of-speech tagger of Apertium will be used to decide when to lowercase the first word of a sentence and to split sentences. On the other hand, Apertium dictionaries have translation-oriented multiword-units (MWUs), coded by linguists. We will use MWU information both in the training corpus and during the translation process, as we expect it to provide better alignments (and therefore, better translation quality) during training. The reason which leads us to expect this improvement is that words are particularized using their context and therefore this avoids frequency interferences between words when they appear in MWUs and the individual words that form those MWUs.

During this work we found the experimental fact that using word division determined by linguistically motivated, human encoded data (in our case, from the Apertium platform), can improve consistently SMT quality in all of our experiments.

Factored translation models (Koehn and Hoang, 2007) use equivalent linguistic data without multi-word translation units. It uses however a different strategy based on training and using separate translation models for lemmas and for parts of speech.

We also include in our proposal a way to manage fixed translations, based on Apertium morphology modules, that allow a more stable handling of numbers, some punctuation marks and fixed translations of proper nouns during translation.

In the following sections we detail the specifics of the tokenizing method proposed (section 2), the experiments carried out to evaluate its performance, the results (section 3) and, finally, conclusions and a description of future work (section 4).

2 Enriching text tokenization with linguistic data

2.1 Baseline: the default tokenization of text in Moses

The default training in Moses is based on texts tokenized in a very crude way: separating words and punctuation, and possibly taking into account some (language-dependent) abbreviations that contain punctuation marks inside. The input texts are lowercased by default and then a *recaser* is trained to attempt to restore the original casing (capitalization) of the text, taking into account the diferent casing in both languages considered in the translation. In the figure 1, in the baseline row, we see how this work is done with an example that will be used throughout this paper.

An alternative way of tokenizing text in Moses, not considered in this work, is using *truecasing*. Truecasing consists in retaining the original case of words but lowercasing only first words of sentences if their most frequent form in texts is lowercased.

2.2 Adding Apertium-based tokenization

We use MWUs for sequences of words that are worth to be considered together rather than separately for a particular purpose. In a similar way, we define multiword translation units (MTUs) as translation units that have MWUs in at least on one of the two languages involved.

The components of Apertium being used for this purpose are the morphological analyzer and the part-of-speech tagger. The morphological analyzer is a module based on finite-state technology; it provides all the possible morphological analyses for a given lexical unit, while also tokenizing the input according to the definition of these lexical units in dictionaries (left-to-right, longest-match, tokenize-as-you-analyse strategy). The part-of-speech tagger uses hidden Markov model (HMM) techniques to determine the best part-of-speech of a given word in its context.

The morphological analyzer of Apertium (`lt-proc -a`) marks the start each lexical unit it recognizes with a circumflex sign (`"^"`) and its end with a dollar sign (`"$"`). MWUs are marked together as if it were regular words, including the blank characters found between individual words. The surface form comes first, and then, the different analyses are written, with the bar character (`"/"`) used as a separator.

The part-of-speech tagger (`apertium-tagger -g -p`) uses a suitably-trained HMM to select the most likely part-of-speech tag (and therefore the most likely analysis) among those provided by the morphological analyser.

In order to allow Moses to use this segmentation, blanks inside MWUs are replaced with tilde (`"~"`) characters. The aim of this preprocessing is to reduce the probability of possible relationships between words identified by automated text alignment process that have not been taken into account in order to properly align a bitext when training a SMT model.

Each multiword unit, in this experiment, is not intended to have a matching multiword in the other side. Multiword units are treated as regular words and the alignment process will decide which correspondence applies for every sentence having the perspective that the SMT engine will decide the most likely translation in each case.

Figure 1 shows the result of tokenizing text using Apertium part-of-speech is shown. Particular part-of-speech tags are used in order to decide whether the first word of a sentence has to be lowercased or not, rather than using the frequency of the word in the text as it is done in truecasing.

Figure 2 shows an example of a parallel sentence of the kind used to train the system. The co-occurrence, in this particular case, of Apertium MWUs gives an idea of how can the specifics of tokenization can affect alignment quality and, therefore the translation quality obtained from the trained models.

Original	A few months ago, the new CEO of Air Berlin, Stephane Richard, announced that the company will base 30% of bonuses on the ”happiness” of their staff. .
Baseline	a few months ago , the new ceo of air berlin , stephane richard , announced that the company will base 30 % of bonuses on the " happiness " of their staff .
Combined	**a˜few** months ago , the new **CEO** of **Air˜Berlin , Stephane Richard ,** announced that the company will base **_NUM2_%** of bonuses on the ˜" happiness "˜ of their staff .

Figure 1: Combined tokenization using Apertium linguistic data. Note the tilde marks grouping multi-word units, and the preservation of the original casing in “Air Berlin”.

	English	**Spanish**
Baseline	we european socialists are in favour of a market economy with a social purpose .	nosotros , los socialistas europeos , estamos a favor de una economía de mercado con fines sociales .
Combined	we European Socialists are **in˜favour** of a **market˜economy** with a social purpose .	nosotros , los socialistas europeos , estamos **a˜favor** de una **economía˜de˜mercado** con fines sociales .

Figure 2: Example of co-occurrence of multiwords in both sides of the training corpus.

2.3 Number preprocessing

Statistical machine translators treat numbers as it were usual words. In general, users do not expect numbers to be deeply transformed as a result of MT processing. They might however require some minor transformations such as those affecting the use of punctuation. For example, the English number 2,345.45 should be written in Spanish as 2.345,45 (with dot and comma reversed), following the conventions of the language.

SMT systems do not deal very well with numbers. Numbers are treated like different words and stored in phrase tables and language models. This representation is not suitable since numbers constitute a regular language that can be perfectly characterized by a regular expression. This fact leads to an enormous variability in the training corpora of SMT systems regardless of the number nature and meaning.

For example, years are generally 2 or 4-digit numbers and temperatures 1, 2 or 3-digit numbers, depending on the particular context of a text. We use a transformation mechanism that tries to keep these facts in mind in order to reduce text complexity while mantaining these differences. Numeric sequences are therefore transformed in a input text into the following entities:

- _NUMZ_: represents the 0 number only when occurs as a 1-digit number.
- _NUMI_: represents the 1 number only when occurs as a 1-digit number.
- _NUM[0-9]+_: represents the rest of sequences of numbers, while the number after _NUM indicates the number of digits found.

The specific treatment of numbers 0 and 1 is done in order to reflect the fact that, not only, but these two numbers are treated specially depending on the language. For example, number 1 appears in singular linguistic contexts and make sense to differentiate it from other 1-digit numbers, while 0 is usually followed by plural forms.

A mapping between these entities and the original numbers is stored in a way that it can be retrieved after running the SMT systems to restore the desired format of the numbers in suitable positions.

An example of this rewriting process is shown in table 1.

Original	Transformed
2004	_NUM4_
2,004	_NUM1_,_NUM3_
0.34	_NUMZ_._NUM2_
3 000	_NUM1_ _NUM3_
0.1	_NUMZ_._NUMI_

Table 1: Some examples of number rewriting.

2.4 Fixed translations

Some inconsistencies appearing in the translations generated by Moses are related to missing or altered proper nouns, punctuation marks, numbers and other kind of fixed-translation patterns or fixed-translation entities.

Taking advantage of the Moses XML Markup feature to indicate fixed translations to the decoder and of the information of the Apertium dictionaries, a module to preprocess patterns and rules has been set up to be used during the translation process.

The main advantage of operating this way is the possibility of having consistent translations of well known expressions without requiring large amount of data containing these expressions even when these expressions are not frequent in the training corpora, and also to fix frequent multiword translations determined by linguists.

The module uses Apertium-like dictionaries with both language-dependent and independent data to mark fixed translations and expressions in the source language that will be forced in the target language. These dictionaries are compiled by the Apertium `lt-comp` program to be turned into finite-state-transducers which can be processed at high-speed by the Apertium `lt-proc` program.

A typical fixed translation dictionary contains:

- list of persons, places and entities that do not have to be translated (*Bush*, *Colorado*, *France Telecom*)
- regular expressions for punctuation marks (exclamation marks, quotes, brackets, etc.)
- regular expressions for numerical entities (dates, amounts of money, decimals, percentages, etc.)
- special characters (currency symbols, ampersands, hash marks, etc.)
- regular expressions for URLs
- regular expressions for e-mail addresses

```
<e>
  <p>
    <l>France<b/>Telecom</l>
    <r>France<b/>Telecom</r>
  </p>
</e>
```

Figure 3: Example of an entry in the fixed translation dictionary to avoid *France* to be translated in isolation from English to Spanish or from French to Spanish when it is part of *France Telecom*.

An entry in the dictionary looks like in the example in figure 3.

An example of fixed translation and the way it is marked in the source language text is provided in figure 4. In this case, *Air Berlin* was marked both as a MWU during the tokenization process according to Apertium morphological analysis and as a fixed translation according to the fixed translation dictionary described in this section. Depending on the training corpora, if *Air Berlin* were not a fixed translation, Moses could try to translate *Air* into the target language.

3 Experiment and results

In order to evaluate the performance of the new tokenizing method compared to the default tokenizer in Moses, the following experiment has been performed:

- **Baseline**: four baseline systems for English–Spanish, Spanish–English, French–Spanish, Spanish–French have been trained using the WMT11[1] baseline system data (Europarl only), instructions and parameters.

[1] http://www.statmt.org/wmt11/

```
<fixed-translation translation="Air~Berlin">Air~Berlin</fixed-translation>
```

Figure 4: Air Berlin is marked using Moses XML Markup feature according to Apertium morphological analysis and fixed translations dictionary.

- **Combined**: systems for the same four language pairs have been trained following the same procedure but with by tokenizing the same training data using the method described in this paper.

The experiment have been carried out using WMT11 baseline system data, instructions and parameters to train baseline models, and then replacing the data by the same data tokenized in the way exposed in this paper. The test set corresponds also to WMT11 task, 2500 sentences from NewsCommentary 2010.

Results for BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and NIST (Doddington, 2002) scores are presented in tables 2, 3 and 4, respectively. The figures show that, for all four systems, those trained with the new tokenizer method outperform the baseline systems. In all cases, at least 0.02 BLEU points are gained in the combined systems. METEOR shows a better improvement in the French–Spanish system than in the other three, which show improvements between 0.01 and 0.02. NIST scores also show a general improvement for all combined systems.

Translator	Baseline	Combined
fr → es	0.27	0.29
es → fr	0.25	0.27
en → es	0.22	0.24
es → en	0.22	0.24

Table 2: BLEU scores for the experiments

This linguistic-motivated tokenizing method proves to be useful to increase the final quality of the translation by making it more consistent with respect to casing, punctuation and other fixed patterns.

Translator	Baseline	Combined
fr → es	0.42	0.45
es → fr	0.40	0.41
en → es	0.38	0.40
es → en	0.24	0.26

Table 3: METEOR scores for the experiments

Translator	Baseline	Combined
fr → es	7.22	7.55
es → fr	6.90	7.21
en → es	6.49	6.95
es → en	6.62	7.02

Table 4: NIST scores for the experiments

4 Conclusions and future work

A new linguistic-based tokenization method to preprocess the texts that are used to train a Moses SMT system has been presented in this paper; the method uses the linguistic data freely available in the Apertium project. This way of combining RBMT resources with SMT has shown to improve SMT results consistently as measured with the standard metrics. The availability of data in the Apertium platform and from other sources makes possible to apply this method to a variety of languages. Additionally, this processing does not conflict with other techniques that may be applied to further improve SMT quality.

In the future, we will continue to explore ways of integration between Apertium and Moses at a deeper level, in order to make this first Apertium–Moses combined system more stable and reliable, in order to obtain a significant improvement in the output translation quality.

Some of the improvements could come from using linguistic information to reorder some sequences of parts of speech between languages with large structural differences or from filtering training corpora using dictionary equivalences in order to remove very unfrequent translations.

5 Acknowledgments

We wish to thank Gema Ramírez Sánchez and professor Mikel L. Forcada for their support during the writing of this paper.

References

Francis M. Tyers and Mikel L. Forcada and Gema Ramírez-Sánchez. 2009. *The Apertium machine translation platform: Five years on*. Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation, pages 3–10, Alacant, Spain.

Felipe Sánchez-Martínez and Mikel L. Forcada and Andy Way. 2009. *Hybrid rule-based – example-based MT: Feeding Apertium with sub-sentential translation units*. Proceedings of the 3rd Workshop on Example-Based Machine Translation, pages 11–18, Dublin, Ireland.

Víctor M. Sánchez-Cartagena and Felipe Sánchez-Martínez and Juan A. Pérez-Ortiz. 2011. *The Universitat d'Alacant hybrid machine translation system for WMT2011*. Proceedings of the 6th Workshop on Statistical Machine Translation, pages 456–463, Edinburgh, United Kingdom.

Philipp Koehn and Hieu Hoang. 2007. *Factored Translation Models*. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 868–876., Praque, Czech Republic.

Philipp Koehn and Hieu Hoang and Alexandra Birch and Chris Callison-Burch and Marcello Federico and Nicola Bertoldi and Brooke Cowan and Wade Shen and Christine Moran and Richard Zens and Chris Dyer and Ondrej Bojar and Alexandra Constantin and Evan Herbst. 2007. *Moses: Open source toolkit for statistical machine translation*. Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180, Prague, Czech Republic.

Kishore Papineni and Salim Roukos and Todd Ward and Wei-Jing Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan.

George Doddington. 2002. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*. Proceedings of the 2nd international conference on Human Language Technology Research, pages 138–145, San Diego, California, USA.

Reordering by Parsing

Jakob Elming and **Martin Haulrich**
Department of International Language Studies and Computational Linguistics
Copenhagen Business School
Copenhagen, Denmark
{jel,mwh}.isv@cbs.dk

Abstract

We present a new discriminative reordering model for statistical machine translation. The model employs a standard data-driven dependency parser to predict reorderings based on syntactic information. This is made possible through the introduction of a reordering structure, which is a word alignment structure where the target word order is transposed onto the source sentence as a path. The approach is integrated in a phrase-based system. Experiments show a large increase in long distance reorderings. Both automatic and human evaluations show substantial increases in translation quality on an English to German task.

1 Introduction

Handling word order differences between languages is one of the main challenges of statistical machine translation (SMT) today. These differences are often most naturally handled at a syntactic level, since they pertain to entire syntactic constituents.

We present a syntactically motivated discriminative reordering model. The model exploits a *reordering structure*, which is a word alignment where the target sentence is unknown. This structure allows us to treat the reordering problem as a dependency parsing problem. We use a standard data-driven dependency parser to predict reorderings instead of dependencies. This is integrated into a phrase-based SMT (PSMT) framework (Koehn et al., 2003).

2 Reordering Structure

Word alignments are often used to display the relation between a translation and its source by linking up equivalent words. Here we transpose the word alignment information to a representation over a single sentence. This can be done by representing the corresponding order of the words of the opposite sentences as a path over the words of the current sentence. In this work we will focus on transposing the word alignment onto the source sentence by annotating it with the order in which the aligned target words occur. This is done in the form of a reordering structure, which is a word alignment, where the target sentence is unknown. The idea of a reordering structure is similar to the underlying concept of *source position target order* (Elming, 2008) or *visit sequence* (Ge, 2010), but the extraction algorithm and conceptual representation is different.

Figure 1 gives a simple example of how a reordering structure is created. The figure contains a source and a target sentence with a word alignment in between and the corresponding reordering structure on the source sentence. The numbers are merely used to explain the correlation between links and edges. They are not part of the structure. The reordering structure is created by following the target words from left to right. The first target word is linked to the first source word, and the graph therefore starts by going to the first word. Then the second target word links to the third source word, so the graph proceeds to this word, and so on. The resulting representation consists of the source sentence annotated with a reordering structure which reflects the word order of the corresponding target sentence.

One requirement for the structure is that all source words partake in an edge. The role of null-linked source words in the structure cannot be uniquely determined from a word alignment. It can either be inserted after the previous word or before the following word in the structure.

We employ a *syntactic closeness measure* to decide between left and right attachment. The distance from the null-linked word up to the first common

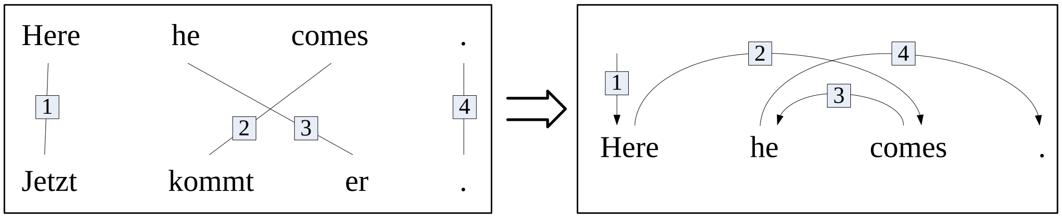


Figure 1: Example of a reordering structure and its underlying word alignment. Numbers are added for explanatory reasons indicating correlation between links and edges. They are not part of the structure.

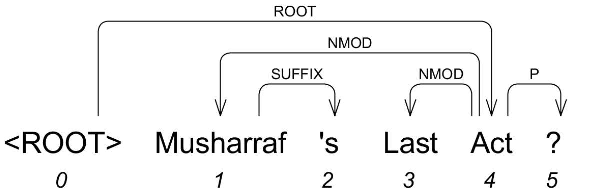


Figure 2: Syntactic dependency structure illustrating syntactic closeness.

node with the left and right neighbor is measured as the number of edges passed on the way. If this is the same for both neighbors, we choose the neighbor with the shortest distance up to the common node. If this is also the same, we attach right. As an example, if we assume *Last* in figure 2 is null-linked, we need to decide whether to insert relative to the left word *'s* or right word *Act*. Common ancestor node is 4 with both neighbors, so distance up from *Last* is the same. We therefore rely on distance up from neighbors, which is 2 passed edges for *'s* and 0 for *Act*. *Act* is therefore syntactically closer, and we attach right. Algorithm 1 illustrates the construction of the reordering structure formally.

```
$sourcePosition\ previous = rootPosition$;
**foreach** *targetPosition* $t = 0, ..., T$ **do**
    **foreach** *sourcePosition* $s$ *linkedTo* $t$ **do**
        add edge from $previous$ to $s$;
        $previous \leftarrow s$;
    **end**
**end**
**foreach** *sourcePosition* $s \in nullLinked$ **do**
    **if** $(s-1)$ *is syntactically closest to* $s$ **then**
        insert $s$ after $(s-1)$;
    **else**
        insert $s$ before $(s+1)$;
    **end**
**end**
```

Algorithm 1: Algorithm for creating reordering structure from word aligned sentence positions.

The measure is linguistically motivated. The common ancestor defines the smallest spanning constituent containing both words. The shorter the path up, the smaller the span, and the syntactically closer the words. If we simply measured the total path length, we might get fooled by a long path down. An example is a noun preceded by a preposition and followed by a relative clause. Here, the noun itself is the common ancestor with the relative pronoun, i.e. it has a 0 distance up, but the down distance may be long. The total path would not classify the noun closest to the relative clause, since the distance to the preceding preposition is 1 up and 0 down.

The advantage of the reordering structure representation is that it is a word alignment representation without explicit reference to the target sentence. Reordering in machine translation can be viewed as a similar challenge, where we want to find the word alignment knowing only the source sentence. The reordering structure provides us with a focus on this problem, since it refers only to the source sentence and therefore may be predicted from this.

The relation between the reordering structure and the word alignment is not reversible. Whereas all reordering structures correspond to a unique word alignment, the reverse is not the case. Certain word alignments are not representable by a reordering structure. In particular, structures where a source word is linked to target words that are separated by target words linking to a different source word cannot be represented without introducing recursion into the structure as exemplified by figure 3. As a consequence, the structure becomes ambiguous, since a single word would have more out-going edges that could be traversed in different orders.

Crego & Yvon (2009) face similar challenges

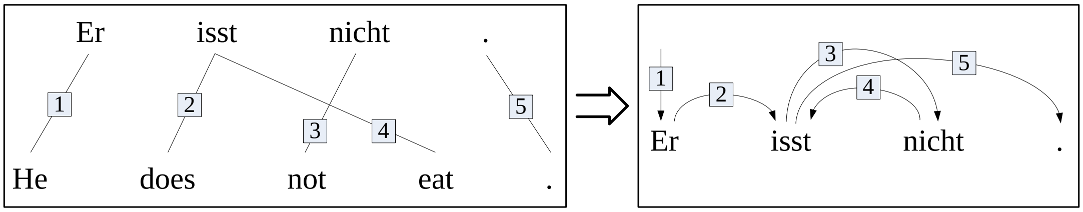


Figure 3: Example of a word alignment that cannot be represented *unrecursively* in a reordering structure. The language pair is here reversed, since we did not find these structures in the direction we are working with.

when monotonizing the parallel sentences. They handle the problem by making a source word clone for each discontinuous unit it is linked to. We do not adopt this approach here, since these structures are not a major concern with PSMT. PSMT has two means for handling word order differences between languages; phrase-internal reordering, where the equivalent words of a phrase pair appear in different orders, and phrase-external reordering, where the phrases are combined in a different order than they appeared in the source sentence. Only phrase-internal reordering can lead to this problematic word alignment in application, since a single source word token cannot participate in more phrases in the same translation. Since phrase-internal reordering is very reliable, the main purpose of the reordering model is to guide the phrase-external reordering, which will not produce these link constellations.

3 Reordering Structure Modelling

In this work, we will pursue the idea that the reordering structure is conceptually similar to an unlabeled syntactic dependency structure. We therefore use the MSTParser (McDonald et al., 2005), a state-of-the-art data-driven dependency parser, to model the reordering structure.

The basic idea is that the parser predicts the most favorable word alignment to the target sentence based on the source sentence. These predictions are made before translation and passed to the decoder. The level of information included in the reordering structure model therefore only depends on what features we are able to design for the parser, and is fully independent of the PSMT system.

The default output of the parser is the most probable reordering structure given its model. This is too restrictive for our purpose. The model would often not be relevant, if it expected a single word order during translation. Especially for longer sentences it would be unlikely to get this exact word order.

		To position			
		1	2	3	4
From position	0	**-0.16**	-1.03	-1.21	-1.39
	1		0.50	**1.01**	0.51
	2	0.91		1.16	**1.48**
	3	1.22	**1.34**		1.14
	4	0.23	0.12	-0.07	

Table 1: Illustration of the edge scores that the parser provides for the English sentence in figure 1. The highest scoring structure in bold.

One of the characteristics of first-order MST parsing is that the score of each edge is independent of the rest of the structure. The parser therefore calculates a matrix of scores for each possible edge in the sentence before searching for the most probable combined structure. We exploit this behavior by emitting the matrix of edge scores instead of the best structure. This way, we can provide the decoder with scores for each possible reordering it can produce.

Table 1 gives an example of such an edge score matrix with the scores that the parser provided for the English sentence in figure 1. As an example, an edge from word 2 to word 4 has a cost of 1.48. Higher scores are better. Position 0 is the root position, which can only have out-going edges. The bold scores mark the most probable structure, which is the structure represented in figure 1.

4 Integration in PSMT

As described in the previous section, the decoder recieves an edge score matrix in addition to the source sentence. This extra information is only used by a *word alignment scoring model*. This model returns a score each time a phrase is added to a translation

System	Lexical reordering	Tune *newstest2008*	Development *newstest2009*	Test *newstest2010*
Baseline	-	13.69	13.20	14.18
	+	13.98	13.74	14.80
Reordering Structure	-	14.04	13.76	14.69
	+	14.48	14.11	14.93
Oracle Reordering Structure	-	16.99	16.34	17.66
	+	17.17	16.67	18.06

Table 2: BLEU evaluation for the systems on different data sets.

hypothesis. Since the model returns a single score, it only introduces one additional parameter to system optimization. The score is calculated as the sum of scores for alignment links of adjacent target word positions within the phrase:

$$s_{wa} = \sum_{i=1}^{n} s(a_{i-1}, a_i) \tag{1}$$

where i is the target word position in a sentence of length n, a_i is the source word position it links to, and $s(a_{i-1}, a_i)$ is the score for target word positions $i-1$ and i being aligned to source word positions a_{i-1} and a_i respectively. That is, the score of an edge going from a_{i-1} to a_i.

The scoring process is exemplified in figure 4. The left box illustrates the end of a translation hypothesis, and the right box illustrates the new phrase being added to this hypothesis. Only the reordering structure being scored at this stage is shown above the translation. Again we indicate the relationship between the word alignment and the edges using indexes. The index 0 shows the final link of the translation hypothesis, which decides where the new phrase links up. That is, the score for adding a new phrase is the sum of the score for connecting to the previous phrase (edge 1) and the scores for the phrase-internal edges (edges 2 and 3). These phrase-internal edges can be computed at phrase retrieval to save computation.

5 Experiments

Experiments were conducted from English to German, a language pair which exhibits substantial word order challenges.

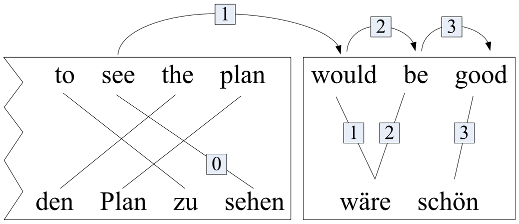


Figure 4: Illustration of the scoring done by the word alignment scoring model when extending the translation hypothesis with a new phrase.

5.1 Data

We use the English-German data from the Workshop on Statistical Machine Translation 2011 (WMT11)[1]. This consists of 3.4/3.3 million words of parallel news data, 46.0/43.7 million words of parallel Europarl data, and 309 million monolingual words of europarl and news. We only use unique sentences from the monolingual data. We use *newstest2008* for tuning, *newstest2009* for development, and *newstest2010* for testing.

5.2 Reordering Structure Model Setup

The reordering structure model is created with the MSTParser[2], a dependency parser based on online discriminative learning. Since the reordering structure will contain many crossing edges, it is necessary to use non-projective parsing. There are no algorithmic modification to the parser. The only modification we make is that we make it emit the edge score matrix for each sentence that it parses. We only train the model on 25,000 sentences from the parallel news data to keep down computational costs. The English side of this subcorpus is dependency parsed using an MSTParser trained on the

[1] http://www.statmt.org/wmt11/translation-task.html

[2] http://sourceforge.net/projects/mstparser/

Penn Treebank converted to dependency structures, and grow-diag-final-and word alignments from creating the PSMT system are used to extract reordering structures in CoNLL format. The dependency parse is used to connect null-linked source words as described in section 2, and it provides word form, part-of-speech tag, and dependency relation features for the reordering structure parser. We did not do extensive feature selection for the reordering structure model, but excluding either of the three information levels decreased performance on a translation task.

5.3 PSMT Setup

All our PSMT systems are created with the Moses toolkit (Koehn et al., 2007). We use the baseline system from WMT11[3] as our baseline with the small modifications that we use truecasing instead of lowercasing and recasing, and allow training sentences of up to 80 words. For our reordering experiments, we expand the baseline Moses system with the word alignment scoring model described in section 4. This is the only change to the baseline system. The baseline system got the best results with a distortion limit of 10, which we used for all experiments. The phrase table and the lexical reordering model is trained on the union of all parallel data with a max phrase length of 7, and the 5-gram language model is trained on the entire monolingual data set.

6 Results

6.1 Automatic Evaluation

Table 2 shows the results from automatic evaluation using the BLEU metric (Papineni et al., 2002). We report on the performance of the baseline and the reordering structure system with and without the lexical reordering model switched on. We use bootstrapping[4] to test the significance of the results (Zhang et al., 2004). For all the data sets, the reordering structure system significantly outperforms the corresponding baseline system.

An interesting observation is that adding either the lexical reordering model or the reordering structure model to the baseline brings an improvement, and adding both improves performance even further. This indicates that the two models target different areas of reordering, and therefore they do not even each other out. Instead we see a cumulative effect where performance is increased even further.

	All edges	Non-monotone edges
Tune	69.47	11.81
Development	69.03	11.88
Test	71.32	14.51

Table 3: Unlabeled attachment scores for the reordering structure model on the data sets.

The final system represented in table 2 called Oracle Reordering Structure gives an indication of the performance that is attainable if the predictions of the reordering structure model are improved. Here the gold standard reordering structure was added as a feature, so the parser obtained a 100% unlabeled attachment score on the data sets. The idea of this system is to see how much there is to gain if we have a perfect reordering structure model. However, the gold standard builds on erroneous automatic word alignments, which means that the "correct" structure may mislead the translation. Also this is the oracle best structure, not the oracle best edge score matrix, which is what is actually used by the system.

Table 3 shows the unlabeled attachment scores for the basic reordering structure model on the data sets. The scores are computed based on the most probable parse for each sentence, and they are reported for all edges and for non-monotone edges, i.e. edges going anywhere else than to the right neighboring word. These non-monotone edges are the most interesting edges, since they represent the reordering, and the prediction of these is very poor. We therefore expect that a fair part of the gain indicated by the Oracle Reordering Structure system is attainable through improvement of the reordering structure model.

6.2 Human Evaluation

In addition to the automatic evaluation, we also perform a small human evaluation using sentence translation ranking (Callison-Burch et al., 2010). We have two native German speakers rank the translations from the baseline system and the reordering structure system relative to each other. We evaluate on the first 100 sentences of the test corpus (*new-*

[3] See a detailed desciption at http://www.statmt.org/wmt11/baseline.html

[4] http://projectile.sv.cmu.edu/research/public/tools/bootStrap/tutorial.htm

	Equally good	Baseline best	RS best
Evaluator 1	50	17	33
Evaluator 2	34	20	46
Average	42.0	18.5	39.5

Table 4: Human evaluation comparing the baseline and the reordering structure (RS) system on the first 100 sentences of the test set.

stest2010). On this subset, the baseline system gets a BLEU score of 10.55, and the reordering structure system gets 10.70.

The evaluators are presented with the source sentence and the two translations in randomized order. They are told to rank the systems from best to worst. Ties are allowed. The evaluators agreed on their judgements in 67 of the 100 sentences. Compared to an expected chance agreement of 1/3, the kappa coefficient is 0.505, which is much in line with findings from WMT10 (Callison-Burch et al., 2010).

Table 4 shows the results from the human evaluations. The translations from the reordering structure system were chosen as better than the baseline system more than twice as often as the reverse. This indicates that the reorderings introduced by the RS system may improve translation quality more than what the BLEU scores reflect. It has previously been reported that BLEU can be insensitive to word order improvements (Callison-Burch et al., 2007).

7 Analysis

An interesting aspect of the effect of the reordering structure model is the amount of word order differences it leads to. This information can be extracted from the word alignment information produced during a translation. Figure 5 shows the amount of reordering created by the baseline and reordering structure model systems on the development data. The figure shows that the reordering structure system introduces a lot more long distance reordering to the translation than the baseline systems. With lexical reordering on, it produces more than twice as many reorderings with a distance of 4 words, and more than 4 times as many reorderings with a distance of 8 words.

Here we also see a cumulative effect of combining the reordering structure model with the lexical reordering model. Together with the baseline system, the lexical reordering model does not introduce much long distance reordering, but combined with the reordering structure model the amount of long distance reordering gets boosted, also compared to the reordering structure model by itself.

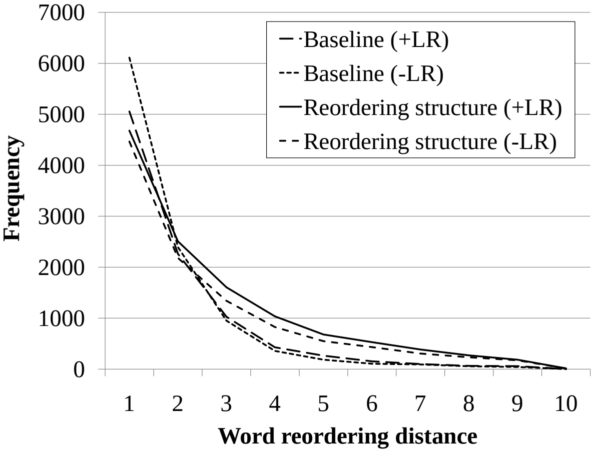


Figure 5: The amount of *non-monotone* decoding done by the baseline and reordering structure systems on the development data. More specifically, the number of target words representing a given reordering distance. LR specifies whether lexical reordering is in use.

8 Related Work

In recent years, the integration of syntactic knowledge into statistical machine translation has received much attention. The main motivation for this has been the need for better reordering of the words during translation. In a framework like synchronous context-free grammars (SCFGs) syntax is incorporated either on the source side (Liu et al., 2006), the target side (Galley et al., 2004), or both sides (Liu et al., 2009), and reordering is handled through the rules that constitute the building blocks for the translation. Such approaches have proven successful especially for language pairs which exhibit much nonlocal reordering (Zollmann et al., 2008; Birch et al., 2009). The hard constraints within the formalisms of these frameworks may however be too restrictive to handle frequently occuring aspects of parallel languages (Wellington et al., 2006; Søgaard and Kuhn, 2009; Galley and Manning, 2010).

In order to avoid such hard constraints introduced by the formalism, we place reordering information in the model to motivate certain word orders rather than prohibit others. That is, we create a reordering model that scores translation in parallel to other scoring models. This is much in line with Chiang (2010), who place syntactic correspondence information in the model as a soft constraint, but their approach is heavily tied to the SCFG framework, whereas our approach is framework independent.

A lot of work has been done on reordering in PSMT. The original approach deterred reordering by applying a *distortion penalty* for each word that is moved across (Koehn et al., 2003). Another approach is *lexicalized reordering*, which conditions the probability of moving a phrase in a certain direction on the lexical content of the phrase (Tillmann, 2004; Koehn et al., 2005). A third approach is *pre-translation reordering*, which reorders the source words in an attempt to assimilate the word order of the target language prior to translation. This can be done by supplying the decoder with a single permutation (Xia and McCord, 2004; Collins et al., 2005; Habash, 2007; Xu et al., 2009) or multiple weighted permutations (Zhang et al., 2007; Li et al., 2007; Elming, 2008; Ge, 2010). The present approach relates to the pre-translation reordering approaches in that it tries to predict the target word order from source sentence syntax. However, in these previous approaches, the source words are reordered prior to translation. This is not done in the current approach – instead, we use a decoder-internal model for scoring all generated reorderings.

The approach utilizes syntactic dependency relations to predict reorderings. This has previously been suggested to provide a better basis for reordering in machine translation due to higher inter-lingual phrasal cohesion than phrase structure (Fox, 2002). Much previous work has included dependency structure information in an SMT system. Quirk et al. (2005) use a source side dependency structure in their treelet SMT system, which translates from subtrees to strings. Galley & Manning (2009) use a dependency parser in a phrase-based setup for assigning a dependency structure to the target side during translation. This allows for the integration of a dependency language model directly into the system. Gimpel & Smith (2009; 2011) treat translation as a monolingual dependency parsing problem, creating a dependency structure over the translation during decoding. No syntactic structure is created during decoding in our approach. Instead the dependency parser is used for the sole purpose of scoring the word order of the target sentence.

9 Conclusion and Future Work

We have introduced a new syntactically motivated discriminative reordering model. The model employs a standard data-driven dependency parser to predict reorderings. This is made possible by introducing a reordering structure. Within the framework of PSMT, we obtain substantial increases in translation quality both measured automatically and by human evaluators on an English to German task.

In the present work, we did very little feature selection and only provided word form, part-of-speech, and dependency relation information for the parser. In the future, we will experiment with additional features to improve the reordering structure model. In particular, we expect that more syntactic features will be beneficial. Also approaches such as second-order and stacked parsing may be helpful, since first-order parsing may be too weak to handle the complexities of the reordering structure. We also want to look closer at the features exploited by the standard MTSParser. These features are optimized to learn dependency structures, and they may not be optimal for learning the reordering structure.

One concern with the approach is that the model is trained against a gold standard which was extracted from automatic word alignment. This means that there will be a lot of noise in the training material. Also when training against the gold standard, all edges are considered equally important, but this may in fact not be the case for translation. Certain reorderings should always apply, while other may be stylistic and optional. A better way of training the model might be to train it as part of optimizing the PSMT system. This way, the system would be optimizing directly towards improving the word order of the translation. Due to the discriminative model's large number of weights, using a discriminative algorithm to optimize the system (Watanabe et al., 2007; Chiang et al., 2008) would be an interesting option. This could either be done by learning from only the data set used for tuning the PSMT system, or by taking the model trained in the present work as a point of departure and revising the weights in the context of optimizing a PSMT system. We expect to pursue this direction in future work.

References

A. Birch, P. Blunsom, and M. Osborne. 2009. A quantitative analysis of reordering phenomena. In *Proceedings of the WMT*.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the WMT*.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of WMT*.

D. Chiang, Y. Marton, and P. Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of EMNLP*.

D. Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of ACL*.

M. Collins, P. Koehn, and I. Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*.

J. M. Crego and F. Yvon. 2009. Gappy translation units under left-to-right smt decoding. In *Proceedings of EAMT*.

J. Elming. 2008. Syntactic reordering integrated with phrase-based smt. In *Proceedings of COLING*.

H. J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP*.

M. Galley and C. D. Manning. 2009. Quadratic-time dependency parsing for machine translation. In *Proceedings of ACL*.

M. Galley and C. D. Manning. 2010. Accurate non-hierarchical phrase-based translation. In *Proceedings of HLT-NAACL*.

M. Galley, M. Hopkins, K. Knight, and D. Marcu. 2004. What's in a translation rule? In *Proceedings of NAACL*, Boston, Massachusetts, USA.

N. Ge. 2010. A direct syntax-driven reordering model for phrase-based machine translation. In *Proceedings of HLT-NAACL*.

K. Gimpel and N. A. Smith. 2009. Feature-rich translation by quasi-synchronous lattice parsing. In *Proceedings of EMNLP*.

K. Gimpel and N. A. Smith. 2011. Quasi-synchronous phrase dependency grammars for machine translation. In *Proceedings of EMNLP*.

N. Habash. 2007. Syntactic preprocessing for statistical machine translation. In *Proceedings of MTSummit*.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*.

P. Koehn, A. Axelrod, A. Birch Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of IWSLT*.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL: Demo and Poster Sessions*, Prague, Czech Republic.

C.-H. Li, M. Li, D. Zhang, M. Li, M. Zhou, and Y. Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of ACL*, Prague, Czech Republic.

Y. Liu, Q. Liu, and S. Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING-ACL*.

Y. Liu, Y. Lü, and Q. Liu. 2009. Improving tree-to-tree translation with packed forests. In *Proceedings of ACL*.

R. T. McDonald, F. Pereira, K. Ribarov, and J. Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of HLT/EMNLP*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, Philadelphia, PA.

C. Quirk, A. Menezes, and C. Cherry. 2005. Dependency treelet translation: syntactically informed phrasal smt. In *Proceedings of ACL*.

A. Søgaard and J. Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *Proceedings of SSST*.

C. Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*.

T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proceedings of EMNLP-CoNLL*.

B. Wellington, S. Waxmonsky, and I. D. Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of ACL*.

F. Xia and M. McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of COLING*. COLING.

P. Xu, J. Kang, M. Ringgaard, and F. Och. 2009. Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of NAACL*.

Y. Zhang, S. Vogel, and A. Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of LREC*.

Y. Zhang, R. Zens, and H. Ney. 2007. Improved chunk-level reordering for statistical machine translation. In *Proceedings of IWSLT*, Trento, Italy.

A. Zollmann, A. Venugopal, F. J. Och, and J. M. Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of COLING*.

Comparing Corpus-based MT Approaches Using Restricted Resources

Monica Gavrila
Hamburg University
Vogt-Koelln Str 30, 22527
Hamburg, Germany
gavrila@informatik.uni-hamburg.de

Natalia Elita
Hamburg University
Vogt-Koelln Str 30, 22527
Hamburg, Germany
elita@informatik.uni-hamburg.de

Abstract

Machine translation (MT) plays an important role in multilingual communication. Dealing with natural language and a diversity of language-pairs, it is not always possible to have sufficient (linguistic) resources for a specific MT approach and a diversity of domains. In this paper we compare a statistical MT system with an example-based one and a hybrid system. For a better overview we include in our comparison also an on-line MT system. We considered for our experiments a small-sized domain-restricted corpus for Romanian and English, in both directions of translation. We also tested which impact part-of-speech information has on the translation results.

1 Introduction

Machine translation (MT) plays an important role in multilingual communication (especially in the World Wide Web environment) and is already an integrated part of current natural language processing (NLP) applications, such as content management systems (CMSs)[1].

Dealing with natural language and a diversity of language-pairs, it is not always possible to have enough (linguistic) resources for a specific MT approach and a large variety of domains. Therefore, we set out focus in this paper on corpus-based MT (CBMT) approaches using a small-size corpus for training. We use for our experiments English-Romanian as language-pair, in both directions of translation.

We present several comparisons between CBMT approaches, in different experimental settings:

- Comparing statistical MT (SMT), example-based MT (EBMT) and hybrid MT (EBMT-SMT) , when no additional linguistic information is added to the corpus. The question which appears is if hybrid systems can overtake the pure CBMT approaches.

- Comparing SMT and EBMT, when part-of-speech (POS) information is added to the data. Usually it is thought that additional linguistic information helps the translation process. The questions we set is what the influence is when small-sized data are involved and which the difference is between the two main CBMT approaches (SMT and EBMT).

For a better overview we compare our results with the ones of an on-line MT system.

Experiments with smaller data (approx. 2.6K sentences) have been presented in the literature in (Popovic and Ney, 2006) for Serbian-English. Comparisons between SMT and hybrid or EBMT approaches are presented in the literature, but usually larger data is used. The marker-based EBMT system described in (Way and Gough, 2005) outperformed the SMT system presented in the same paper. In (Smith and Clark, 2009) the hybrid EBMT-SMT system is outperformed by a Moses-based SMT system. SMT and EBMT approaches for Romanian

[1] For example in the ATLAS (Applied Technology for Language-Aided CMS) project (http://www.atlasproject.eu/).

an English are shown in (Ignat, 2009) and (Irimia, 2009), respectively.

The paper is organized as follows: after the short introduction we will present the MT systems employed. In Section 3 we describe the data used in the experiments and we give a brief description of Romanian. Section 4 shows the automatic evaluation results and their interpretation. The paper ends with conclusions and further work.

2 The MT Systems

In this section we present the CBMT systems used: a Moses-based SMT system (**Mb_SMT**), a pure EBMT system ($Lin-EBMT^{REC+}$) and a hybrid (EBMT-SMT) MT system (OpenMaTrEx). For comparison reasons we also translated our test dataset with an on-line MT system: Google translate.

2.1 The SMT System: Mb_SMT (A)

The pure SMT system (**Mb_SMT**) follows the description of the baseline architecture given for the EMNLP 2011 6th Workshop on SMT[2]. **Mb_SMT** uses Moses[3], an SMT system that allows the user to automatically train translation models for the language pair needed, considering that the user has the necessary parallel aligned corpus. More details about Moses can be found in (Koehn et al., 2007). We used in our experiments SRILM (Stolcke, 2002) for building the language model (LM) and GIZA++ (Och and Ney, 2003) for obtaining the word alignment information. We made two changes to the specifications of the SMT workshop: we left out the tuning step[4] and we built an LM of order 3, instead of 5[5].

2.2 The EBMT Systems: $Lin-EBMT^{REC+}$ (B)

The EBMT system in this paper ($Lin-EBMT^{REC+}$) has been developed at the University of Hamburg. It combines the linear EBMT approach with the template-based one – see (McTait, 2001) for the definitions of the EBMT approaches and templates. It is based on surface-forms and uses no linguistic resources, with the exception of the parallel aligned corpus. It contains all the three steps of an EBMT system[6]: matching, alignment and recombination. Before starting the translation, training and test data are pre-processed in the same way as in **Mb_SMT**, i.e. tokenization, lowercasing etc. In order to reduce the search space in the matching process, we use a word index. The matching procedure is an approach based on surface-forms, focusing in finding recursively the longest common substrings. If during the matching procedure the test sentence is found in the training corpus, its translation represents the output. Otherwise, the alignment and recombination steps are performed. The alignment information is extracted from the GIZA++ output of the **Mb_SMT** system. The longest TL aligned subsequences are used further in the recombination step, which is based on 2-gram information and word-order constraints. In $Lin-EBMT^{REC+}$ ideas from the template-based EBMT approach are incorporated in the recombination step, by extracting and imposing several types of word-order constraints. More information about the system, templates and how combinations of constraints influence the results is presented in (Gavrila, 2011).

2.3 The Hybrid System: OpenMaTrEx (C)

The hybrid EBMT-SMT system we used is OpenMaTrEx: a free open-source EBMT system based on the marker hypothesis. This hypothesis (Green, 1979) is a universal psycholinguistic constraint which states that natural languages are '*marked*' for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes.

OpenMaTrEx consists of a marker-driven chunker, several chunk aligners, and two engines: one is based on the simple proof-of-concept monotone recombinator (called Marclator[7]) and the other uses a Moses-based decoder (called MaTrEx[8]).

From the two modes (Marclator and MaTrEx)

[2] `www.statmt.org/wmt11/baseline.html`.

[3] `www.statmt.org/moses/`.

[4] Leaving out the tuning step is motivated by the size of the data in this paper and the results we obtained in experiments which are not the topic of this paper, when comparing SMT with and without tuning. Not all tests with tuning showed an improvement.

[5] The change has been motivated by results presented in (Rousu, 2008)

[6] The steps of an EBMT system are firstly described in (Nagao, 1984).

[7] `www.openmatrex.org/marclator/`.

[8] `www.sf.net/projects/mosesdecoder/`.

in which OpenMaTrEx can be run, we chose for this paper the hybrid MT architecture, the MaTrEx mode. In this mode the system wraps around the Moses statistical decoder, using a hybrid translation table containing marker-based chunks as well as statistically extracted phrase pairs. For our experiments we followed the training and translation steps as described in (Dandapat et al., 2010).

The markers for English have been already contained in OpenMaTrEx. They were derived from the Apertium English-Catalan dictionaries[9]. The markers for Romanian were created from scratch during the experiments presented in this paper. Morpho-syntactic specifications from MULTEXT-East[10] and Wikipedia[11] were used to derive the markers. There are currently 366 Romanian and 307 English makers. More about the Romanian markers can be found in (Gavrila and Elita, 2011).

2.4 The On-line System: Google Translate (D)

For comparison reasons we included an on-line MT System – Google Translate (`translate.google.com`) – in our experiments. The system is a free statistically-based machine translation service, provided by Google Inc. It translates a section of text, document or webpage, from one source language (SL) into the target language (TL). While Google Translate is nominated as an SMT system on `Wikipedia.org`, on the Google support webpage[12] it is only stated that it uses the "*state-of-the-art technology*", without reference to any specific MT approach.

3 The Corpus

For our experiments we used a domain restricted, small-sized corpus: RoGER. It is a parallel corpus, aligned at sentence level. It is domain-restricted, as the texts are from a users' manual of an electronic device.

The languages included in the development of this corpus are Romanian (ro), English (en), German and Russian. The corpus has been manually compiled and verified. It is not annotated and diacritics are ignored. The initial text was preprocessed by replacing numbers, websites and images with "*meta-notions*" as follows: numbers by *NUM*, pictures by *PICT* and websites by *WWWSIT*E. In order to simplify the translation process, some abbreviations were expanded.

The corpus contains 2333 sentences for each language. The average sentence length is eleven tokens for English, Romanian and German and nine for Russian. More statistical data about the corpus is presented in Table 1. Punctuation signs are considered as tokens. More about the RoGER corpus can be found in (Gavrila and Elita, 2006)

From the corpus, 133 sentences have been randomly extracted as the test data, the remaining 2200 sentences being used as training data.

We considered two experimental settings: one when no additional linguistic information is added to the corpus (Experimental setting I) and one when part-of-speech (POS) information is incorporated in the corpus (Experimental setting II). While former setting uses all four MT system mentioned in Section 2, the latter employs only **Mb_SMT** and $Lin-REC^{REC+}$. This happens as only these two MT systems work with the modified corpus, with no real impact on the algorithms or other resources. However, some POS information is indirectly included in the OpenMaTrEx algorithm in the form of markers.

For the Experimental setting II we annotated the corpus by means of the text processing web services described on the website of the Research Institute for Artificial Intelligence of the Romanian Academy (RACAI)[13]. The website provides on-line web services for text processing (such as tokenization, sentence splitting, POS Tagging and lemmatization), factored translation and language identification. More information about the web-services can be found in (Tufis et al., 2008). We concatenated the POS information to the word as **WORD+"*POS*"+POS**, where "*POS*" is a delimiter. A word with POS information (**WORD+"*POS*"+POS**) is considered during the translation as one token for the corpus-based MT ap-

[9] `www.apertium.org/?id=whatisapertium\&lang=en`.

[10] `nl.ijs.si/ME/V4/msd/html/msd-ro.html`.

[11] `ro.wikipedia.org/wiki/Parte_de_vorbire`.

[12] `translate.google.com/support/?hl=en`.

[13] `http://www.racai.ro/webservices/TextProcessing.aspx` - last accessed on June 27th, 2011.

Feature	**English**	**Romanian**	**German**	**Russian**
No. tokens	26096	25850	27142	22383
Voc.* size	2012	3104	3031	3883
Voc. (*Word-frequency higher than two*)	1231	1575	1698	1904

Table 1: The RoGER corpus – Some statistics (**Voc.=vocabulary*).

proaches involved. From the information provided by the web services we only used one of the POS tags[14]

Statistical information on the data for Experimental setting I is shown in Table 2. The statistical information about the training and test data which contains POS information is presented in Table 3 (Experimental setting II).

Data SL	**No. of words**	**Voc. size**	**Average sentence length**
en-ro			
Training	27889	2367	12.68
Test	1613	522	12.13
ro-en			
Training	28946	3349	13.16
Test	1649	659	12.40

Table 2: RoGER statistics (Experimental setting I).

Data SL	**No. of words**	**Voc. size**	**Average sentence length**
en-ro			
Training	27816	2815	12.64
Test	1610	564	12.11
ro-en			
Training	28954	4133	13.16
Test	1651	735	12.41

Table 3: RoGER statistics when additional POS information is added (Experimental setting II).

3.1 Language Characteristics: Romanian

As English is the language mostly used in NLP, we will present several characteristics of Romanian in this subsection.

Romanian is a morphologically rich language, having less resources when compared with other European languages. It is an Eastern Romance language, with grammar and basic vocabulary closely related to those of its relatives (e.g. Italian, Spanish, French). It has been influenced by several other languages, such as the Slavic languages, Hungarian and Turkish. This influence is encountered especially at lexical level.

Among the language-specific characteristics induced by its Latin origin are the following: a 3-gender system, double negation and pronoun-elliptic sentences. Also, as in all Romance languages, Romanian verbs are highly inflected (according to person, number, tense, etc.) Another Latin element that has survived in Romanian while having disappeared from other Romance languages is the morphological case differentiation in nouns, albeit reduced from the original seven to only three forms (nominative/accusative, genitive/dative and vocative).

It is the only Romance language where definite articles are attached to the end of the noun or the adjective as enclitics, depending on the position of the adjective before or after the noun. This phenomenon is encountered in some Slavic languages (Bulgarian, Macedonian), in Scandinavian languages and in Albanian.

4 Experimental Results

We evaluated our translations using two automatic evaluation metrics based on n-grams: BLEU and NIST. Due to lack of data and further translation possibilities, the comparison with only one reference translation is considered in these experiments.

Although criticized, BLEU (bilingual evaluation understudy) is the score mostly used in the last years for MT evaluation. It measures the number of n-grams, of different lengths, of the system output that appear in a set of reference translations. More de-

[14] The C-TAG: The first tag after the lemma provided by the web services.

tails about BLEU can be found in (Papineni et al., 2002).

The NIST Score, described in (Doddington, 2002), is similar to the BLEU score in that it also uses n-gram co-occurrence precision. If BLEU considers a geometric mean of the n-gram precision, NIST calculates the arithmetic mean. Another difference is that n-gram precisions are weighted by the n-gram frequencies.

The evaluation scores for all four MT systems (Experimental setting I) are shown in Table 4. In this table several explanations are needed: **A** is **Mb_SMT**, **B** $Lin-EBMT^{REC+}$, **C** OpenMaTrEx and **D** Google translate.

Score	**A**	**D**	**C**	**B**
		en-ro		
BLEU	0.4386	0.4782	0.3934	0.3085
NIST	6.5599	6.9334	5.9725	5.5322
		ro-en		
BLEU	0.4765	0.5241	0.4428	0.3668
NIST	6.8022	7.4478	6.4124	6.2991

Table 4: Evaluation results for RoGER (no POS Information).

It can be seen that for all cases the pure SMT system is better than the hybrid system. The EBMT system is the last. The on-line MT system overtakes all MT systems we trained.

Table 5 shows how POS information influences the translation results of **Mb_SMT** (System **A**) and $Lin-EBMT^{REC+}$ (System **B**)

Score	**A**	**B**
	en-ro	
BLEU	0.3879	0.2916
NIST	5.8047	5.0893
	ro-en	
BLEU	0.4618	0.3559
NIST	6.3533	6.0039

Table 5: Evaluation results for RoGER (additional POS information).

A comparison between the results of the two settings (with and without additional POS in the corpus) is shown in Figure 1.

For this specific data the results which contain POS information are lower than the ones without additional information. There are two reasons for these results: either POS information is affecting negatively the translations or the automatic scores cannot capture the improvement. Therefore, we should manually analyze part of the results. The negative impact can be due to incorrect results of the web-services (incorrect POS attached) or increase of data sparseness, which has a direct impact on the statistical approaches and the word alignment.

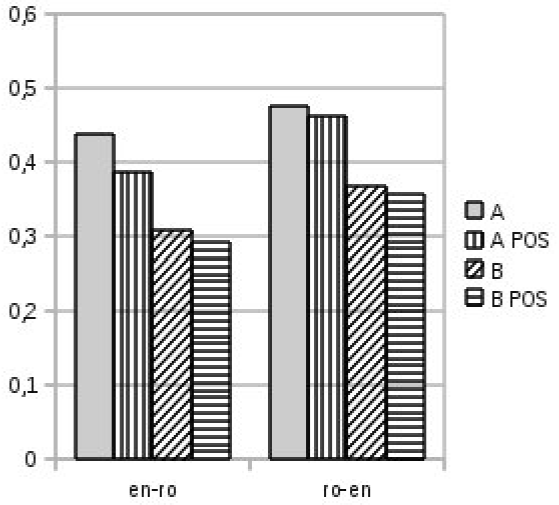


Figure 1: Comparison of the Evaluation Results

For a better overview on the results we compared the tokens[15] of the translations with those in the references. The results are shown in Table 6 in which "*Common tokens*" (CT) are tokens which the reference and the translation have in common and "*Ordered common tokens*" (O.CT) are common tokens between the translation and its reference, which have the same order in both sentences.

For example, the following two sentences:
*I decided **to go** home **by** bus.*
*We **go to** the theater **by** car.*
have three "*common tokens*" (*to*, *go*, *by*) and two "*ordered common tokens*" (*go*, *by*).

The percentage values in Table 6 are calculated from the total number of tokens in the reference translation. The results for **Mb_SMT** are closer to the reference translation. Moreover, the use of POS information influences negatively the values.

We manually analyzed the results of **Mb_SMT**

[15] In this context token means word, number or punctuation sign.

Desc.	**Ref.**	**A**	**B**
		en-ro	
Total	495	490	466
CT	-	352 (71.11%)	302 (61.01%)
O. CT	-	343 (69.29%)	244 (49.29%)
		en-ro and POS	
Total	490	472	480
CT	-	273 (55.71%)	257 (52.45%)
O. CT	-	267 (54.49%)	211 (43.06%)

Table 6: Comparison between the translations and their references (Ref.=reference, Desc.=description).

and $Lin - EBMT^{REC+}$ from the point of view of adequacy[16] and fluency[17]. Although not fully relevant,as only one human evaluator was available, but still with possible impact on further research, the average results for adequacy and fluency are presented in Table 7. The evaluation scale for adequacy and fluency is the one described in (LDC, 2005):

Adequacy: 1=None, 2=Little, 3=Much, 4=Most, 5=All.

Fluency: 1=Incomprehensible, 2= Disfluent, 3=Non-native, 4=Good, 5=Flawless

Evaluation	**A**	**B**
	en-ro	
Adequacy	4.22	3.64
Fluency	4.08	3.44
	en-ro and POS	
Adequacy	4.1	3.66
Fluency	3.74	3.3

Table 7: System analysis: adequacy and fluency (average values).

These results confirm the automatic evaluation scores and previous analyses.

The test scenario was kept as realistic as possible. Therefore, we have not excluded test sentences already in the training corpus: common users do not analyze the texts before translating them. Next to tests sentences included in the training data, also out-of-vocabulary (OOV) words have a direct impact on the translation results. An overview of these two aspects in our data is shown in Table 8.

Corpus	**No. of OOV-Words (% from voc.* size)**	**Sentences in the corpus**
	en-ro	
Test	60 (11.49%)	37 (27.81%)
Test (POS)	74 (13.12%)	37 (27.81%)
	ro-en	
Test	84 (12.75%)	34 (25.56%)
Test POS	116 (15.78%)	34 (25.56%)

Table 8: Analysis of the test data sets (Experimental settings I and II) (**voc.=vocabulary*).

As expected, the number of OOV-words increases when POS information is included in the data. Also the number increases when Romanian is the source language. This happens due to the characteristics of the language.

5 Conclusions and Further Work

In this paper we presented several CBMT experiments with different approaches using a small-sized domain-restricted corpus.

Analyzing the results it can be concluded that not always additional linguistic information improves the MT results. Also combining different approaches does not always lead to better results. The training and test data themselves, the impact of additional information (such as increase of data sparseness) directly influence the translations. For under-resourced language-pairs or lower-resourced domains it can be enough just the use of a pure SMT system.

For a better understanding of the results further (manual) analysis is required. Moreover, we need to run more tests with different language-pairs and corpora. Some further results in this direction can be found in (Gavrila and Elita, 2011).

References

Dandapat, Sandipan and Mikel L. Forcada and Declan Groves and Sergio Penkale and John Tinsley and Andy Way. 2010 OpenMaTrEx: A Free/Open-Source

[16] Adequacy refers to the degree to which information present in the original is also communicated in the translation.

[17] Fluency refers to the degree to which the output is well formed according to the rules of the target language.

Marker-Driven Example-Based Machine Translation System *IceTAL'10*, pages 121–126.

Doddington, George. 2002 Automatic evaluation of machine translation quality using n-gram co-occurrence statistics *Proceedings of the second international conference on Human Language Technology Research*, 138–145, San Francisco, CA, USA Morgan Kaufmann Publishers Inc., San Diego, California.

Irimia, Elena. 2009 EBMT Experiments for the English-Romanian Language Pair *In Proceedings of the Recent Advances in Intelligent Information Systems*, 91–102, ISBN 978-83-60434-59-8.

Ignat, Camelia. 2009 Improving Statistical Alignment and Translation Using Highly Multilingual Corpora *PhD Thesis*, INSA - LGeco- LICIA, Strasbourg, France.

Gavrila, Monica and Natalia Elita. 2006 Roger - un corpus paralel aliniat *In Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române Workshop Proceedings*, 63–67 December, Publisher: Ed. Univ. Alexandru Ioan Cuza, ISBN: 978-973-703-208-9.

Gavrila, Monica. 2011 Constrained recombination in an example-based machine translation system *Proceedings of the EAMT-2011: the 15th Annual Conference of the European Association for Machine Translation*, May, Leuven, Belgium.

Gavrila, Monica and Natalia Elita. 2011 Experiments with Small-size Corpora in CBMT *Proceedings of RANLP Student Research Workshop* September, Hissar, Bulgaria.

Green, T.R.G. 1979 The necessity of syntax markers: Two experiments with artificial languages *Journal of Verbal Learning and Verbal Behavior*, Volume 18, Number 4, 481 – 496. ISSN 0022-5371.

Koehn, Philipp and Hieu Hoang and Alexandra Birch and Chris Callison-Burch and Marcello Federico and Nicola Bertoldi and Brooke Cowan and Wade Shen and Christine Moran and Richard Zens and Chris Dyer and Ondrej Bojar and Alexandra Constantin and Evan Herbst. 2007 Moses: Open Source Toolkit for Statistical Machine Translation *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, June, Prague, Czech Republic

Linguistic Data Consortium. 2005 Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5. Technical report `http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf`

McTait, Kevin. 2002 *Translation Pattern Extraction and Recombination for Example-Based Machine Translation* PhD Thesis, Center for Computational Linguistics, Department of Language Engineering, PhD Thesis, UMIST.

Nagao, Makoto. 1984 A Framework of a Mechanical Translation between Japanese and English by Analogy Principle *Proceedings of the international NATO symposium on Artificial and human intelligence*, 173–180 New York, NY, USA, Elsevier North-Holland, Inc., ISBN 0-444-86545-4, Lyon, France.

Och, Franz Josef and Hermann Ney. 2003 A Systematic Comparison of Various Statistical Alignment Models *Journal of Computational Linguistics*, Volume 29, Number, pages 19–51

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002 BLEU: a method for automatic evaluation of machine translation *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Session: Machine translation and evaluation*, 311–318 Philadelphia, Pennsylvania, Publisher: Association for Computational Linguistics Morristown, NJ, USA.

Popovic, Maja and Hermann Ney. 2006 Statistical machine translation with a small amount of bilingual training data. *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: Strategies for developing machine translation for minority languages*, 25-29, Genoa, Italy, May.

Rousu, Juho. 2008 SMART Project: Workpackage 3 advanced language models. On-line material, `http://www.smart-project.eu/files/SMART-Y1-review-WP3.v1.pdf` (last accessed on September 5th, 2011).

Smith, James and Stephen Clark. 2009 EBMT for SMT: A new EBMT-SMT hybrid. *Proceedings of the 3rd International Workshop on Example-Based Machine Translation* 3–10, Editors Mikel L. Forcada and Andy Way, Dublin, Ireland.

Stolcke, Andreas. 2002 SRILM - An Extensible Language Modeling Toolkit *Proc. Intl. Conf. Spoken Language Processing*, 901–904 September, Denver, Colorado.

Tufis, Dan and Radu Ion and Alexandru Ceausu and Dan tefnescu. 2008 RACAI's Linguistic Web Services *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008* Marrakech, Morocco, May ELRA - European Language Resources Association. ISBN 2-9517408-4-0

Andy Way and Nano Gough. 2005 Comparing example-based and statistical machine translation *Natural Language Engineering*, 1:295309, September.

Deep evaluation of hybrid architectures: simple metrics correlated with human judgments

Gorka Labaka, Arantza Díaz de Ilarraza, Kepa Sarasola
University of the Basque Country
gorka.labaka@ehu.es, jipdisaa@ehu.es, kepa.sarasola@ehu.es

Cristina España-Bonet, Lluís Màrquez
Universitat Politècnica de Catalunya
cristinae@lsi.upc.edu, lluism@lsi.upc.edu

Abstract

The process of developing hybrid MT systems is guided by the evaluation method used to compare different combinations of basic subsystems. This work presents a deep evaluation experiment of a hybrid architecture that tries to get the best of both worlds, rule-based and statistical. In a first evaluation human assessments were used to compare just the single statistical system and the hybrid one, the rule-based system was not compared by hand because the results of automatic evaluation showed a clear disadvantage. But a second and wider evaluation experiment surprisingly showed that according to human evaluation the best system was the rule-based, the one that achieved the worst results using automatic evaluation. An examination of sentences with controversial results suggested that linguistic well-formedness in the output should be considered in evaluation. After experimenting with 6 possible metrics we conclude that a simple arithmetic mean of BLEU and BLEU calculated on parts of speech of words is clearly a more human conformant metric than lexical metrics alone.

1 Introduction

The process of developing hybrid MT systems is guided by the evaluation method used to compare different combinations of basic subsystems. Direct human evaluation is more accurate but unfortunately it is extremely expensive, so automatic metrics have to be used in prototype developing. However the method should evaluate different systems with the same criteria, and these criteria should be as close as possible to human judgment.

It is well known that rule-based and phrase-based statistical machine translation paradigms (RBMT and SMT, respectively) have complementary strengths and weaknesses. First, RBMT systems tend to produce syntactically better translations and deal with long distance dependencies, agreement and constituent reordering in a better way, since they perform the analysis, transfer and generation steps based on syntactic principles. On the bad side, they usually have problems with lexical selection due to a poor handling of word ambiguity. Also, in cases in which the input sentence has an unexpected syntactic structure, the parser may fail and the quality of the translation decrease dramatically. On the other side, phrase-based SMT models usually do a better job with lexical selection and general fluency, since they model lexical choice with distributional criteria and explicit probabilistic language models. However, phrase-based SMT systems usually generate structurally worse translations, since they model translation more locally and have problems with long distance reordering. They also tend to produce very obvious errors, which are annoying for regular users, e.g., lack of gender and number agreement, bad punctuation, etc. Moreover, SMT systems can experience a severe degradation of performance when applied to corpora different from those used for training (*out-of-domain* evaluation).

It is also well known that the BLEU metric (Papineni et al., 2002) is actually the most used metric in statistical MT. But several doubts have arisen around BLEU (Melamed et al., 2003; Callison-Burch et al., 2006; Koehn and Monz, 2006). In addition to the fact that it is extremely difficult to interpret what is being expressed in BLEU (Melamed et al., 2003), improving its value neither guarantees an improvement in the translation quality (Callison-Burch et al., 2006) nor offers as much correlation with human judgment as was believed (Koehn and Monz, 2006). Those problems have also been detected when translating to Basque (Mayor, 2007; Labaka, 2010).

In the last few years, several new evaluation metrics have been suggested to consider a higher level of linguistic information (Liu and Gildea, 2005; Popović and Ney, 2007; Chan and Ng, 2008), and different methods of metric combination have been tested. Due to its simplicity, we decided to use the idea presented by Giménez and Màrquez (2008), where the different simple metrics are combined by means of the arithmetic mean.

In this work we present some surprising results we have achieved in a deep evaluation of a hybrid architecture. In a first step we used human evaluation to compare just the single statistical system and the hybrid one, we did not compare the rule-based system by hand because the results of automatic evaluation showed a clear disadvantage. But a second and wider evaluation experiment surprisingly showed that according to human evaluation the best system was the rule-based, the one that achieved the worst results using automatic evaluation. We tried to make a diagnosis of this phenomenon, and then based on this we finally found a simple but more human conformant metric that we plan to use in training new versions of our hybrid system.

In the next section of this paper we describe the hybrid system. Section 3 presents the evaluation experiments: the corpora used in them, the first experiment comparing just the single statistical system and the hybrid one, and the second and wider evaluation experiment which compares the all three systems. Then Section 4 describes the process of searching for other automatic metrics being more human conformant. And finally, the last section is devoted to conclusions and future work.

2 The hybrid system, SMatxinT

Statistical Matxin Translator, SMatxinT in short, is a hybrid system controlled by the RBMT translator and enriched with a wide variety of SMT translation options (España-Bonet et al., 2011).

The two individual systems are a rule-based Spanish-Basque system called Matxin (Alegria et al., 2007) and a standard phrase-based statistical MT system based on Moses which works at the morpheme level allowing to deal with the rich morphology of Basque (Labaka, 2010).

The initial analysis of the source sentence is done by Matxin. It produces a dependency parse tree, where the boundaries of each phrase are marked. In order to add hybrid functionality two new modules are introduced to the RBMT architecture (Figure 1): the tree enrichment module, which incorporates SMT additional translations to each phrase of the syntactic tree; and a monotonous decoding module, which is responsible for generating the final translation by selecting among RBMT and SMT partial translation candidates from the enriched tree.

The tree enrichment module introduces two types of translations for the syntactic constituents given by Matxin: 1) the SMT translation(s) of every phrase, and 2) the SMT translation(s) of the entire subtree containing that phrase. For example, the analysis of the test fragment "*afirmó el consejero de interior*" (said the Secretary of interior) gives two phrases: the head "*afirmó*" (said) and its children "*el consejero de interior*" (the Secretary of interior). The full rule-based translation is "*Barne Sailburua baieztatu zuen*" and the full SMT translation is "*esan zuen herrizaingo sailburuak*". SMatxinT considers these two phrases for the translation of the full sentence, but also the SMT translations of their constituents ("*esan zuen*" and "*herrizaingo sailburuak*"). However, short phrases may have a wrong SMT translation because of a lack of context. To overcome this problem SMatxinT also uses the translation of a phrase extracted from a longer SMT translation ("*herrizaingo sailburuak*" in the previous example). So, in order to translate "*afirmó el consejero de interior*" the system has produced 5 distinct phrases, a number that can be increased by considering a n-best list of SMT outputs.

After tree enrichment, the transfer and generation steps of the RBMT system are carried out in a usual way, and a final monotonous decoder chooses among the options. A key aspect for the performance of the system is the election of the features

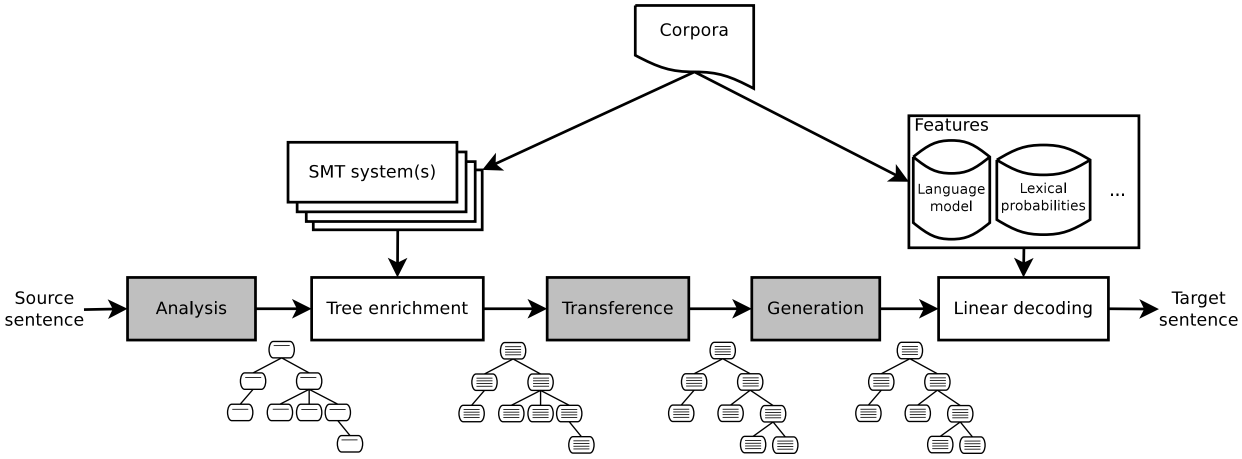


Figure 1: General architecture of SMatxinT. The RBMT modules which guide the MT process are the grey boxes.

for this decoding. The results we present here are obtained with a set of eleven features. Three of them are the usual SMT features (language model, word penalty and phrase penalty). We also include four features to show the origin of the phrase and the consensus among systems (a counter indicating how many different systems generated the phrase, two binary features indicating whether the phrase comes from the SMT/RBMT system or not, and the number of source words covered by the phrase generated by both individual systems simultaneously). Finally, we use the lexical probabilities in both directions in two forms: a similar approach to IBM-1 probabilities modified to take unknown alignments into account and a lexical probability inferred from the RBMT dictionary. We refer the reader to España-Bonet et al. (2011) for further details.

3 Experiments

In our experiments we evaluate both individual systems and the final hybrid: SMT, Matxin and SMatxinT. The language pair of application is dictated by the rule-based system and, in this case, Matxin works with the Spanish-to-Basque translation. Basque and Spanish are two languages with very different morphologies and syntaxes.

3.1 Bilingual and monolingual corpora

The corpus built to train the SMT system consists of four subsets: (1) six reference books translated manually by the translation service of the University of the Basque Country (EHUBooks); (2) a collection of 1,036 articles published in Spanish and Basque by the Consumer Eroski magazine[1] (Consumer); (3) translation memories mostly using administrative language developed by Elhuyar[2] (ElhuyarTM); and (4) a translation memory including short descriptions of TV programmes (EuskaltelTB). Table 1 shows some statistics on the corpora, giving some figures about the number of sentences and tokens.

		sentences	tokens
EHUBooks	Spanish Basque	39,583	1,036,605 794,284
Consumer	Spanish Basque	61,104	1,347,831 1,060,695
ElhuyarTM	Spanish Basque	186,003	3,160,494 2,291,388
EuskaltelTB	Spanish Basque	222,070	3,078,079 2,405,287
Total	Spanish Basque	491,853	7,966,419 6,062,911

Table 1: Statistics on the bilingual collection of parallel corpora.

The training corpus is then basically made up of administrative documents and descriptions of TV programs. For development and testing we extracted some administrative data for the *in-domain* evaluation and selected one collection of news for the *out-of-domain* study, totaling three sets:

Elhuyardevel and *Elhuyartest*: 1,500 segments each, extracted from the administrative documents.

[1] http://revista.consumer.es

[2] http://www.elhuyar.org/

NEWStest: 1,000 sentences collected from Spanish newspapers with two references.

Additionally, we collected a 21 million word monolingual corpus, which together with the Basque side of the parallel bilingual corpora, builds up a 28 million word corpus. This monolingual corpus is also heterogeneous, and includes text from two sources of news: the Basque corpus of Science and Technology (ZT corpus) and articles published by Berria newspaper (Berria corpus).

3.2 First experiment of evaluation

According to the automatic evaluation, carried out in the previous article and extended in Table 4, the rule-based Matxin system is clearly the worst system obtaining the worst scores for both metrics (BLEU and TER) in both test corpora. On the other hand, the evaluation of the hybrid system varies depending on the test set. On the in-domain corpora (Elhuyar test set), the BLEU score achieved by SMatxinT is slightly worse than the scores obtained by the single SMT system, but better according to TER (Snover et al., 2006) evaluation. The distinct behavior between metrics and the small differences do not allow us to define a clear preference between statistical and hybrid systems. On the contrary, on the out-domain corpora (NEWS test set), SMatxinT consistently archives better scores than any other system.

Based on these results, we stated that the low in-domain performance of the Matxin penalizes the hybrid system, preventing it to overcome the single SMT system. But, in the out-domain test set, where the scores of Matxin were not so far from the rest of the systems, our hybridization technique was able to combine the best of both systems obtaining the best translation. In order to verify this assertion, we carried out an human evaluation, where we asked four evaluators to determine the preference between the hybrid and the SMT translations of 100 sentences randomly chosen from the NEWS test set. The figures obtained corroborated that the hybrid system outperforms the single SMT system in the out-domain corpora.

3.3 Deeper evaluation: Human evaluation to compare the three systems

In order to get a more detailed insight of the performance of our systems, we recently extended this manual evaluation to the rest of the systems and test corpora. That way, we selected another 100 sentences from the Elhuyar test set and asked the same four evaluators to assess the preference between the three system pairs (SMT-Matxin, SMT-SMatxinT, Matxin-SMatxinT).

Surprisingly, according to this manual evaluation the best system is the rule-based Matxin system, the worst ranked one using automatic evaluation. Even for in-domain evaluation it is clearly better than the statistical system and of similar quality as the hybrid one, that is slightly superior to the statistical system. For out-domain evaluation the differences are very clear: the rule-based Matxin system clearly outperforms the hybrid system and this one outperforms the statistical system.

This can be seen in Table 2. The table shows the number of times that a system is better than the other for those sentences where there was full agreement among evaluators (*Agreement*) and for the full subset (*All*). Results are given for the three system pairs on the two test sets, the in-domain and the out-of-domain ones.

We confirmed these surprising results of manual evaluation by examining some examples where BLEU scores did not reflect the difference of quality between translation outputs. Let us analyze the example shown in Table 3, that is, the translation of the source sentence *"Legasa cuenta ya con un convenio sobre la recuperación de bienes comunales."*. The table shows the source sentence with its meaning in English together with two translation references and the output given by the two individual systems.

In this example, the output of the rule-based system is adequate, but BLEU is unable to recognize some linguistic equivalences: *jadanik* and *jada* are synonymous, as well as *berreskuratzearen inguruan* and *berreskuratze gainean.* Similarly *herri ondasunak* and *herri-ondasunen* are almost the same because the "-" is optional, and using *Legasa* instead of *Legasak* is a common error easy to understand. The following segments are quasi equivalents: *hitzarmena du* and *kontatzen du hitzarmen batekin*. All these correspondences are trivial for humans but invisible for the BLEU metric.

On the other hand, the output of the statistical system is harder to understand. By using *Legasako* instead of *Legasak*, the sentence becomes difficult to

			System 1	**Tied**	**System 2**
Elhuyar (in-domain)	SMT vs. SMatxinT	Agreement	5 (9.4%)	**31 (58.5%)**	17 (32.1%)
		All	25 (12.5%)	**109 (54.5%)**	66 (33.0%)
	SMT vs. **Matxin**	Agreement	14 (23.7%)	19 (32.2%)	**26 (44.1%)**
		All	41 (20.5%)	79 (39.5%)	**80 (40.0%)**
	SMatxinT vs. Matxin	Agreement	19 (28.8%)	**24 (36.4%)**	23 (34.8%)
		All	59 (29.5%)	**82 (41.0%)**	59 (29.5%)
NEWS (out-domain)	SMT vs. **SMatxinT**	Agreement	15 (21.4%)	22 (31.4%)	**33 (47.2%)**
		All	40 (20.0%)	74 (37.0%)	**86 (43.0%)**
	SMT vs. **Matxin**	Agreement	11 (17.7%)	13 (21.0%)	**38 (61.3%)**
		All	32 (16.0%)	64 (32.0%)	**104 (52.0%)**
	SMatxinT vs. **Matxin**	Agreement	19 (26.4%)	13 (18.1%)	**40 (55.5%)**
		All	49 (24.5%)	54 (27.0%)	**97 (48.5%)**

Table 2: Manual evaluation for random subset of 100 sentences of each test corpus.

Source	*Legasa cuenta ya con un convenio sobre la recuperación de bienes comunales.*
(English)	*Legasa already has a convention on the recovery of community property.*
Ref. 1	*Legasak hitzarmena du jada herri ondasunak berreskuratzearen inguruan.*
Ref. 2	*Legasak badauka ondasun komunalak berreskuratzeari buruzko hitzarmena.*
Matxin	*Legasa jadanik kontatzen du hitzarmen batekin herri-ondasunen berreskuratze gainean.*
SMT	*legasako hitzarmena du dagoeneko berreskuratzeari buruzko ondasunak komunalak.*
SMatxinT	*dagoeneko legasako hitzarmena berreskuratzeari buruzko ondasun komunalak*

Table 3: Example where an understandable translation obtained by Matxin is penalized by BLEU, but the confusing SMT translation gets a good BLEU score.

understand, and the same happens with the strange end of the sentence. However, this translation obtains a good evaluation score because every word but one is in the references.

4 Searching for human conformant automatic metrics

In view of the large difference between the results obtained by standard automatic metrics and the manual evaluation, and considering that the human evaluators value syntactical correctness more than the common lexical metrics (such as BLEU and TER) do, we considered the possibility of using metrics that use a higher level of linguistic information (Liu and Gildea, 2005; Popović and Ney, 2007; Giménez and Màrquez, 2007; Chan and Ng, 2008). Thus, in addition to the standard BLEU and TER, we applied these same metrics over the sequences of syntactic categories, parts of speech (PoS), resulting BLEU_PoS and TER_PoS. Table 4 shows how the metrics that use linguistic information obtain more similar results to those achieved by the manual evaluation. Thus, in our out-domain evaluation the metrics that use PoS information show the same preference between systems than the human assessment. That is, Matxin gets the best results, followed by SMatxinT and SMT. Similarly, in the in-domain test set, the human preference of SMatxinT over the statistical system is clearer with this type of metrics. Despite this, PoS based metrics can not fully compensate the high penalty that Matxin receives and this system remains the lowest ranked in the Elhuyar test set (in-domain), although the distance is shorter.

However, those results are provably biased by the fact that both SMT and SMatxinT systems are optimized to rise their BLEU score. Thus, they get a high lexical matching to the reference, at the expense of the syntactical correctness. Similarly, the use of metrics that only take into account a even more specific aspect of translation, such as the coincidence of PoS, are not suitable to be used as the unique metric for the whole developing cycle. Using such metrics on SMT parameter optimization, for example, could lead to get translations whose lexical correction is fully ignored. So this kind of met-

		BLEU	TER	BLEU_PoS	TER_PoS	comb_BLEU	comb_all
Elhuyar (in-domain)	**Matxin**	5.25	84.51	25.63	52.82	15.44	7.88
	SMT	**14.53**	71.60	30.78	48.82	22.65	11.53
	SMatxinT	14.48	**70.50**	**31.96**	**47.07**	**23.22**	**11.82**
Elhuyar (in-domain) hand evaluated sentences	**Matxin**	5.85	84.95	26.68	52.19	16.27	8.29
	SMT	12.75	75.58	30.15	49.37	21.45	**11.38**
	SMatxinT	**13.37**	**75.09**	**31.39**	**48.63**	**22.38**	**11.38**
NEWS (out-domain)	**Matxin**	11.65	72.39	**39.19**	**42.40**	**25.42**	**12.93**
	SMT	14.45	70.18	31.09	48.65	22.77	11.59
	SMatxinT	**15.08**	**67.72**	34.55	45.56	24.82	12.62
NEWS (out-domain) hand evaluated sentences	**Matxin**	11.01	73.55	**38.74**	**43.07**	**24.88**	**12.65**
	SMT	11.32	73.08	29.56	50.49	20.44	10.41
	SMatxinT	**13.64**	**70.42**	35.34	46.82	24.49	12.45

Table 4: Automatic scores of all individual and hybrid systems.

rics should be combined with metrics that also take into account other aspects of the translation, as lexical matching. In the literature different methods of metric combination have been tested. Among other methods, one can find those based on linear combinations (Padó et al., 2009; Liu and Gildea, 2007; Giménez and Màrquez, 2008), regression based algorithms (Paul et al., 2007; Albrecht and Hwa, 2008) or a variety of supervised machine learning algorithms (Quirk et al., 2005; Amigó et al., 2005).

Due to its simplicity and the results achieved, we decided to use the idea presented by Giménez and Màrquez (2008), where the different metrics are combined just by means of the arithmetic mean. This method of combination, despite its simplicity, obtained competitive results on the MetricsMATR shared task (Callison-Burch et al., 2010). Thus we have defined two metrics that combine lexical information with PoS information: (1) one that combines the four metrics (BLEU, TER, BLEU_PoS and TER_PoS) we tested and (2) another one that combines only BLEU with BLEU_PoS.

BLEU and BLEU_PoS are quality measures (higher score means higher quality) while TER and TER_PoS are error measure (lower score means higher quality). Due to the different nature of the metrics and to be able to combine all of these four metrics by means of the arithmetic mean, we had to modify the values of TER to become quality measures. Thus, the new metrics are calculated using the following formulas:

$$Comb\_BLEU = (BLEU + BLEU\_PoS)/2$$

$$\begin{aligned} Comb\_All = &(BLEU + BLEU\_PoS + (100 - TER) \\ &+ (100 - TER\_PoS))/4 \end{aligned}$$

The two metrics that combine lexical metrics with PoS information obtained results similar to those based only on PoS, in terms of preference between systems. In the same way, BLEU_PoS and TER_PoS, Comb_BLEU and Comb_All established the same preference order as the manual evaluation, except in the case of Matxin in the in-domain test set. But, unlike those metrics based only on PoS information, the combined metrics are more suitable as they allow a better syntactic adequacy while they maintain correct lexical matchings.

In addition to this correlation at the document level, we also wanted to check the correlation of each metric at sentence level where manual assessments were set. For each sentence in which both human assessments agree, we have compared the result with the preference for each metric. To define which is the preference for each metric, we considered that the automatic metric prefers a translation if one of the translations gets a score 10% higher than the other. In cases where the relative difference is not higher than 10%, we consider that the automatic metric is not able to discriminate between the two translations. Table 5 shows the percentage of sentences where each automatic metric's preference coincides with the one set by both human evaluators (we discard the cases in which human evaluations have not agreed).

		BLEU	TER	BLEU_PoS	TER_PoS	comb_BLEU	comb_all
Elhuyar (in-domain)	SMT vs. SMatxinT	34 (64%)	33 (62%)	33 (62%)	30 (56%)	**35 (66%)**	31 (58%)
	SMT vs. Matxin	23 (39%)	23 (39%)	25 (42%)	22 (37%)	24 (41%)	**26 (44%)**
	SMatxinT vs. Matxin	25 (38%)	**29 (44%)**	**29 (44%)**	27 (41%)	25 (38%)	28 (42%)
NEWS (out-domain)	SMT vs. SMatxinT	35 (50%)	31 (44%)	36 (51%)	**38 (54%)**	**38 (54%)**	34 (49%)
	SMT vs. Matxin	31 (50%)	29 (47%)	**42 (68%)**	38 (61%)	39 (63%)	**42 (68%)**
	SMatxinT vs. Matxin	38 (53%)	38 (53%)	39 (54%)	36 (50%)	**46 (64%)**	36 (50%)

Table 5: Sentence by sentence correlation between human evaluation and automatic metrics.

These figures show that the metrics based on linguistic information (both, those that only uses PoS information and those that combine it with lexical information) get more coincidences than those that only use lexical information (BLEU or TER).

5 Conclusions

In this work we present an in-depth evaluation of SMatxinT, a hybrid system that is controlled by the RBMT translator and enriched with a wide variety of SMT translation options. The results of the human evaluation, where the translation of the two individual systems and SMatxinT were compared in pairs, established that Matxin, the RBMT system, achieved the best performance followed by SMatxinT, while the SMT system generated the worst translations.

Those results, very far from what the automatic metrics (BLEU and TER) show, corroborate the already known inadequacy of the metrics that measure only the lexical matching for comparing systems that use so different translation paradigms. This kind of metrics are biased in favor of the SMT, as it happens in our evaluation, where the statistical system achieves the best results in the in-domain evaluation, even when it generates the worst translations according to the manual assessment.

To address these limitations of the metrics that are only based on lexical matching, we defined a couple of metrics that seek to ensure the syntactic correctness, calculating the same expressions but at the PoS level. These metrics, which are able to assess the syntactic correctness, have shown a higher level of agreement with human assessments both at document and sentence level.

Nevertheless, the metrics that assess specific aspects of the translation (such as PoS matching) do not ensure the absolute quality of the translation, and should be combined with regular lexical matching metrics. At the time of combining these metrics, we opted for simplicity and we used the arithmetic mean. This method, despite its simplicity, has already shown its suitability before.

Our combined metrics are simple and able to maintain a higher correlation with manual evaluation than the usual lexical metrics, while ensure the lexical matching.

We are planning to use this simple combination of metrics in developing new versions of our hybrid system. Simultaneously we are adapting linguistic tools to the Asiya Open Toolkit[3] to test other new evaluation metrics that consider a higher level of linguistic information.

Acknowledgments

This work has been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247914 (MOLTO project, FP7-ICT-2009-4-247914), the Spanish Ministry of Science and Innovation (OpenMT-2 project, TIN2009-14675-C03-01) and the Local Government of the Basque Country (Berbatek project, IE09-262).

References

Joshua Albrecht and Rebecca Hwa. 2008. Regression for machine translation evaluation at the sentence level. *Machine Translation*, pages 1–27.

Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Aingeri Mayor, and Kepa Sarasola. 2007. Transfer-based MT from spanish into basque: Reusability, standardization and open source. *Lecture Notes in Computer Science*, 4394:374–384.

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo. 2005. Qarla: a framework for the evalua-

[3] http://nlp.lsi.upc.edu/asiya

tion of text summarization systems. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 280–289, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the International Conference of European Chapter of the Association for Computational Linguistics (EACL)*, pages 249–256.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 17–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62, Columbus, Ohio, June. Association for Computational Linguistics.

Cristina España-Bonet, Gorka Labaka, Arantza Díaz de Ilarraza, Lluis Màrquez, and Kepa Sarasola. 2011. Hybrid machine translation guided by a rulebased system. In *Proceedings MT Summit XIII*, Xiamen, China, Septemper.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 256–264, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio, June. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *In Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*, pages 102–121.

Gorka Labaka. 2010. *EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation*. Ph.D. thesis, University of the Basque Country.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 25–32.

Ding Liu and Daniel Gildea. 2007. Source-language features and maximum correlation training for machine translation evaluation. In *HLT-NAACL'07*, pages 41–48.

Aingeru Mayor. 2007. *Matxin: erregeletan oinarritutako itzulpen automatikoko sistema*. Ph.D. thesis, Euskal Herriko Unibertsitatea.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 61–63, Morristown, NJ, USA. Association for Computational Linguistics.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Chris Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 297–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2007. Reducing Human Assessments of Machine Translation Quality to Binary Classifiers. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Maja Popović and Hermann Ney. 2007. Word error rates: decomposition over pos classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 48–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 271–279, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231.

A Radically Simple, Effective Annotation and Alignment Methodology for Semantic Frame Based SMT and MT Evaluation

Chi-kiu Lo and **Dekai Wu**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
jackielo,dekai @cs.ust.hk

Abstract

We introduce a radically simple yet effective methodology for annotating and aligning semantic frames inexpensively using untrained lay annotators that is ideally suited for practical semantic SMT and evaluation applications. For example, recent work by Lo and Wu (2011) introduced MEANT and HMEANT, which are state-of-the-art metrics that evaluates translation meaning preservation via Propbank style of semantic frames. For such applications, however, we argue that the Propbank annotation are too complex and detailed, since they are aimed at training linguists to annotate semantic frames with gold standard accuracy. Instead, we believe that annotating semantic frames for such purposes should be as intuitive as understanding the basic event structure of a sentence, which any untrained human does effortlessly. We propose a simplified set of annotation guidelines consisting of half a page plus three annotated examples. Together with a graphical user interface designed to facilitate the annotation and comparison process by guiding untrained humans step by step, only 5 to 15 minutes are needed to train lay annotators. This allows the lay annotators to focus on understanding the translation to provide consistent and efficient annotation and comparison. The methodology is 'cloud' based to be truly platform independent, installation-free and portable.

1 Introduction

We present a practical alternative to linguistically sophisticated but expensive methodologies semantic frame annotation and alignment, designed in particular with an eye to semantic statistical machine translation (SMT) and MT evaluation. Our approach contrasts with, for example, the complex guidelines for Propbank annotation (Bonial *et al.*, 2010) used to train linguists to annotate semantic frames with gold standard accuracy. Though excellent for their intended purpose, Propbank style guidelines are long and full of linguistic terminology, making them highly unsuitable for training lay persons.

Our efforts are motivated by the increasing needs of recent work on semantic SMT and semantic MT evaluation. In semantic SMT for example, the SRL-for-SMT work of Wu and Fung (2009a) and Wu and Fung (2009b) relies on cross-lingual matching of semantic role labels. In semantic MT evaluation, the metrics MEANT and HMEANT from Lo and Wu (2011a,b,c) are also based on SRL matching.

New research directions of this kind demand quick, inexpensive, relative accurate semantic frame annotation and alignment. We argue that the methodology for annotating semantic frames for such purposes should be as easily intuitive as comprehending the basic event structure of a sentence — which any untrained native speaker does naturally and effortlessly.

Our alternative methodogy achieves this by combining (1) a streamlined, highly simplified and intuitive set of annotation guidelines with (2) an easy-to-use graphical user interface that guides untrained lay annotators step-by-step through the annotation process within (3) a convenient 'cloud' based platform that flexibly supports distributed workflows involving physically separated annotators working on any standard browser.

The streamlined annotation guidelines consist of a mere half-page of instructions — mostly whitespace — supplemented with three annotated examples for reference. The simplicity of the guidelines allows lay annotators to focus on understanding the translation to provide consistent and efficient annotation and comparison. Training an annotator typically takes on the order of five minutes. Despite (or perhaps because) of the simplicity, interannotator agreement is nevertheless quite high.

A graphical user interface is specifically designed to address the risk of annotation inconsistency that arises from using unskilled humans rather than linguistic experts to annotate semantic frames. The guidelines are incorporated into a GUI that guides annotators to label semantic predicate argument structure. The system guides the annotators to first identify the predicate of a frame, and then specify the span and the role of its associated arguments one by one. Every time when the annotators label a predicate, they start the process of annotating a new semantic frame in the sentence. Each annotated frame is marked up with a different color, so that annotators can clearly distinguish the multiple semantic frames within a single sentence.

Convenient annotation workflows across distributed locations are facilitated by the 'cloud' based approach. The cross-platform web interface is accessible from any mordern Javascript-enabled browser. Annotation of translation is currently supported in any language encoded in UTF8 with left to right orthography. Text is expected to be segmented into sentences, reflecting the common assumption of nearly all present day MT systems.

In the following sections, we first contrast our approach with related work on the process methodology for Propbank annotation. We then propose a concrete set of annotation guidelines. Next, we describe the design of a graphical user interface specifically tailored to guide lay annotators step by step through the process of annotating semantic frames with our simplified set of role labels. Following this, we propose a set of guidelines for aligning and comparing semantic frames for translations, again designed to be easy for lay annotators and yet sufficiently accurate. We also describe the design of the graphical user interface for alignment of semantic frames. Finally, we present experimental results on timing lay annotators, demonstrating the efficiency and low cost of this methodology (which has been shown elsewhere to produce state-of-the-art results for semantic MT evaluation).

2 Related Work

The Propbank annotation guidelines (Bonial *et al.*, 2010) are aimed at training linguists to annotate semantic frames to gold standard accuracy, and are unnecessarily long and technical for lay persons. Propbank requires annotators to determine the word sense for each predicate and is built on top of the syntactic structure in the sentences, using the software tool Jubilee (Choi *et al.*, 2010) to support the complex Propbank annotation and viewing process. Thus, the annotator training cost of Propbank annotation is disproportionately high for applications such as semantic MT evaluation.

Recent works in semantic SMT and MT evaluation show an increasing demand for low-cost semantic frame annotation and comparison. In semantic SMT, for example, Wu and Fung (2009a) and Wu and Fung (2009b) apply SRL to SMT decoding, using an SRL based reordering model that returns improved translations containing fewer semantic role confusion errors. The SRL based reordering model relies on cross-lingual SRL matching. In semantic MT evaluation, a new generation of automatic and semi-automatic MT evaluation metrics proposed by Lo and Wu (2011a,b,c) captures similarities and differences between the reference translation and MT output semantic structures. This approach also relies on SRL matching between reference translation and MT output.

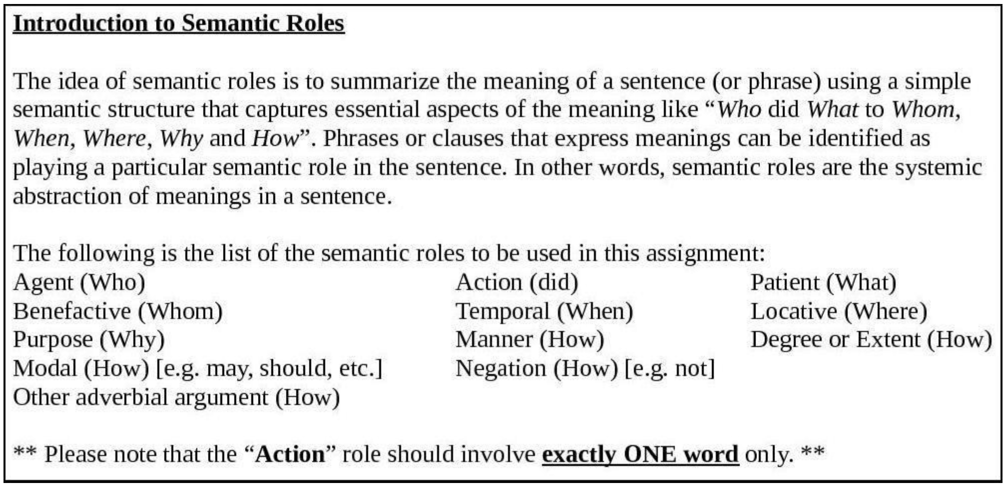
The Propbank annotation guidelines consist of 70 pages, of which 59 pages are annotation instructions and 11 pages cover the menu for the annotation tools. The annotation instructions detail the annotation process, the definition of the argument labels, exception handling for tagging, the handling of null elements in syntax trees and the handling of special cases and spoken data. Since Probank is built on top of the syntactic structure of the sentences, Propbank annotators, i.e. readers of the guidelines, are expected to have prior knowledge of word senses, syntactic structure annotations, and other linguistic information (e.g. null elements). However, all the details in the Propbank annotation guidelines are only necessary when the goal of annotation is to provide consistent and high quality gold standard semantic frame annotation. In contrast, extracting semantic information for practical applications such as semantic SMT and evaluation should be as intuitive as understanding the basic event structure of a sentence which any untrained human does effortlessly.

Since Propbank aims to provide gold standard semantic frame annotation, the annotations are subsequently adjudicated. Therefore, Jubilee, the Propbank instance annotation editor has complex use cases and consists of two modes: the 'normal' mode and the 'gold mode'.

The normal mode is used by annotators to determine the word sense for each predicate in the sentence and annotate the arguments with semantic role labels. Since the Propbank annotation is built on top of the syntactic structure of the sentences and requires annotators to first determine the word sense of the predicate, the normal mode consists of three panels — the treebank view, the frameset view and the argument view. Annotators must navigate around these panels in the different steps of annotation.

One the other hand, the gold mode is used by the adjudicators who select the most appropriate annotation of the instance as the gold standard or correct the annotations if necessary. To determine which annotation of the instance is the most appropriate as the gold standard, in addition to the three panels in the normal mode, the gold mode includes one more panel showing all the annotations for the instance. Similarly, adjudicators must navigate between all the four panels in the different step of adjudication.

The complexity of Jubilee is only necessary when the goal of annotation is providing consistent and high quality gold standard semantic frame annotation. In contrast, for practical applications such as semantic SMT and evaluation, the semantic frame annotation tool should be



Introduction to Semantic Roles

The idea of semantic roles is to summarize the meaning of a sentence (or phrase) using a simple semantic structure that captures essential aspects of the meaning like "*Who* did *What* to *Whom*, *When*, *Where*, *Why* and *How*". Phrases or clauses that express meanings can be identified as playing a particular semantic role in the sentence. In other words, semantic roles are the systemic abstraction of meanings in a sentence.

The following is the list of the semantic roles to be used in this assignment:
Agent (Who) Action (did) Patient (What)
Benefactive (Whom) Temporal (When) Locative (Where)
Purpose (Why) Manner (How) Degree or Extent (How)
Modal (How) [e.g. may, should, etc.] Negation (How) [e.g. not]
Other adverbial argument (How)

** Please note that the "**Action**" role should involve **exactly ONE word** only. **

Figure 1: Instruction of semantic frame annotation for MT evaluation

straightforward and require minimal training instructions in using the tool itself. A software tool supporting these kinds of annotation should be easy to use so that lay annotators can concentrate on evaluating the meaning of the translation and provide consistent annotations for evaluation.

3 Annotating Semantic Frames

To minimize the labor cost of running the semantic MT evaluation metric so that it can be driven by untrained monolingual human, the instructions for annotating semantic frames have to be clear, simple and intuitive. MEANT (Lo and Wu, 2011a) adopted Propbank SRL style predicate-argument framework, which captures the basic event structure in a sentence. The original Propbank annotation specification is designed for readers with strong linguistic background who can distinguish different word senses of predicates. We present the intuitive guidelines and step-by-step guided interface that make semantic role labeling, i.e. identifying the basic event structure—"who did what to whom, when, where and why" (Pradhan *et al.*, 2004) — a task that even untrained monolingual readers can do.

3.1 Simplified set of labels and minimal guidelines

In contrary to the 89 pages of Propbank annotation guidelines, we simplified the instructions of annotation into half of a page intuitively. We first clearly state the objective of semantic role labeling using lay person terminologies. Then, according to the basic event structures—"who did what to whom, when, where and why", we simplified the set of Propbank style semantic role labels into a set of 10 to 12 role labels. Figure 1 shows the half-page instructions with the simplified set of roles.

The "did" event which corresponds to the predicate in the semantic frame is defined as "Action".

The "who" event which corresponds to the subject of the predicate (i.e. ARG0) in the semantic frame is defined as"Agent".

The "what" event which corresponds to the object of the predicate (i.e. ARG1) in the semantic frame, (in other words, "the argument which undergoes the change of state or is being affected by the action" (Bonial *et al.*, 2010)), is defined as "Patient".

The "whom" event which corresponds to the benefactive argument of the predicate (i.e. ARG2) in the semantic frame is defined as "Benefactive".

The "when" event which corresponds to the temporal argument of the predicate (i.e. ARGM-TMP) in the semantic frame is defined as "Temporal".

The "where" event which corresponds to the locative argument of the predicate (i.e. ARGM-LOC, ARGM-DIR) in the semantic frame is defined as "Locative".

The "why" event which corresponds to the cause or purpose argument of the predicate (i.e. ARGM-CAU, ARGM-PRP) in the semantic frame is defined as "Purpose".

Since the "how" event which corresponds to the more detailed modifiers of the predicate, Lo and Wu (2011b) presented experiments on different variants of sub-categorizing the "how" event.

To concretize the lay human annotators' understanding of the role labels, three annotated examples were provided. The examples were shown in the order of advance-

Example 1:

```
I send a message to you on Saturday at home for making appointment .
```

Agent:	I
Action:	send
Patient:	a message
Benefactive:	to you
Temporal:	on Saturday
Locative:	at home
Purpose:	for making appointment

Example 2:

```
A few days ago , the National Development Bank successfully issued
30 billion yen of samurai bonds to Japan 's capital market .
```

Agent:	the National Development Bank
Action:	issued
Patient:	30 billion yen of samurai bonds
Temporal:	A few days ago
Locative:	Japan 's capital market
Manner:	successfully

Example 3:

```
South Korea 's Ministry of Agriculture and Forestry said this
evening that an Asan City duck farm reported to the relevant
department on the 11th that since the 5th of this month , the
number of egg production of over 9,000 ducks in the duck farm had
fallen sharply .
```

Agent 1:	South Korea 's Ministry of Agriculture and Forestry
Action 1:	said
Patient 1:	an Asan City duck farm reported to the relevant department on the 11th that since the 5th of this month , the number of egg production of over 9,000 ducks in the duck farm had fallen sharply
Temporal 1:	this evening
Agent 2:	an Asan City duck farm
Action 2:	reported
Patient 2:	since the 5th of this month , the number of egg production of over 9,000 ducks in the duck farm had fallen sharply
Benefactive 2:	the relevant department
Temporal 2:	on 11th
Agent 3:	the number of egg production of over 9,000 ducks in the duck farm
Action 3:	fallen
Temporal 3:	since the 5th of this month
Manner 3:	sharply

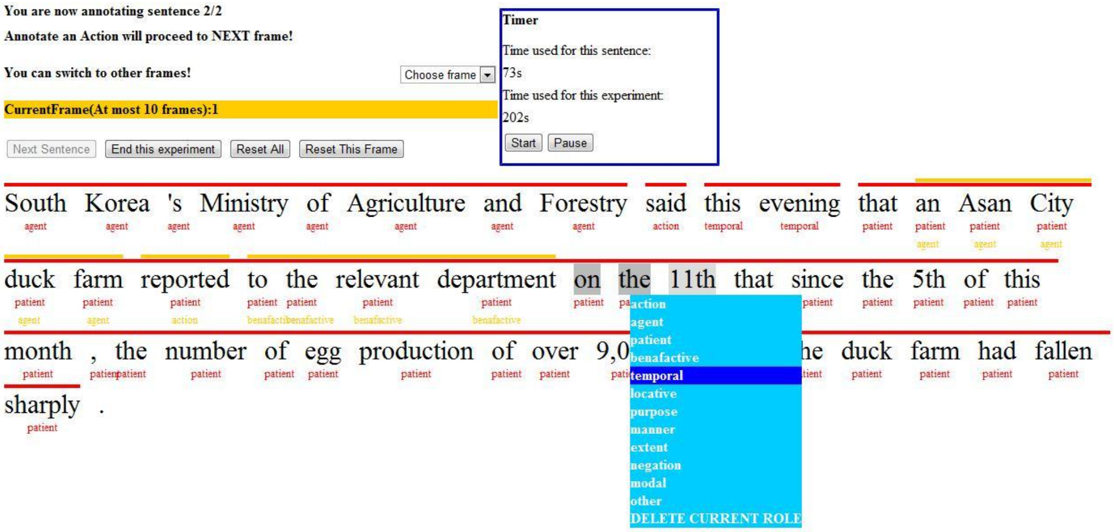
Figure 2: Annotated examples

ment of semantic structures. The first two examples contained one predicate only and the last example contained three predicates. Figure 2 shows the three annotated examples provided to annotators to concretize their understanding of the simplified set of semantic role labels.

3.2 Semantic frame annotation web interface

Annotators are allowed to view and annotate only translations that are assigned to them. Therefore, users have to login to the system. A login page is shared by the annotation web interface and the comparison web interface that is introduced in later section. After logging in, annotators are sent to the annotation dataset claiming page, where they can see the list of datasets that is assigned to themselves and the list of datasets that they have already annotated. Figure 3 shows the task claiming page.

Figure 4 shows the annotation page with an annotation in progress. The page can be divided into three panels.



You are now annotating sentence 2/2

Annotate an Action will proceed to NEXT frame!

You can switch to other frames!

CurrentFrame(At most 10 frames):1

Timer

Time used for this sentence:

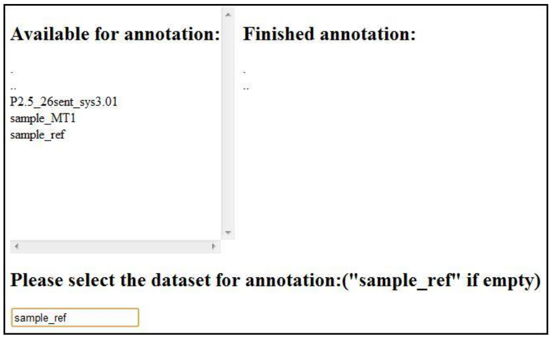
73s

Time used for this experiment:

202s

South Korea 's Ministry of Agriculture and Forestry said this evening that an Asan City duck farm reported to the relevant department on the 11th that since the 5th of this month , the number of egg production of over 9,0 he duck farm had fallen sharply .

Figure 4: Semantic frame annotation web interface



Available for annotation:

.

..

P2.5_26sent_sys3.01

sample_MT1

sample_ref

Finished annotation:

.

..

Please select the dataset for annotation:("sample_ref" if empty)

sample_ref

Figure 3: Annotation task claiming page

The top left corner is the information and control panel where annotators receive information about the progress and control the annotation process. The top right corner is the timer panel. The lower panel is the annotation panel.

The first line on the top left corner shows the progress of the annotation task. In this screen shot, the annotator is annotating the second sentence in a dataset of 2 sentences. The next line reminds the annotators to annotate the action first to start annotating a new frame. This is designed according the linguistic formation of predicate-argument structure of semantic frame. The third line on the top left corner reminds the annotators if they find any error in previous annotated frames, they can choose the corresponding frame from the associated combo box at the end of the line. The fourth line is colored to show the annotator clearly which frame they are currently editing. Following the first four lines, there are four buttons. When the annotator finishes annotating the current sentence, he/she should either click “Next Sentence” if there are more sentences in the data set for annotation, or “End this experiment” if there is no more sentence in the data set. “Reset All” allows the annotators to remove all annotations in all frames of the current sentence. “Reset this Frame” allows the annotators to remove all annotations in the current frame.

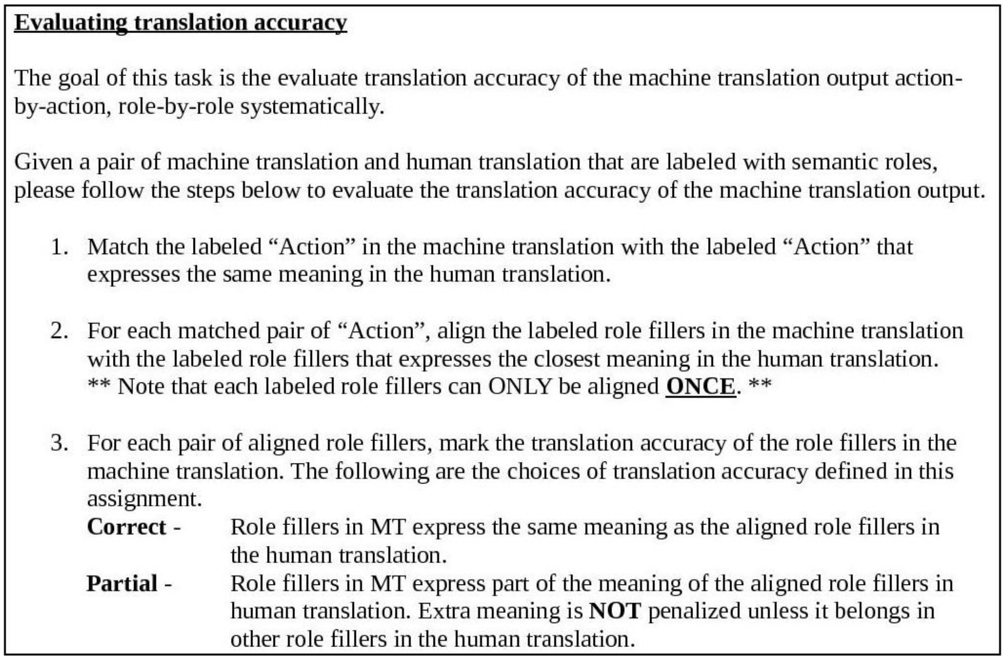
On the top right corner, there is the timer showing the time used for the current sentence and the current task. There are two buttons in the timer, “Start” and “Pause”. The sentence will be covered up if the timer is not started to ensure accurate timing.

The lower half of the page is the annotation panel. The current sentence is shown in the annotation panel. The colored lines above the sentence indicate the span of the semantic role. The colored labels below the sentence indicate the label of the semantic role. One frame is represented by one color. the annotations in all frames are shown to the annotators at the same time in the same panel so that the annotators can see the whole event structure they annotated and verify the annotations easily.

The annotators click on the word token at the begining of a role span and click on the work token at the end of the same role span to specify the span of the semantic role. After that, a pop up menu will be shown to let annotators to determine the role label. After selecting the role label, the pop up menu will be hidden again and the annotators can continue annotating other roles or frames.

4 Aligning/comparing Semantic Frames

After annotating the semantic frames, we must then determine the translation accuracy of the role fillers. To overcome the disadvantages of resorting to excessively permissive bag-of-words matching or excessively restrictive exact string matching, human judges were employed



Evaluating translation accuracy

The goal of this task is the evaluate translation accuracy of the machine translation output action-by-action, role-by-role systematically.

Given a pair of machine translation and human translation that are labeled with semantic roles, please follow the steps below to evaluate the translation accuracy of the machine translation output.

1. Match the labeled "Action" in the machine translation with the labeled "Action" that expresses the same meaning in the human translation.

2. For each matched pair of "Action", align the labeled role fillers in the machine translation with the labeled role fillers that expresses the closest meaning in the human translation. ** Note that each labeled role fillers can ONLY be aligned **ONCE**. **

3. For each pair of aligned role fillers, mark the translation accuracy of the role fillers in the machine translation. The following are the choices of translation accuracy defined in this assignment.
   Correct - Role fillers in MT express the same meaning as the aligned role fillers in the human translation.
   Partial - Role fillers in MT express part of the meaning of the aligned role fillers in human translation. Extra meaning is **NOT** penalized unless it belongs in other role fillers in the human translation.

Figure 5: Semantic frame comparison guidelines for MT evaluation

to evaluate the correctness of each role filler translation between the reference and machine translations. However, with the ultimate goal of automating this step, the definition of translation correctness in meaning must be well-defined. Moreover, to facilitate a finer-grained measurement of translation utility, the definition of translation correctness must also be finer-grained. We present the fine-grained but well-defined choices of translation correctness and minimal guidelines for semantic MT evaluation.

4.1 Fine-grained but well-defined choices of correctness and minimal guidelines

To avoid the inconsistency among human judges , instead of adopting 5-point or 7-point scales used in translation adequacy judgment, we define the translation correctness of role fillers as three cardinal marks, i.e. "correct", "partial" and "incorrect". Since predicate verb is exactly one word, either the machine translation express the same action or not the same action, we only define "correct" and "incorrect' for predicate. Figure 5 shows the fine-grained but well-defined choices of translation correctness.

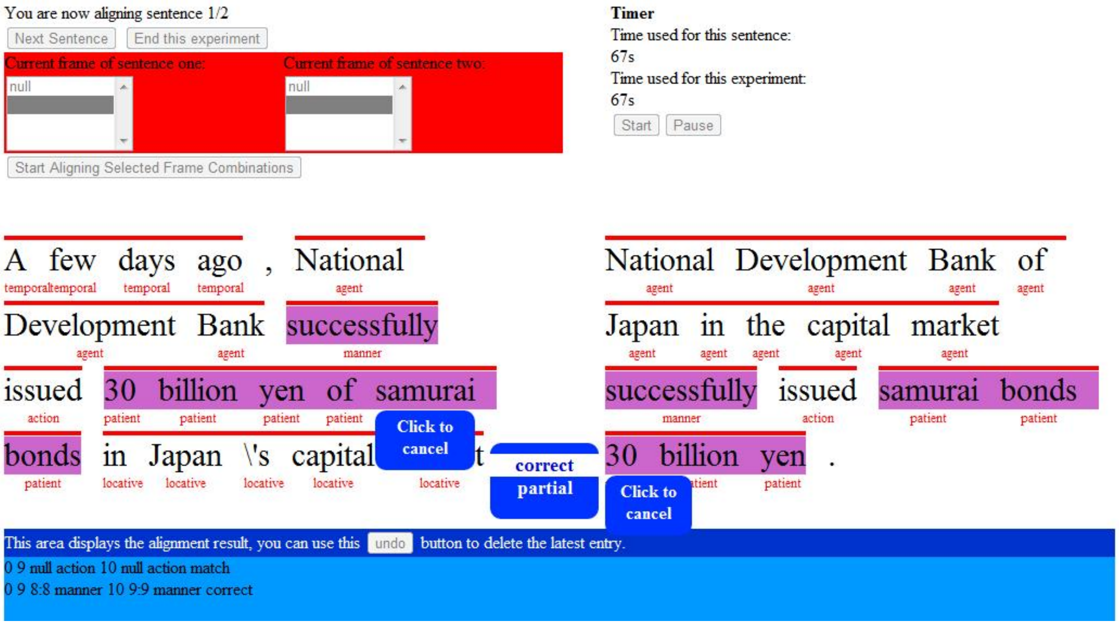
> Role fillers in MT, that express the same meaning as that in the reference translation, is considered as a "correct" translation.
>
> Role fillers in MT, that express a part of the meanings of that in the reference translation, is considered as a "partial" correct translation. Extra meaning is not penalized unless it belongs in another role.

We also assume that a wrongly translated predicate means that the entire semantic frame is incorrect; therefore, the "correct" and "partial" argument counts are collected only if their associated predicate is correctly translated in the first place.

4.2 Semantic frame comparison web interface

Similar to the annotation web interface, human judges are allowed to view and judge only translations that are assigned to them. After logging in, human judges are sent to the comparison task claiming page, where they can see the list of datasets that is assigned to themselves and the list of datasets that they have already compared.

We assume that a wrongly translated predicate means that the entire semantic frame is incorrect; therefore, human judges are required to pick a pair of correctly translated predicate in the reference translation and the machine translation before judging the translation accuracy of the arguments associated with it. After picking a pair of matched predicates, the annotated machine translation and reference translation are shown to the human judges simultaneously. The reference translation is shown on the left and the machine translation is shown on the right. Human judges will then align each argument in the machine translation with one argument in the reference translation which expresses meaning that is closest to each other and mark the translation correctness of that argument. Figure 6 shows the comparison page when a comparison is in progress.



You are now aligning sentence 1/2

Timer
Time used for this sentence:
67s
Time used for this experiment:
67s

This area displays the alignment result, you can use this undo button to delete the latest entry.

```
0 9 null action 10 null action match
0 9 8:8 manner 10 9:9 manner correct
```

Next Sentence
End this experiment
Current frame of sentence one:
null
Current frame of sentence two:
null
Start Aligning Selected Frame Combinations
Start
Pause
A few days ago , National Development Bank successfully issued 30 billion yen of samurai bonds in Japan \'s capital
temporaltemporal temporal temporal agent agent agent manner action patient patient patient patient patient locative locative locative locative locative
Click to cancel
correct
partial
National Development Bank of Japan in the capital market successfully issued samurai bonds 30 billion yen .
agent agent agent agent agent agent agent agent agent manner action patient patient patient patient
Click to cancel

Figure 6: Semantic frame comparison web interface

5 Experiments

To assess the efficiency of the guidelines and the interface, we measured the time required by human judges to perform either the semantic frame annotation and comparison task, on two different data sets.

We also analyzed the inter-annotator agreement to show that despite of the simplicity of the annotation guidelines, the annotators are nevertheless quite consistent to each other.

Lo and Wu (2011a, b, c) have already presented state-of-the-art results in semantic MT evaluation using the proposed methodology. That is, semantic MT evaluation metrics using low-cost lay annotators for semantic frame annotation correlates with human adequacy judgement higher than automatic fluency-oriented metric, BLEU, and non-automatic expensive metric, HTER.

5.1 Setup

We had two set of data samples annotated and compared. Each sample was randomly drawn from a translation evaluation corpus containing Chinese input sentences, English reference translations, and the machine translation outputs from three different state-of-the-art systems. A set of 35 sentences drawn from the subset of the DARPA GALE program Phase 2.5 newswire evaluation dataset in which both the Chinese and English sentences have been annotated with PropBank semantic role labels. Another set of samples was drawn from the NIST MetricsMaTr meta-evaluation dataset (Callison-Burch *et al.*, 2010), with 39 sentences of the broadcast news genre. We employed Chinese-English bilinguals to annotate the semantic roles using the proposed annotation guidelines. Each translation is annotated by at least two annotators to support the consistency analysis.

5.2 Results on efficiency

The collected timing data is detailed in Table 1 in terms of sentences, frames, roles and words. The training on the annotation guidelines and briefing on the graphical user interface require typically 5 to 10 minutes of preparation, at most 15 minutes, including any necessary time for annotators or judges asking questions.

The results bear out the efficiency of our methodology, in spite of the fact that annotation was performed solely by inexpensive computer science undergraduate students with no linguistic background training. The time used for annotating semantic frames averaged about 1-1.5 minutes per sentence, depending on the complexity of the sentences—much less time than required for gold standard Propbank annotation. The time used for comparing the role fillers between the semantic frames in the reference and machine translations, similarly, averaged under 2 minutes per sentence.

Furthermore, note that these timing figures are for completely unskilled non-experts. In fact, the time required tends to decrease even further as annotators gain experience.

5.3 Results on consistency

With the easy-to-use graphical user interface, the annotations from different annotators are even more consistent

Table 1: Timing statistics for human semantic role annotation and role filler comparison tasks, for both the MetricsMaTr and GALE samples. t/s, t/f, t/r, and t/w indicate time per sentence, frame, role, word, respectively.

	#frames	#roles	#words	min t/s	max t/s	avg t/s	avg t/f	avg t/r	avg t/w
MetricsMaTr REF annotation	1.85	6.86	12.69	15.00	485.00	127.12	68.59	18.53	5.01
MetricsMaTr MT annotation	1.39	5.19	10.59	2.00	428	75.94	54.40	14.54	3.49
MetricsMaTr MT comparison	——	——	——	5.00	183	26.75	5.05	1.35	0.33
GALE REF annotation	2.79	11.07	21.44	18.00	416.00	131.30	47.13	11.71	3.06
GALE MT annotation	2.49	7.46	15.53	4.00	376	96.22	38.99	11.03	2.68
GALE MT comparison	——	——	——	9.00	401	141.33	41.61	13.10	4.89

than that reported in Lo and Wu (2011a). The IAA on role identification is 78% for reference translation and 75% for MT output. The IAA on role classification is 70% and 69% for reference translation and MT output respectively. By guiding the annotators step by step through the process of annotation, the IAA on both tasks show a 1-4% improvement from that reported in Lo and Wu (2011a). The high IAA suggests that the simple and intuitive annotation guidelines are in general sufficient for practical application such as semantic SMT and MT evaluation.

6 Web Access to the System

For research uses, please register for the full cloud based interface at `http://www.cs.ust.hk/~dekai/meant`.

7 Conclusion

We have presented a new, radically simple yet effective methodology for inexpensively annotating semantic frames using minimally trained lay annotators, that we believe to be ideal for practical semantic SMT and evaluation applications. Instead of using skilled linguists to annotate gold standard Propbank semantic frame annotation, we showed that annotating semantic frames for MT evaluation can be as intuitive as understanding the basic event structure of a sentence, which any untrained human does naturally and effortlessly. We simplified the annotation guidelines into half a page plus three annotated examples. We described a graphical user interface for both semantic frame annotation and semantic frame alignment/comparison, that guides untrained humans step by step. Restricted guidelines with this easy-to-use GUI allow untrained humans to focus on understanding the translation to provide consistent and efficient annotation and comparison. Our convenient 'cloud' based implementation of this is platform independent, installation-free, and portable as it is developed using technologies supported by any mordern web browser. Thus, we have presented in detail a semantic frame annotation and alignment methodology with minimal labor cost.

References

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. PropBank Annotation Guidelines. In *http://www.ldc.upenn.edu/Catalog/docs/LDC2011T03/propbank/english-propbank.pdf*, 2010.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Pryzbocki, and Omar Zaidan. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, 15-16 July 2010.

Jinho D. Choi, Claire Bonial, and Martha Palmer. Propbank Instance Annotation Guidelines Using a Dedicated Editor, Jubilee. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, 2010.

Chi-Kiu Lo and Dekai Wu. MEANT: An Inexpensive, High-Accuracy, Semi-Automatic Metric for Evaluating Translation Utility based on Semantic Roles. In *Proceedings of the Joint conference of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT-11)*, 2011.

Chi-Kiu Lo and Dekai Wu. SMT vs. AI redux: How semantic frames evaluate MT more accurately. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2011.

Chi-Kiu Lo and Dekai Wu. Structured vs. Flat Semantic Role Representations for Machine Translation Evaluation. In *Proceedings of the 5th Workshop on Syntax and Structure in Statistical Translation (SSST-5)*, 2011.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of the 2004 Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04)*, 2004.

Dekai Wu and Pascale Fung. Can Semantic Role Labeling Improve SMT? In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT-2009)*, pages 218–225, Barcelona, Spain, May 2009.

Dekai Wu and Pascale Fung. Semantic Roles for SMT: A Hybrid Two-Pass Model. In *Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT-09)*, pages 13–16, 2009.

Word Translation Disambiguation without Parallel Texts[*]

Erwin Marsi André Lynum Lars Bungum Björn Gambäck
Department of Computer and Information Science
NTNU, Norwegian University of Science and Technology
Sem Sælands vei 7–9, NO–7491 Trondheim, Norway
{emarsi,andrely,larsbun,gamback}@idi.ntnu.no

Abstract

Word Translation Disambiguation means to select the best translation(s) given a source word in context and a set of target candidates. Two approaches to determining similarity between input and sample context are presented, using n-gram and vector space models with huge annotated monolingual corpora as main knowledge source, rather than relying on large parallel corpora. Experiments on SemEval's Cross-Lingual Word Sense Disambiguation task (2010 English→German part) show some models on average surpassing the baselines, suggesting that translation disambiguation without parallel texts is feasible.

Index Terms: word sense disambiguation, vector space models, n-gram language models

1 Introduction

One of the challenges in translating a word is that, according to a translation dictionary or some other translation model, a source language word normally has several translations in the target language. For instance, the English word *plant* may be translated as the German word *Fabrik* in the context of industry, but as *Pflanze* in the context of nature. Hence contextual information is required to resolve ambiguities in word translation. This task is known as Word Translation Disambiguation (WTD).

The currently predominant paradigm for data-driven machine translation is phrase-based statistical machine translation. In phrase-based MT the task of WTD is not explicitly addressed, but instead the influence of context on word translation probabilities is implicitly encoded in the model, both in the phrasal translation pairs learned from parallel text and stored in the phrase translation table (collocating words in the immediate context of an ambiguous source word are likely to end up together in a translation phrase, thus helping to disambiguate possible translations candidates) and in the target language model (usually n-gram models which tend to prefer collocations and other local dependencies).

One potential problem with this approach is that the amount of context taken into account is rather small. It is clear that word translation disambiguation often depends on cues from a wider textual context, for instance, elsewhere in the same sentence, paragraph or the document as a whole. This is beyond the scope of most phrase-based SMT approaches, which work with relatively small phrases. Another drawback of phrase-based MT (and of most data-driven MT approaches) is dependence on large aligned parallel text corpora for training purposes, a both scarce and expensive resource.

The work described here has been carried out in the context of the project PRESEMT (Pattern REcognition-based Statistically Enhanced MT; www.presemt.eu) which emphasises flexibility and adaptability towards new language pairs. A key part is to avoid relying on large and expensive parallel corpora, as such corpora are not available for the majority of language pairs; but to instead rely on very small purpose-built parallel corpora, widely available linguistic resources such as bilingual dic-

[*] This research has received funding from the European Community's 7th Framework Programme under contract nr 248307 (PRESEMT). Thanks to Els Lefever for responding to questions and request regarding the CL-WSD data sets.

tionaries, and huge monolingual corpora that can for example be easily mined from the web and automatically annotated with existing resources such as POS taggers. This combination of linguistically oriented resources and large corpora makes the system a hybrid MT system, combining data driven approaches and linguistic resources.

The next section details the word translation disambiguation task and introduces the data sets and evaluation measures used. Sections 3 and 4 then describe the n-gram and vector space modelling, respectively, followed by the experimental setup and ways to transform the vector space in Section 5. The actual experimental results are given in Section 6. Section 7 sets the work in context of efforts by others, before Section 8 discusses the results.

2 Task and data

The task addressed in this work is correctly translating a single word in context, or more formally:

Word Translation Disambiguation (WTD) *Given a source word in its context (e.g., a sentence) and a set of target word candidates (e.g., from a bilingual dictionary), the task of Word Translation Disambiguation is to select the best translation(s).*

This is akin to word glossing or word-for-word translation provided that all translation candidates can be retrieved from a bilingual dictionary. WTD can thus be regarded as a ranking and filtering task. It is different, however, from full word translation, because it is assumed that all possible translations are given in advance, which is not the case in the more general task of full word translation. Full word translation can be regarded as a two-step process: (1) generation of word translation candidates, (2) word translation disambiguation. Any solution to WTD would partly solve full word translation and is therefore worthwhile to pursue.

This paper describes two approaches to WTD: First, n-gram language modelling where a surface representation of the Target Language (TL) sentence is constructed and the paths through these contexts are scored by the model. Second, vector space modelling using similarity based on the lexical semantics of the TL context to rank translation candidates according to semantic distance of the content.

AGREEMENT in the form of an exchange of letters between the European Economic Community and the Bank for International Settlements concerning the mobilization of claims held by the Member States under the medium-term financial assistance arrangements
{bank 4; bankengesellschaft 1; kreditinstitut 1; zentralbank 1; finanzinstitut 1}

1) The Office shall maintain an electronic data bank with the particulars of applications for registration of trade marks and entries in the Register. The Office may also make available the contents of this data **bank** *on CD-ROM or in any other machine-readable form.*
{datenbank 4; bank 3; datenbanksystem 1; daten 1}

(b) established as a band of 1 km in width from the banks of a river or the shores of a lake or coast for a length of at least 3 km.
{ufer 4; flussufer 3}

Table 1: Some contexts for the English word *bank* with possible German translations in the CL-WSD trial data

2.1 Data

There is a recent data set well suited for evaluating WTD systems. The 2010 exercises on Semantic Evaluation (SemEval-2) featured a Cross-Lingual Word Sense Disambiguation (CL-WSD) task (Lefever and Hoste, 2010) based on the English Lexical Substitution task from SemEval-2007. There systems had to find an alternative (synonym) substitute word or phrase for a target word in its context (McCarthy and Navigli, 2007). The CL-WSD task basically extends lexical substitution across languages, i.e., instead of finding substitutes for a word in the same language, its possible translations in another language have to be found. Although originally conceived in the context of word sense disambiguation, it is a word translation task.

While the source language in the CL-WSD data is English, there are five target languages: Dutch, French, Spanish, Italian and German. The trial set consists of 5 nouns (20 sentence contexts per noun, 100 instances in total per language), and the test set of 20 nouns (50 sentence contexts per noun, 1000 instances in total per language). Table 1 provides examples of contexts for the English word *bank* and its possible German translations from trial data.

The CL-WSD data sets were constructed in a two-step process. First, a "sense inventory" of all possibank, bankanleihe, bankanstalt, bankdarlehen, bankengesellschaft, bankensektor, bankfeiertag, bankgesellschaft, bankinstitut, bankkonto, bankkredit, banknote, blutbank, daten, datenbank, datenbanksystem, euro-banknote, feiertag, finanzinstitut, flussufer, geheimkonto, geldschein, geschäftsbank, handelsbank, konto, kredit, kreditinstitut, nationalbank, notenbank, sparkasse, sparkassenverband, ufer, weltbank, weltbankgeber, west-bank, westbank, westjordanien, westjordanland, westjordanufer, westufer, zentralbank

Table 2: All German translation candidates for English *bank* as extracted from the CL-WSD trial gold standard

ble translations of a given source word was created, based on the Europarl corpus (Koehn, 2005), where alignments involving the relevant source words were manually checked. The corresponding target words were manually lemmatised and clustered into translations with a similar sense. Second, trial and test data were extracted from two independent corpora (JRC-ACQUIS and BNC). For each source word, four human translators picked the contextually appropriate sense cluster and chose up to three preferred translations it. Translations are thus restricted to those appearing in Europarl, probably introducing a slight domain bias. Each translation has an associated count indicating how many annotators considered it adequate in the given context. The spread of this count varies widely between different sentences, ranging from reasonably tight agreements on one or two candidates (with some other receiving a few votes) to sentences annotated with a long list of candidates (most receiving only one vote).

It is important to understand that the work in this paper addresses only part of the CL-WSD task: since the focus here is on WTD, it can be assumed that a perfect solution to finding translation candidates already exists. In practice this is accomplished by extracting all possible translations from the gold standard; e.g., for the English lemma *bank*, all translation candidates occurring in the trial gold standard for German are listed in Table 2.

2.2 Evaluation measures

The CL-WSD shared task employed two evaluation measures: the Best and Out-Of-Five scores (Lefever and Hoste, 2010). The Best criterion is intended to measure how well the system succeeds in delivering the best translation, i.e., the one preferred by the majority of annotators. The Out-Of-Five (OOF) criterion measures how well the top five candidates from the system match the top five translations in the gold standard. However, in WTD experiments, the Best measure has some deficiencies, most importantly that it is not normalized between 0 and 1. This results in a very uneven spread of scores, both among different target words and among the individual test sentences for each word, making it difficult — or not even meaningful — to judge differences in system performance by looking at average scores. Hence rather than using the original Best score, we adopt the normalized variant proposed by Jabbari et al. (2010), here referred to as Best$_{JHG}$.

For each sentence t_i, let H_i denote the set of human translations. For each t_i there is a function freq_i returning the count of how many annotators chose it for each term in H_i and a value $\mathit{maxfreq}_i$ for the maximum count. The pairing of H_i and freq_i constitutes a multiset representation of the human answer set. Let $|S|^i$ denote the multiset cardinality of S according to freq_i, i.e., $\sum_{a \in S} \mathit{freq}_i(a)$, the sum of all counts in S. For the first example in Table 1: $H_1 = \{$bank, bankengesellschaft, kreditinstitut, zentralbank, finanzinstitut$\}$; $\mathit{freq}_1(\text{bankengesellschaft}) = 4$, $\mathit{freq}_1(\text{bank}) = 1$, etc; $\mathit{maxfreq}_1 = 4$; and $|H_1|^1 = 8$.

The Best$_{JHG}$ measure is defined as follows

$$\text{Best}_{JHG}(i) = \frac{\sum_{a \in A_i} \mathit{freq}_i(a)}{\mathit{maxfreq}_i \times |A_i|} \qquad (1)$$

where A_i is the set of translations for test item i produced by the system. The optimal score of 1.0 is achieved by returning a single translation whose count is $\mathit{maxfreq}_i$, with proportionally lesser credit given to answers in H_i with smaller counts. In principle a system can output several candidates in order to "hedge its bets", but there is a penalty for non-optimal translations, so the best strategy appears to be to output just one. The systems in our experiment always produced a single translation for the Best$_{JHG}$ score, so $|A_i| = 1$ always. In the first example of Table 1, the system output $A_1 = \{\text{bank}\}$ would give $\text{Best}_{JHG}(1) = 1.0$ whereas $A_1 = \{\text{bankengesellschaft}\}$ would give $\text{Best}_{JHG}(1) = 0.25$ and $A_1 = \{\text{ufer}\}$ would give $\text{Best}_{JHG}(1) = 0.0$.

The Out-Of-Five (OOF) criterion is defined as:

$$OOF(i) = \frac{\sum_{a \in A_i} freq_i(a)}{|H_i|^i} \quad (2)$$

In this case systems are allowed to submit up to five candidates of equal rank. It is a recall-oriented measure with no additional penalty for precision errors, so there is no benefit in outputting less than five candidates. With respect to the previous example from Table 1, the maximum score is obtained by system output $A_1 = \{$bank, bankengesellschaft, kreditinstitut, zentralbank, finanzinstitut$\}$, which gives $OOF(1) = (4+1+1+1+1)/8 = 1$, whereas $A_1 = \{$bank, bankengesellschaft, nationalbank, notenbank, sparkasse$\}$ would give $OOF(1) = (4+1)/8 = 0.625$. One remaining problem with the OOF measure is that the maximum score is not always one, i.e. not normalized, because sometimes the gold standard contains more than five translation alternatives.

For assessing overall system performance, the average of Best$_{JHG}$ or OOF scores across all test items for a single source word is taken. In addition, the CL-WSD task employed a "mode" variant of both scores. These were not used in the evaluations for reasons explained by Jabbari et al. (2010). All experiments use TL context to rank translation candidates for a given word in the source sentence, but for the SemEval CL-WSD data the target language sentence is not given, which means that a suitable context has to be constructed in order to perform disambiguation. This is done by collecting all translation candidates for all words in the sentence. These translation candidates are put in a bag of words from which the words' appropriate feature vectors are constructed.

3 N-gram models

Utilising n-gram language models (LMs) to rank target contexts is motivated by their widespread use and that a naive approach to order translation candidates (TC) is a useful comparison for other models. The advantage of n-gram modelling is its conceptual simplicity and practical availability. Only one model is needed to process all trial and test words.

Adapted to the WTD task, an LM can predict the likelihood of a target context being part of the language. TC sentences are constructed by combining each TC with every possible translation of their context. The shortest TC sentence is the TC itself, and if the LM is queried for all TC candidates, the most frequent would turn out on top. For the English *bank*, the most likely German candidate is *Bank*. The n-gram model should rank TC sentences of the right sense higher, because co-located phrases like *the West Bank* and *Gaza Strip* are reflected in higher n-gram probabilities of their corresponding TC sentences. This applies when the n-gram model finds the TC with the content-bearing word in the right place (when word-to-word translation is correct), unlike for multi word expressions with different surface forms in German and English.

The LM was built from sentence-separated lemmatised parts of deWac, a large monolingual web corpus of German containing over 1,627M tokens (Baroni and Kilgarriff, 2006). For each TL context, a huge number of n-grams to query the model were compiled. With a 5-gram model, a possible 4 words preceding and succeeding the word to be translated could be tested. The results of various context lengths were kept in a 2-dimensional matrix, where each index represents words ahead of, and after the TC word. Results from different context lengths are extracted, until enough TC are found (often 5). If the [-4,1] entry (4 words before, 0 after) is ranked highest, the TC represented by these n-grams would be used exclusively in output, if the limit was reached. If not, the algorithm moves on to the next matrix entry. Because of the naive word-by-word translation, few n-gram candidates of higher order were found. Ranking by no surrounding context leads to the same answer for all instances of the word, with the most frequent TL sense first.

4 Vector space modelling

A simple idea underlies the approach to WTD: given a source word in context and a number of translation candidates, search in a large TL corpus for context samples exemplifying the translation candidates. Thus, given the English word *bank* and its possible German translations *Bank, Datenbank, Ufer, ...* retrieve sentences containing *Bank*, those containing *Datenbank*, those containing *Ufer*, etc. Next search these context samples for the one most similar to the given source word context. The best TC is the one associated with this context sample.

Two basic issues need to be addressed in this approach. First, matching a given context in the source language against any context samples in the TL is obviously complicated by the difference in language. We take the straight forward approach of carrying out a word-by-word translation of the source context by means of a translation dictionary. However, there are alternative solutions to this issue conceivable, e.g., by using an existing MT system for translating the source context, or by translating the TL contexts to the source language instead.

The second issue is how to measure similarity of textual contexts, a key issue in many language processing tasks. Numerous approaches have been proposed, ranging from simple measures for word overlap and approximate string matching (Navarro, 2001), through WordNet-based and corpus-based measures (Mihalcea et al., 2006), to elaborate combinations of deep semantic analysis, word nets, domains ontologies, background knowledge and inference (Androutsopoulos and Malakasiotis, 2010). The approach to similarity taken here is that of Vector Space Models (VSM) for words (Salton, 1989). These models are based on the assumption that the meaning of a word can be inferred from its usage, i.e., distribution in text (Harris, 1954): words with similar meaning tend to occur in similar contexts.

Vector space models for words are created as high-dimensional vector representations through a statistical analysis of the contexts in which words occur. Similarity between words is defined as similarity between their context vectors in terms of some vector similarity measure, e.g., cosine similarity. A major advantage of this approach is the balance of reasonably good results with a simple model. In addition, it does not require any external knowledge resources besides a large text corpus and is fully unsupervised (human annotations are not needed).

Vector space modelling is applied to disambiguation as follows: first training and test instances are converted to feature vectors in a common multidimensional vector space. Next this vector space is reshaped by applying one or more transformations. The motivation for a transformation can be, e.g., to reduce dimensionality, to reduce data sparseness, to promote generalization or to possibly induce latent dimensions. Finally, for each of the vectors in the test corpus, the N most similar vectors are retrieved from the training corpus using cosine similarity, and translation candidates are predicted from the target words associated with these vectors.

5 Experimental setup

The preliminary experiments in this paper cover the German part of the CL-WSD trial data, i.e., 5 nouns with 20 sentence contexts per noun, 100 instances. We intend to run experiments on the larger CL-WSD test data set, as well as on other language pairs, once our WTD approach has sufficiently stabilized on a couple of successful models. Since the CL-WSD task offers no training data, a *training corpus* was constructed in the following steps:

Context sampling: For each translation candidate of a source word, examples of its use in context were obtained. Up to 5000 contexts per translation candidate were sampled from deWac through the web API of the SketchEngine (Kilgarriff et al., 2004). Sentences containing more than 75 tokens were skipped.

Linguistic processing: Context sentences were tokenized, lemmatised and part-of-speech tagged using the TreeTagger for German (Schmid, 1994).

Vocabulary creation: A vocabulary of terms was created over all samples sentences for all translation candidates of a single source word. First, stop words were removed according to a list of 134 German stop words. Next, function words were removed based on the POS tag, leaving mostly content words. Regular expressions were used for removing ill-formed tokens. Finally, frequency-based filtering was applied, removing all terms occurring less than 10 times, and terms occurring in more than 5% of the samples.

Vector encoding: Each context sample was encoded as a labeled (sparse) feature vector, where the features are the vocabulary terms and the feature values are the counts of these terms in the context sample at hand. The vector was labeled with the translation candidate it is a sample of. All vectors for all translation candidates of a single source word were collected in a (sparse) matrix.

The CL-WSD trial data was processed in a similar way to obtain a *test corpus*, with preprocessing carried out by the TreeTagger for English (Schmid, 1994). The test sentences were then translated word-for-word by look-up of the lemma plus POS combination in an English-German dictionary with over 900K entries obtained by reversing an existing German-English dictionary. If multiple translations for an English word were found, all were included in the sentence translation. Finally, the test sentence translations were encoded as (sparse) feature vectors in the same way as the training contexts, using the same vocabulary. As a result all German translations outside of the vocabulary were effectively deleted.

The vector space models were implemented in Gensim (Řehůřek and Sojka, 2010), an efficient VSM framework in Python. It provides a number of models for transforming vector space. In addition we implemented the Summation and PMI models. The following transformations were evaluated:

Bare vector space model. Does not apply any transformation to the feature space.

Term Frequency*Inverse document frequency (Jones, 1972) effectively gives more weight to terms that are frequent in the context but do not occur in many other contexts.

Pointwise Mutual Information (Church and Hanks, 1990) measures the association between translations candidates and context terms, and should give higher weight to terms with more discriminative power.

Latent Semantic Indexing reduces the dimensionality of the vector space by applying a Singular Value Decompostion (Deerwester et al., 1990). It is claimed to model the latent semantic relations between terms and address problems of synonymy and polysemy, hence increasing similarity between conceptually similar context vectors, even if those vectors have few terms in common.

Random Projection (also called Random Indexing). Another way to reduce the dimensionality of the vector space by projecting the original vectors into a space of nearly orthogonal random vectors. RP is claimed to result in substantially smaller matrices and faster retrieval without significant loss in performance (Sahlgren and Karlgren, 2005).

Summation model. Sums all context vectors for the same translation candidate, resulting in a centroid vector for each translation candidate. It is attractive from a computational point of view because the resulting matrix is relatively small.

For each of the 20 vectors in the test corpus for a English word, the training corpus is searched for the most similar vectors and the associated labels provide the German translations. Cosine similarity is used to calculate vector similarity. For scoring on the Best$_{JHG}$ measure, we use the single best matching vector in the training corpus. For scoring OOF, first the n best matching vectors are retrieved ($n = 1000$ in the experiments). Next the cosine similarities of all vectors with the same label are summed and the five labels with the highest summed cosine similarity constitute the output.

6 Results

Two baselines were employed. The first baseline (MostFrequentBaseline) does not rely on parallel corpora. It consists of simply selecting the translation candidate whose lemma occurs most frequently in the deWaC corpus. It therefore completely ignores the context of the words. This results in low scores on the Best$_{JHG}$ measure, although the OOF scores for bank and occupation are high. The low scores may be due to differences between predominant translations in Europarl and in deWaC. Another factor which may reduce the efficiency of target side frequencies is that the word counts can be "polluted" because a certain German word is also the translation of another very frequent English word, a problem discussed by Koehn and Knight (2000).

The second baseline (MostFrequentlyAligned) does rely on parallel corpora and was also used in the CL-WSD shared task. It is constructed by taking the translation candidate most frequently aligned to the source word in the Europarl corpus with manually corrected source word alignments. As expected, the Best$_{JHG}$ scores are consistently much higher than those of the first baseline. However, this is not so with regard to the OOF scores, which are lower than the first baseline for bank and occupation.

The simple n-gram model was employed in three different orders, uni- tri and pentagram models, but without exploring all possible priorities of context lengths (skewing to before- or after context). On average the higher-order models performed better.

	Bank	*Movement*	*Occupation*	*Passage*	*Plant*	Mean
RP (300)	15.83	17.50	11.25	5.42	20.00	14.00
LSI (200)	30.42	11.25	21.25	9.17	20.42	18.50
SumModel	**43.75**	17.50	**37.92**	7.92	**43.75**	**30.17**
PMI	32.08	**21.25**	26.67	2.92	38.33	24.25
TF*IDF	20.00	11.67	35.83	3.33	23.33	18.83
BareVSM	28.33	10.00	37.08	9.58	17.08	20.42
5-gram model	25.00	12.92	27.08	14.17	15.42	18.92
3-gram-model	10.00	16.67	24.17	11.67	6.67	13.84
1-gram-model	42.50	5.00	2.50	1.67	3.33	11.00
MostFreqAlignBaseline	6.25	19.17	35.83	**15.00**	40.00	23.25
MostFreqBaseline	1.25	5.00	2.50	1.67	10.26	4.14

Table 3: Best$_{JHG}$ scores for different models (underlined=above both baselines; bold=highest)

Results for different models in terms of the Best$_{JHG}$ score and Out-of-five scores are listed in Table 3 and Table 4. Regarding system scores, several general observations can be made. To begin with, the scores on passage tend to be lower than those on bank, occupation and plant. To a lesser extent, the same holds for scores on movement, keeping in mind that max OOF score on movement is also lower. Seemingly no correlation with the number of translation candidates though, as passage has 42 whereas bank and plant have 40 and 60 respectively. Furthermore, even though most models often outperform both baselines on some words, there is no model that consistently outperforms both baselines on all five words, although the SumModel comes close, it has a problem with passage. Looking at the mean scores over all five words, however, the SumModel outperforms both baselines. This is a promising result considering that model is smallest and does not rely on parallel text.

In a similar vein, no model consistently outperforms all others. For instance, even though SumModel yields high OOF scores on four out of five words, PMI scores higher on plant. LSI seems to provide no improvements over the BareVSM. RP performed badly, which may be related to implementation issues. TF*IDF seems to give slightly worse results in comparison to BareVSM. A possible explanation is that its feature weighting is unrelated to vector labels, so it may actually reduce the weight of discriminative context words. PMI, which does take the vector label into account, gives a slight improvement over BareVSM on the Best$_{JHG}$ score.

7 Related work

Koehn and Knight compare different methods to train word-level translation models for German-to-English translation of nouns, three of which also rely on a translation dictionary in combination with monolingual corpora (Koehn and Knight, 2000; Koehn and Knight, 2001). The first is identical to our MostFrequent baseline, the second uses a target LM to pick the most probable word sequence, and the third relies on monolingual source and target language corpora in combination with the Expectation Maximization (EM) algorithm to learn word translation probabilities. Performance of the latter two is reported to be comparable to that of using a standard SMT model trained on a parallel corpus. Our SVM approach is different in that it models a much larger contexts, i.e., full sentences. Similarly, Monz and Dorr (2005) employ an iterative procedure based on EM to estimate word translation probabilities. However, rather than relying on an n-gram LM, they measure association strength between pairs of target words, which they claim is less sensitive to word order and adjacency, and therefore data sparseness, than higher n-gram models. Their evaluation is only indirect as application of the method in a cross-lingual IR setting.

Rapp proposes methods for extracting word translations from unrelated monolingual corpora, based on the idea that words that frequently co-occur in the source language also have translations that frequently co-occur in the target language (Rapp, 1995; Rapp, 1999). His use of distributional similarity between translations in the form of a vector space is

	Bank	*Movement*	*Occupation*	*Passage*	*Plant*	Mean
MaxScore	95.60	82.62	93.58	89.57	83.22	88.92
RP (300)	24.80	12.65	22.70	8.82	21.63	18.12
LSI (200)	47.07	12.61	35.40	17.03	35.61	29.54
SumModel	**52.59**	**28.01**	42.03	17.72	32.54	**34.58**
PMI	41.00	16.33	38.41	15.47	**38.52**	29.95
TF*IDF	37.76	12.31	27.72	12.16	25.00	22.99
BareVSM	47.88	13.86	40.83	14.60	28.33	29.10
5-gram model	31.75	23.01	**37.73**	15.06	26.55	26.82
3-gram model	27.14	23.01	36.81	17.70	22.16	25.42
1-gram-model	22,92	14.17	24.39	6.63	20.04	17.63
MostFreqAlignBaseline	23.23	20.34	32.78	**27.25**	21.06	24.93
MostFreqBaseline	31.69	14.17	40.02	6.63	20.04	22.51

Table 4: Out-of-five (OOF) scores for different models (underlined=above both baselines; bold=highest)

similar to our approach. However, his goal is to bootstrap a bilingual lexicon, whereas our goal is to disambiguate. As a result, Rapp's input consists of a source word in isolation for which contexts are retrieved from a source language corpus, while our input consists of a source word in a particular context. Other work on lexical bootstrapping from monolingual corpora inspired by Rapp's work include Fung and Yee (1998) and Fung and McKeown (1997).

The submissions to the SemEval 2010 CL-WSD workshop presented a number of relevant approaches to the WTD task (van Gompel, 2010; Silberer and Ponzetto, 2010; Vilariño Ayala et al., 2010). All submitted systems, however, relied on using parallel text. Still most systems were unable to outperform the MostFrequentlyAligned baseline. Something our systems do, but a direct comparison is not fair because we only address the subtask of disambiguation and not the task of finding translation candidates.

8 Discussion and conclusion

While it is hard to draw a general conclusion on the basis of these preliminary experiments, it is our experience that it is difficult to find an approach that generalises well over any word or context for the WTD task. In our experiments, increases in performance for one set of target words were generally accompanied by reduction in performance for other words. This leads one to speculate that there are hidden variables governing the disambiguation behaviour of words such that a classification of words according to such hidden variables yield a more evenly distributed performance increase. For n-gram models the expected improvement in performance with higher-order models is observed.

In sentence space we have explored re-sampling subsets of the sentences and combining all sentences by summing all the matrix rows (sum). Attempts to cluster the sentences through for k-means and within-between cluster distances have largely been unsuccessful. Plans for future work include evaluation of the best models on the CL-WSD test data set and in the context of the full PRESEMT system.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, May.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 87–90, Trento, Italy, April. ACL.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 414–420, Morristown, NJ, USA. ACL.

Zellig Harris. 1954. Distributional structure. *Word*, 10:146–162. Reprinted in Z. Harris, *Papers in Structural and Transformational Linguistics*, Reidel, Dordrecht, Holland 1970.

Sanaz Jabbari, Mark Hepple, and Louise Guthrie. 2010. Evaluation metrics for the lexical substitution task. In *Proceedings of the 2010 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 289–292, Los Angeles, California, June. ACL.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of Euralex*, pages 105–116, Lorient, France, July.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, pages 711–715. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Philipp Koehn and Kevin Knight. 2001. Knowledge sources for word-level translation models. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 27–35.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand, September.

Els Lefever and Véronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden, July. ACL.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. ACL.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21th National Conference on Artifical Intelligence*, Boston, Massachusetts, July. AAAI.

Christof Monz and Bonnie J. Dorr. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th International Conference on Research and Development in Information Retrieval*, pages 520–527, Salvador, Brazil, August. ACM SIGIR.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, March.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, MIT, Cambridge, Massachusetts, June. ACL.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, Madrid, Spain, July. ACL.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 45–50, Valetta, Malta, May. ELRA. Workshop on New Challenges for NLP Frameworks.

Magnus Sahlgren and Jussi Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(2), June. Special Issue on Parallel Texts.

Gerard Salton. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, Reading, Massachusetts.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the 1st International Conference on New Methods in Natural Language Processing*, pages 44–49, University of Manchester Institute of Science and Technology, Manchester, England, September.

Carina Silberer and Simone Paolo Ponzetto. 2010. UHD: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 134–137, Uppsala, Sweden, July. ACL.

Maarten van Gompel. 2010. UvT-WSD1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 238–241, Uppsala, Sweden, July. ACL.

Darnes Vilariño Ayala, Carlos Balderas Posada, David Eduardo Pinto Avendaño, Miguel Rodríguez Hernández, and Saul León Silverio. 2010. FCC: Modeling probabilities with GIZA++ for Task 2 and 3 of SemEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 112–116, Uppsala, Sweden, July. ACL.

A New Hybrid Machine Translation Approach Using Cross-Language Information Retrieval and Only Target Text Corpora

Nasredine Semmar
CEA, LIST, Vision and Content Engineering Laboratory
18 route du Panorama
Fontenay-aux-Roses, F-92265, France
nasredine.semmar@cea.fr

Dhouha Bouamor
CEA, LIST, Vision and Content Engineering Laboratory
18 route du Panorama
Fontenay-aux-Roses, F-92265, France
dhouha.bouamor@cea.fr

Abstract

Parallel corpora play a vital role in Statistical Machine Translation. Non-availability of these corpora is a major barrier for adding new languages pairs. In this paper, we propose a new hybrid approach for English-French machine translation combining a cross-language search engine and a statistical language model trained from a monolingual corpus. The cross-language search engine returns the translation candidates ordered by their relevance and the language model of the target language is used to disambiguate the translation. This approach has been evaluated and compared to Moses. We used 100000 French sentences of the Europarl corpus to train the language model, 1103 English-French sentences of the Arcade-II corpus as the translation reference and the BLEU score. The obtained scores are 21.33% for our approach and 21.45% for Moses. The experimental results also showed that our approach provides better translation performance in terms of grammatical coherence.

1 Introduction

Parallel corpora play a vital role for training translation models in Statistical Machine Translation (SMT). Non-availability of these corpora, morphology and syntactic structure differences between source and target languages are the major challenges for adding new languages pairs for SMT engines. We present, in this paper, a new hybrid approach for machine translation which uses only a monolingual corpus in the target language. This approach is based on a cross-language search engine which returns for each sentence to translate a set of translation candidates extracted from the monolingual corpus already indexed. A statistical language model is then used to identify the correct translation.

This paper is organized as follows. In section 2, some related work is presented. Section 3 describes the implementation of our hybrid machine translation approach. In section 4, some experimental results are reported and discussed. Section 5 concludes our study and presents our future work.

2 Related Work

There are two main approaches for machine translation (Trujillo, 1999) (Hutchins, 2005):

- Rule-based approaches.
- Corpus-based approaches.

The rule-based approaches regroup word-to-word translation, syntactic translation with transfer rules and interlingua which uses an intermediate semantico-syntactic representation to generate translations into any target language.

The corpus-based machine translation approaches use statistics and probability calculation in order to identify equivalences between texts in the corpus (Koehn, 2010). This probability calculation depends on two measures. The first is the probability that the words in the target language are translations of the words in the source language (translation model). The second is the probability that these words are correctly combined in the target language (language model). Probability that a given word in the target text is a translation of a given word in the source text is calculated on the basis of a sentence-aligned parallel corpus. The language model consists of probabilities of sequences of words based on a monolingual corpus in the target language.

Rule-based approaches require manual development of bilingual lexicons and linguistic rules, which can be costly, and which often do not generalize to other languages. Corpus-based approaches are effective only when large amounts of parallel text corpora are available.

Hybrid approaches combine the strengths of rule-based and corpus-based machine translation strategies (Somers, 2005). (Koehn et al. 2010) presented an extension of the state-of-the-art phrase-based statistical machine translation models in order to integrate additional linguistic information such as lemmas, part-of-speech, and morphological properties of words. The authors reported that experiments showed gains over standard phrase-based models, both in terms of automatic scores (gains of up to 2% BLEU), as well as a measure of grammatical coherence.

Our hybrid approach for machine translation is based on a new paradigm which consists in using a cross-language search engine to extract translated texts from a monolingual corpus and combining linguistic information with a statistical language model in order to generate the correct translation.

3 Machine Translation Based on Cross-language Information Retrieval

Cross-language information retrieval consists in providing a query in one language and searching documents in different languages (Grefenstette, 1997), and the goal of machine translation is to produce for each sentence in the source language its equivalent in the target language. Cross-language information retrieval using linguistic analysis for indexing and interrogation and rule-based machine translation are closely related domains. Both use bilingual lexicons and automatic text analysis.

The machine translation prototype implementing our approach is composed of two modules: A cross-language search engine and a text generator (Figure 1):

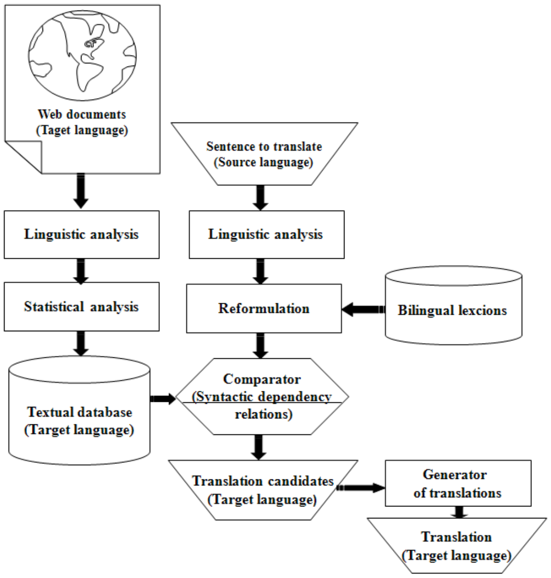


Figure 1: Machine translation using cross-language information retrieval

3.1 Cross-language Information Retrieval

The cross-language search engine (Semmar et al., 2006) is used to provide a collection of sentences in the target language. These sentences are considered are translation candidates. The search engine uses a weighted Boolean model, in which sentences in the target language are grouped into classes characterized by the same set of concepts composed of words. This search engine is composed of a multilingual analyzer, a statistical analyzer, a reformulator and a comparator.

Multilingual Analysis

The multilingual analysis is built using a traditional architecture (LIMA) (Besançon et al., 2010) and includes a morphological analyzer, a part-of-speech tagger and a syntactic analyzer. The linguistic analyzer produces a set of normalized lemmas, a set of named entities and a set of dependency relations between words.

Statistical Analysis

The role of the statistical analysis is to attribute to each word or a compound word a weight according to the information it provides to choose the target sentences relevant to the sentence to translate. The weight is maximum for words appearing in one single sentence and minimum for words appearing in all the sentences. This weight is used by the comparator to compare intersection between the sentence to translate and indexed sentences. Our search engine uses a weighted Boolean model, in which sentences are grouped into classes characterized by the same set of concepts. The classes constitute a discrete partition of the database. For example, if the sentence to translate is "*nuclear waste*" on a database containing only sentences on nuclear plants, the statistical model indicates that sentences containing the compound word "*nuclear waste*" are more relevant than sentences containing the words "*nuclear*" and "*waste*". Sentences containing the words "*nuclear*" and "*waste*" are more relevant than sentences containing only the word "*waste*".

Query Reformulation

Reformulation consists in inferring new words from the original query (sentence to translate) words according to lexical and semantic knowledge (synonyms, etc.). The reformulation can be used to increase the quality of the retrieval in a monolingual interrogation (Debili, 1989). It can also be used to infer words in other languages. The query terms are translated using bilingual dictionaries. Each term of the query is translated into several terms in target language. The translated words form the search terms of the reformulated query. The links between the search terms and the query concepts can also be weighted by a confidence value indicating the relevance of the translation. Reformulation can be achieved on the word or on the word with a specific part of speech and can also be used to transform the syntactic structure of the sentence to translate into the target language. This reformulation uses an English-French bilingual lexicon composed of 220000 entries to translate words, and a set of rules to transform syntactic structures from the source language to the target language.

Comparison of the sentence to translate with indexed sentences

The comparator computes intersections between words and the syntactic structure of the sentence to translate and words and syntactic structures of the indexed sentences. This comparator provides a relevance weight for each intersection and returns the translation candidates. These translation candidates could be sub-sentences composed of only some words corresponding to the translation of just a part of the sentence to translate. Linguistic information such as lemmas, grammatical categories, gender, number and syntactic dependency relations are associated with the words of the translation candidates.

3.2 Text Generation

Our text generation approach is based on a syntactic analysis. This approach consists, on the one hand, in composing the sub-sentences returned by the comparator of the cross-language search engine in order to build a dependency syntactic structure in the target language which covers the sentence to translate, and, on the other hand, in producing a correct sentence in the target language by using the syntactic structure of the translation candidate.

The text generator is composed of two modules: a reformulator and a flexor. The reformulator uses the parts of sentences to match the translation hypothesis. Some linguistic rules are used to assemble the new hypothesis in a lattice of translations. This lattice contains linguistic information for each word of the translation. A statistical model is learned on a monolingual lemmatized corpus which contains linguistic information. This model scores the lattice in order to find the best syntactic hypothesis in the target language. The lattice is implemented by using the AT&T FSM toolkit (Mohri et al., 2002). The language model is learned with the CRF++ toolkit (Kudo and Matsumoto, 2001). The flexor transforms the lemmas of the target language sentence into plain words. We use the linguistic information returned by the cross-language search engine to produce the right form of the lemma. This flexor consists in transforming the lemma of a word into the surface form of this word by using the grammatical category, the gender and the number of the word. For example, the lemma "*avoir*" (verb) in present simple and third person singular will be transformed into the form "*a*". Sometimes, we obtain several forms for the same lemma. To disambiguate, we use a statistical language model based on CRF that has been previously trained on a monolingual corpus. This disambiguation provides the right flexion of the lemma and therefore the best translation.

4 Experiment Results and Discussion

To evaluate the performance of our machine translation approach, we indexed the first 100000 French sentences of the Europarl[1] corpus and we used a subset of Arcade-II[2] corpus composed of 1103 sentences in English and French as the translation reference. In order to compare the translation results of our approach with the results of the open source baseline system Moses, we used the same Europal bilingual corpus composed of the first 100000 sentences in English and French to train the language and translation models and we considered the same 1103 sentences of Arcade-II as a test corpus. We also considered that there is only one reference per test sentence and we used the BLEU score to evaluate the translation quality of the two systems. Our translation approach obtained a score of 21.33% and Moses obtained a score of 21.45%. These two scores are very close and are satisfactory taking into account that only 100000 sentences are used to train these two systems.

In order to show the relevance of using a deep linguistic analysis in machine translation, we used Google Translate[3] to translate into French the sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital.*". Google Translate proposes the translation "*Administrations de sécurité sociale en Grèce sont appelant à l'indépendance à l'égard de l'investissement de capitaux.*". Thereby, the compound word "*Social security funds*" has been translated by the compound word "*Administrations de sécurité sociale*" and the expression "*are calling for*" has been translated as "*sont appelant*".

As we can see, our translation prototype proposes the compound word "*fonds de la sécurité sociale*" as a translation for the compound word "*Social security funds*" and the expression "*appellent à*" as a translation for the expression "*are calling for*". These translations are better than those provided by Google Translate.

Table 1 shows the translation results ordered by their relevance given by our machine translation approach for the English sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital.*".

Relevance	Translation candidate
1	les fonds de la sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des capitaux.
2	les fonds de sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des capitaux.
3	les fonds de la sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des fonds.
4	les fonds de sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des fonds.
5	les fonds de le sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des capitaux.

Table 1: The first five translations returned for the English sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital.*"

5 Conclusion and Future Work

This paper proposed a new hybrid approach for English-French machine translation combining a cross-language search engine and a statistical language model trained from a monolingual

[1] The Europarl parallel corpus is available on http://www.statmt.org/europarl.

[2] The Arcade-II parallel corpus was produced within the French national project Arcade-II (Evaluation of sentence and word alignment tools), as part of the Technolangue programme funded by the French Ministry of Research and New Technologies (MRNT).

[3] This experimentation has been done in March 2011. At present, Google Translate proposes a better translation.

corpus. The results we obtained showed that it is possible to improve machine translation performance by combining a good bilingual lexicon with a large statistical language model. In addition, using a deep linguistic analysis on the sentence to translate and also on the indexed sentences allowed the search engine to present relevant translations on the top of the list of the translation candidates. In order to confirm these results, we are currently working on a large evaluation of our approach and in the same time we are adapting it for a new language pair English-Arabic.

Acknowledgments

This research work is supported by the FINANCIALWATCH (QNRF NPRP: 08-583-1-101) project.

References

Besançon R., De Chalendar G., Ferret O., Gara F., Laib M., Mesnard O., and Semmar N. 2010. A Deep Linguistic Analysis for Cross-language Information Retrieval LIMA :A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. Proceedings of LREC 2010.

Debili, F., Fluhr, C., and Radasoa, P. 1989. About reformulation in full text IRS. Information Processing & Management, Infortmation Processing and Mmanagement, Elsevier.

Grefenstette G. 1999. Cross-language information retrieval. Boston: Kluwer Academic Publishers.

Hutchins J. 2005. Machine Translation: General Overview. The Oxford Handbook of Computational Linguistics, Oxford University Press, Oxford, UK.

Koehn P. 2010. Statistical Machine Translation. Cambridge University Press.

Koehn P., Haddow B., Williams P., and Hoang H. 2010. More Linguistic Annotation for Statistical Machine Translation. Proceedings of the Fifth Workshop on Statistical Machine Translation and MetricsMATR.

Kudo T. and Matsumoto Y. 2001. Chunking with support vector machines. Meeting of the North American chapter of the Association for Computational Linguistics (NAACL), 1–8.

Mohri M., Pereira, F., and Riley M. 2002. Factored Translation Models Weighted Finite-State Transducers in Speech Recognition. Computer Speech and Language, 16(1):69-88.

Semmar N., Laib M., and Fluhr C. 2006. A Deep Linguistic Analysis for Cross-language Information Retrieval. Proceedings of LREC 2006.

Somers H. 2005. Machine Translation: Latest Developments. The Oxford Handbook of Computational Linguistics, Oxford University Press, Oxford, UK.

Trujillo A. 1999. Translation Engines: Techniques for Machine Translation. Springer-Verlag Series on Applied Computing.

Centre d'Innovació
Barcelona Media
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
DCU

ML4HMT 2011

Shared Task on **Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation**

19th November 2011

Barcelona Media
BARCELONA

Sponsors

META

SEVENTH FRAMEWORK PROGRAMME

Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT-2011)

(http://www.dfki.de/ml4hmt/)

Barcelona (Spain) · Saturday, November 19th, 2011

The "Shared Task on Optimising the Division of Labour in Hybrid MT" is an effort to trigger systematic investigation on improving state-of-the-art Hybrid MT, using advanced machine-learning (ML) methodologies. The main focus of the shared task is trying to answer the following question: *Can Hybrid/System Combination MT techniques benefit from extra information (linguistically motivated, decoding and runtime) from the different systems involved?*

Participants of the challenge are requested to build hybrid translations by combining the output of several MT systems of different types. Four participating combination systems, each following a different solution strategy, have been submitted to the shared task. We have computed automated metric scores and conducted an extensive manual evaluation campaign to assess the quality of the hybrid translations. Interestingly, the system winning nearly all the automatic scores only reached a third place in the manual evaluation. Vice versa, the winning system according to manual rankings ranked last place in the automatic metric scores based evaluation. This clearly indicates that more systematic investigation of hybrid system combination approaches, both on a system level and on the evaluation of such systems, needs to be undertaken.

We will work on an updated version of the corpus for the next edition of this shared task, and we will further focus on the integration of advanced machine learning techniques as these are expected to support better exploitation of our corpus' data properties.

We are looking forward to an interesting workshop and want to thank the participants for their efforts during the ML4HMT-2011 Shared Task.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.:249119).

We thank the organisers of LIHMT 2011 for their support.

Organisation committee

Toni Badia (Pompeu Fabra University, Spain)

Christian Federmann (German Research Center for Artificial Intelligence, Germany)

Josef van Genabith (Dublin City University, Ireland)

Maite Melero (Barcelona Media Innovation Center, Spain)

Eleftherios Avramadis (German Research Center for Artificial Intelligence, Germany)

Pavel Pecina (Dublin City University, Ireland)

Marta R. Costa-jussà (Barcelona Media Innovation Center, Spain)

Venue

Barcelona Media
Av. Diagonal, 177, 9th floor
Barcelona, Spain

Programme

09:15 Welcome

09:30 Toni Badia (BM) “Introduction to the ML4HMT Shared Task Workshop”

09:40 Patrick Lambert (LIUM) “The MANY System @ML4HMT-2011”

10:30 Tsuyoshi Okita (DCU) “DCU System Combination @ML4HMT-2011”

11:00 Eleftherios Avramidis (DFKI) “DFKI System Combination with sentence ranking @ML4HMT-2011”

11:30 *Coffee break*

12:00 Christian Federmann (DFKI) “DFKI System Combination using Syntactic Information @ML4HMT-2011”

12:30 Christian Federmann (DFKI) “Comparison of overall results @ML4HMT-2011”

12:40 Alon Lavie (CMU) “MEMT: Alignment-based MT System Combination with Linguistic and Statistical Features”

13:10 Discussion Panel chair: Patrick Lambert (LIUM), Alon Lavie (CMU), Cristina España-Bonet (UPC) and Christian Federmann (DFKI). Topics include:

> (i) Two Hybrid paradigms: Multi- vs Single-system
> (ii) In the Multi-system approach: can Hybrid/System Combination MT techniques benefit from extra information (linguistically motivated, decoding and runtime) from the different systems involved?
> (iii) Evaluation in the Multi-system approach: do we evaluate the output in isolation or do we use evaluation information from the different systems involved?

14:00 *Lunch*

The ML4HMT workshop is supported by META-NET

About META-NET

META-NET is a Network of Excellence dedicated to fostering the technological foundations of a multilingual European information society. Language Technologies will:

- enable communication and cooperation across languages,
- secure users of any language equal access to information and knowledge,
- build upon and advance functionalities of networked information technology.

A concerted, substantial, continent-wide effort in language technology research and engineering is needed for realising applications that enable automatic translation, multilingual information and knowledge management and content production across all European languages. This effort will also enhance the development of intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots.

To this end META-NET is building the Multilingual Europe Technology Alliance (META). Bringing together researchers, commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders. META will prepare the necessary ambitious joint effort towards furthering language technologies as a means towards realising the vision of a Europe united as one single digital market and information space.

META NET

http://www.meta-net.eu/

Table of Contents

Machine Translation System Combination with MANY for ML4HMT

Loïc Barrault and Patrik Lambert
LIUM, University of Le Mans
Le Mans, France.
FirstName.LastName@lium.univ-lemans.fr

Abstract

This paper describes the development of a baseline machine translation system combination framework with the MANY tool for the 2011 ML4HMT shared task. Hypotheses from French–English rule-based, example-based and statistical Machine Translation (MT) systems were combined with MANY, an open source system combination software based on confusion networks decoding currently developed at LIUM. In this baseline framework, the extra information about the MT systems provided for the shared task was not used. The system combination yielded significant improvements in BLEU score when applied on system combination data.

1 Introduction

The "Machine Learning for Hybrid Machine Translation" (ML4HMT) workshop proposed a shared task which objective was to investigate whether system combination or hybrid machine translation techniques could benefit from extra information (linguistically motivated, decoding and runtime) from the different systems involved. Thus the focus was to improve the combination of several types of MT systems (rule-based, example-based and statistical) thanks to the extra information corresponding to each type of system.

The LIUM computer science laboratory participated in this shared task providing a baseline for it, that is a system combination withouth using any of the extra information provided by the organisers about each MT system. The one-best system outputs were combined using the MANY[1] (Barrault, 2010) framework, an open source system combination software based on Confusion Networks (CN).

The MANY toolkit was run with all default options. These options, and more generally the various steps involved in the combination system, are described in Section 2. The data available for the shared task and the results obtained are presented in Section 3.

2 System description

MANY is a system combination software (Barrault, 2010) based on the decoding of a lattice made of several Confusion Networks (CN). This is a widespread approach in MT system combination (Rosti et al., 2007; Shen et al., 2008; Karakos et al., 2008; Rosti et al., 2009). MANY can be decomposed in two main modules: an alignment module and a decoder (see Figure 1), which are described in the next sections. A last section deals with parameter tuning.

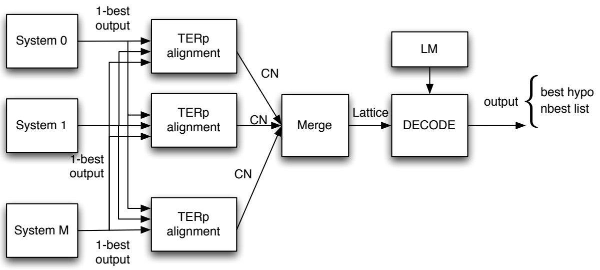


Figure 1: System combination based on confusion network decoding.

[1] MANY is available at the following address http://www-lium.univ-lemans.fr/~barrault/MANY

Alignment Module

The alignment module is actually a version of TERp (Snover et al., 2009) which has been modified to add some functionalities, such as alignment between a sentence and a confusion network. The alignment with TERp uses different costs (which corresponds to an exact match, an insertion, a deletion, a substitution, a shift, a synonym match and a stem match) to compute the best alignment between two sentences. In the case of confusion networks, the match (substitution, synonyms, and stems) costs are considered when the word in the hypothesis matches (is a substitution, a synonyms or a stems of) at least one word of the considered confusion sets in the CN.

The role of the alignment module is to incrementally align the hypotheses against a backbone in order to create a confusion network, as depicted in Figure 2. Each hypothesis acts as backbone, the remaining hypotheses being aligned and merged to it beginning with the nearest in terms of TER and ending with the more distant one. If there are $M + 1$ hypotheses to combine, $M + 1$ confusion networks are generated. Those confusion networks are then connected together into a single lattice by adding a first and last node. The probability of the first arcs (later named priors) must reflect how well such system provides a well structured hypothesis.

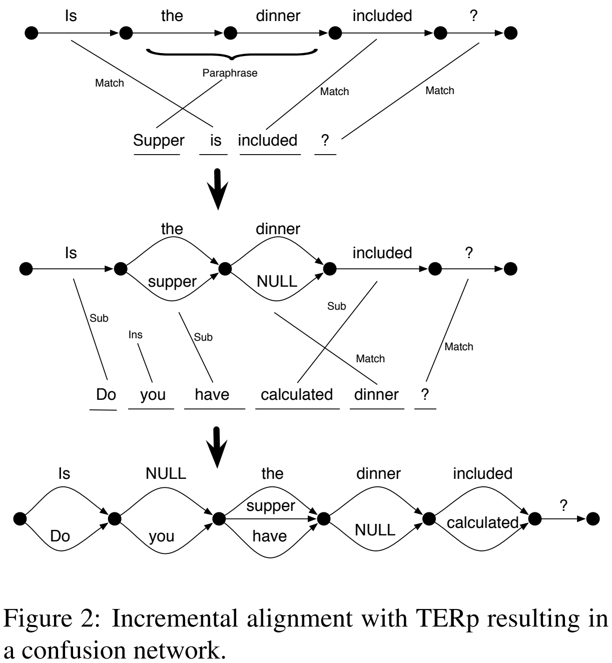


Figure 2: Incremental alignment with TERp resulting in a confusion network.

Decoder

The decoder is based on the token pass algorithm and it accepts as input the lattice previously created. The probabilities computed in the decoder can be expressed as follow :

$$log(P_W) \;=\; \sum_i \alpha_i \, log\Big(h_i(t)\Big) \qquad (1)$$

where t is the hypothesis, the α_i are the weights of the feature functions h_i. The following features are considered for decoding:

- The language model probability: the probability given by a 4-gram language model.
- The word penalty: penalty depending on the size (in words) of the hypothesis.
- The null-arc penalty: penalty depending on the number of null-arcs crossed in the lattice to obtain the hypothesis.
- System weights: each word receive a weight corresponding to the sum of the weights of all systems which proposed it.

At the beginning, only one token is created at the first node of the lattice. Then this token spreads over the consecutive nodes, accumulating the score on the arc it crosses, the language model probability of the word sequence generated so far and null or length penalty if applicable. The number of tokens can increase really quickly to cover the whole lattice, and, in order to keep it tractable, only the $Nmax$ best tokens are kept (the others are discarded), where $Nmax$ can be set at the start. Other methods to restrict the number of tokens (like pruning based on score or other heuristics) can easily be implemented in this software, but this has not been implemented yet.

Tuning

According to recent experiments (Barrault, 2011), it is better to consider the tuning of the alignment module parameters and the decoder parameters in two distinct steps.

By default, TERp costs are set to 0.0 for match and 1.0 for everything else. These costs are not optimal, since a shift in that case will hardly be possible. However, tuning these costs (with Condor, a numerical optimizer based on Powell's algorithm, (Berghen and Bersini, 2005)) never showed significant improvements so far. Thus the default configuration in the current version of MANY is to keep default TERp weights for alignment.

Decoder feature functions weights were optimized with MERT (Och, 2003). The 300-best list created at each MERT iteration was appended to the n-best lists created at previous iterations. This proved to be a more reliable tuning than previous tuning of decoder weights performed with Condor (Barrault, 2011).

3 Shared Task

The task consisted in combining the outputs of the following five MT systems: Joshua (hierarchical), Lucy (rule-based), Metis (working with a monolingual target corpus and a bilingual dictionary only), Apertium (rule-based) and Matrex (combination of example-based and phrase-based SMT features). Outputs of these MT systems were provided on a development set to tune the combination framework, and on a test set to produce the combination output to be evaluated. We took as input of our combination system the one-best plain text output extracted from the xml file for each MT system. The original case was preserved (lower case for the Joshua output and true case for the rest of systems) and the texts were tokenized. Statistics of the development (dev) and test sets calculated on the reference after tokenization are presented in Table 1.

NAME	#sent.	#words
dev	1025	23908
test	1026	25863

Table 1: ML4HMT shared task corpora : number of sentences and running words (after tokenization) calculated on the reference.

Language model. The English target language model has been trained on the only data set allowed for the shared task, namely the News Commentary corpus provided for the MT shared task of the Sixth Workshop of Statistical Machine Translation (WMT 2011).[2] This corpus contains 180k running words of quality commentary articles about the news. We used the SRILM toolkit (Stolcke, 2002) to train a 4-gram back-off language model with Kneser-Ney (Kneser and Ney, 1995) smoothing.

LM weight	Word penalty	Null penalty
0.032	0.23	0.010

Joshua	Lucy	Metis	Apertium	Matrex
-0.013	-0.27	+0.014	-0.21	-0.22

Table 2: Parameters obtained with tuning decoder parameters with MERT.

System	BLEU	TER	METEOR
Joshua	13.80	67.30	52.71
Lucy	22.70	61.97	57.62
Metis	9.09	80.02	41.36
Apertium	21.61	62.88	55.25
Matrex	20.18	60.18	56.55
MANY	24.36	58.55	56.25

Table 3: Automatic scores on the test set for the single MT hypotheses and their combination with MANY.

Tuning. The alignment module was run on the dev set MT hypotheses without tuning, keeping the default TERp weights (0 for exact match and 1 for the other costs). Decoding of the resulting lattice of confusion networks was tuned using MERT to obtain the set of decoder feature functions weights which provides the best scoring combination output on the dev set. The optimum set of parameters obtained is presented in Table 2. The system thus gave a higher weight to words coming from the hypothese proposed by Lucy, then by Matrex, Apertium, Joshua, and it weighted negatively words proposed by Metis.

Evaluation. The test set hypotheses were incrementally aligned with TERp default costs, a lattice was created with the resulting confusion networks, and decoding was conducted with the weights presented in Table 2. This produced the final combination output, which was evaluated on the test set against the reference, as well as the MT hypotheses.

[2] http://www.statmt.org/wmt11/

The evaluation results are shown in Table3. The combination with MANY improves the best single system BLEU score (Lucy) by 1.6 points, the best single system TER score (Matrex) by 1.6 points, but its METEOR score is 1.3 points below the one of the best single system (Matrex).

Another remark about the results is that the ranking of the systems resulting from the weights obtained during tuning (Table 2), namely Lucy/Matrex/Apertium/Joshua/Metis, is consistent with the METEOR score ranking, and close to the BLEU or TER rankings.

4 Conclusions and perspectives

We ran the MANY system combination toolkit on five MT systems of different types provided for the ML4HMT workshop shared task. The combination achieved a better BLEU score and TER score than the best single system (with a 1.6 point gain in both cases), but a worse METEOR score. We emphasize that in the current version, although MANY can benefit from various information sources, the decision taken by the decoder mainly depends on a target language model. Thus the decision to restrict the size of the authorized monolingual training corpus was a severe limitation. In the future, we want to estimate good confidence measure to use in place of the systems priors. These confidences measures have to be related to the system performances, but also to the complementarity of the systems considered.

Finally, we want to give some ideas of how extra information about the MT systems could be taken into account within MANY. The decoder could benefit from information related to the hypothesis, such as the phrase pairs used and their probabilities, or the language model probabilities of each n-gram. The search space could be extended with synonyms, paraphrases or other types of information.

5 Acknowledgements

This work has been partially funded by the European Union under the EuroMatrix Plus project (http://www.euromatrixplus.net, IST-2007.2.2-FP7-231720).

References

Loïc Barrault. 2010. MANY : Open source machine translation system combination. *Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for Machine Translation*, 93:147–155.

Loïc Barrault. 2011. Many improvements for WMT. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 135–139, Edinburgh, Scotland.

Frank Vanden Berghen and Hugues Bersini. 2005. CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, 181:157–175, September.

Damianos Karakos, Jason Eisner, Sanjeev Khudanpur, and Markus Dreyer. 2008. Machine translation system combination using ITG-based alignments. In *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*, pages 81–84, Columbus, Ohio, USA, June 16-17.

Kneser and Ney. 1995. Improved backing-off for m-gram language modeling. In *IEEE Inte. Conf. on Acoustics, Speech and Signal Processing*, pages 49–52, Detroit, MI, May.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, Sapporo, Japan.

A.-V.I. Rosti, S. Matsoukas, and R. Schwartz. 2007. Improved word-level system combination for machine translation. In *Association for Computational Linguistics*, pages 312–319.

A.-V.I. Rosti, B. Zhang, S. Matsoukas, , and R. Schwartz. 2009. Incremental hypothesis alignment with flexible matching for building confusion networks: BBN system description for WMT09 system combination task. In *EACL/WMT*, pages 61–65.

Wade Shen, Brian Delaney, Tim Anderson, and Ray Slyh. 2008. The MIT-LL/AFRL IWSLT-2008 MT System. In *International Workshop on Spoken Language Translation*, Hawaii, U.S.A, 69–76.

M. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2009. TER-Plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation Journal*.

A. Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, CO.

DCU Confusion Network-based System Combination for ML4HMT

Tsuyoshi Okita
Dublin City University
Glasnevin, Dublin 9, Ireland
tokita@computing.dcu.ie

Josef van Genabith
Dublin City University
Glasnevin, Dublin 9, Ireland
josef@computing.dcu.ie

Abstract

This paper describes a system combination module in the MaTrEx (Machine Translation using Examples) MT system developed at Dublin City University. We deployed this module to the evaluation campaign for the ML4HMT task, achieving an improvement of 2.16 BLEU points absolute and 9.2% relative compared to the best single system.

1 Introduction

This paper describes a system combination module in the MaTrEx (Machine Translation using Examples) MT system (Du et al., 2009; Okita et al., 2010b) developed at Dublin City University. We deployed this module to the evaluation campaign for the ML4HMT task.

System combination techniques often rely on the Minimum Bayes Risk decoder (MBR decoder) (Kumar and Byrne, 2002) with and without confusion network. Our system combination approach uses the MBR decoder with the confusion network (Bangalore et al., 2001; Matusov et al., 2006; Du et al., 2009). One notable addition in this paper is in the optimization procedure (presented in Section 2) which considers all the possible combinations of given inputs and may result in excluding the outputs of some of the systems participating in system combination architecture. As far as we know, there is no paper yet which discusses in detail how to best select from the provided set of single best translations. This paper also seeks to explain the mechanism why this selection works.

The alternative approach which does not use the confusion network tends to address the problem when the MBR decoder has to handle larger n in its n-best lists (Tromble et al., 2008; DeNero et al., 2009).

The remainder of this paper is organized as follows. Section 2 describes the system combination strategy we used in this evaluation campaign. In Section 3, our experimental results are presented. In Section 4, we discuss why one inferior system is better removed in the overall system combination strategy. We conclude in Section 5.

2 Our System Combination Strategy

Let E be the target language, F be the source language, and $M(\cdot)$ be an MT system which maps some sequence in the source language F into some sequence in the target language E. Let $\mathcal{E}$ be the translation outputs of all the MT systems. For a given reference translation E, the decoder performance can be measured by the loss function $L(E, M(F))$. Given such a loss function $L(E, E')$ between an automatic translation E' and the reference E, a set of translation outputs $\mathcal{E}$, and an underlying probability model $P(E|F)$, a MBR decoder is defined as in (1) (Kumar and Byrne, 2002):

$$\begin{aligned}\hat{E} &= \arg\min_{E'\in\mathcal{E}} R(E') \\ &= \arg\min_{E'\in\mathcal{E}} \sum_{E'\in\mathcal{E}} L(E,E')P(E|F) \qquad (1)\end{aligned}$$

where $R(E')$ denotes the Bayes risk of candidate translation E' under the loss function L. We use BLEU (Papineni et al., 2002) as this loss function L.

system	MT output seqs				prob	expected matches
1	a	a	a	c	0.30	expected-matches(aaac)=0.3*4+0.2*0+0.2*0+0.2*0+0.1*0=1.2
2	b	b	c	d	0.20	expected-matches(bbcd)=0.3*0+0.2*4+0.2*3+0.2*3+0.1*1=2.1
3	b	b	b	d	0.20	expected-matches(bbbd)=0.3*0+0.2*3+0.2*4+0.2*2+0.1*2=2.0
4	b	b	c	f	0.20	expected-matches(bbcf)=0.3*0+0.2*3+0.2*2+0.2*4+0.1*0=1.8
5	f	f	b	d	0.10	expected-matches(ffbd)=0.3*0+0.2*1+0.2*2+0.2*0+0.1*4=1.0
system	MT output seqs				prob	expected matches
1	a	a	a	c	0.33	expected-matches(aaac)=0.33*4+0.22*0+0.22*0+0.22*0+0.00*0=1.32
2	b	b	c	d	0.22	expected-matches(bbcd)=0.33*0+0.22*4+0.22*3+0.22*3+0.00*1=**2.20**
3	b	b	b	d	0.22	expected-matches(bbbd)=0.33*0+0.22*3+0.22*4+0.22*2+0.00*2=1.98
4	b	b	c	f	0.22	expected-matches(bbcf)=0.33*0+0.22*3+0.22*2+0.22*4+0.00*0=1.98
5	-	-	-	-	0.00	

Table 1: Motivating examples. MBR decoding can be schematically described as the expectation of the number of matching between the MT output sequence and some sequence, as is described in this table. The upper row shows the MT output sequences consisting of 5 systems, while the lower row shows the MT output sequences consisting of 4 systems. In this case, the expected matches of “bbcd” for 4 systems (lower row) are better than those for 5 systems (upper row). This suggests that it may be better to remove extremely bad MT output from the inputs of system combination.

We now introduce the idea of searching for the optimal subset $\mathcal{E}_0$ among $\mathcal{E}$ (where $\mathcal{E}$ is the translation outputs of all the MT systems participating in the system combination). The motivating example is shown in Table 1. In this example, five MT output sequences “aaac”,”bbcd”,”bbbd”,”bbcf”, and “ffbd” are given. Suppose that we calculate the expected matches of “bbcd”, which constitute the negative quantity in Bayes risk. If we use all the given MT outputs consisting of 5 systems, the expected matches sum to 2.1. If we discard the system producing ”ffbd” and only use 4 systems, the expected matches improve to 2.20. As a conclusion, it is not always the best solution to use the full set of given MT outputs, but to remove some MT output can be a good strategy. This suggests to consider all the possible subsets of the full set of MT outputs, as is shown in (2):

$$\hat{E} = \arg\min_{\mathcal{E}_i \subseteq \mathcal{E}} \sum_{E' \in \mathcal{E}_i} L(E, E')P(E|F) \quad (2)$$

where $\mathcal{E}_0 \subseteq \mathcal{E}$ indicates that we choose $\mathcal{E}_0$ from all the possible subsets of $\mathcal{E}$ (or a power set of $\mathcal{E}$). [1]

We now move on to obtain each value of $\arg\min_{E' \in \mathcal{E}_i} \sum_{E' \in \mathcal{E}_i} L(E, E')P(E|F)$ and consider a confusion network which enables us to combine several fragments from MT outputs. In the first step, we select the sentence-based best single system via a MBR decoder. Note that single system outputs are often used as the backbone of the confusion network. For example in Table 2, system t1 is selected as the backbone. Note that the backbone determines the general word order of the confusion network.

In the second step, based on the backbone which is selected in the first step, we build the confusion network by aligning the hypotheses with the backbone. In this process, we used the TER distance (Snover et al., 2006) between the backbone and the hypotheses. We do this for all the hypotheses sentence by sentence. Note that in this process, deleted words are substituted as NULL words (or ϵ-arcs). For example in Table 2, the lower half shows an example of a confusion network. hyp(t2), . . . , hyp(t5) are aligned according to the backbone(t1). Note that $*$ denotes ϵ-arcs, (D) denotes deletion, (I) denotes insertion, and (S) denotes substitution following the terminology in the TER distance literature. The right most column in Table 2 in the rows of the confusion network, that is 57.14, 71.43, and so forth, shows the TER score for this example.

In the third step, the consensus translation is extracted as the best path in the confusion network. The most primitive approach (Matusov et al., 2006) is to select the best word $\hat{e}_k$ by the word posterior probability via voting at each position k in the con-

[1] A power set of $\mathcal{E} = \{1, 2\}$ is $\{\{1, 2\}, \{1\}, \{2\}, \emptyset\}$.

segment 782		
Input t1	since the a team of almost 1000 policemen is in charge of security .	
Input t2	since the previous day an equipment of almost 1000 policewomen is being in charge of the safety .	
Input t3	from the previous day a team from almost 1000 police officer himself is using of the security	
Input t4	from the previous day a team of almost 1000 police is occupying of the security .	
Input t5	since the day before a team of almost 1 policemen is pursuing security .	
backbone(t1)	since the a team of almost 1000 policemen is in charge of security .	
hyp(t2)	since the previous(I) day(I) an(S) equipment(S) of almost 1000 policewomen(S) is being(I) in charge of the(I) safety(S) .	57.14
hyp(t3)	from(S) the(I) previous(I) day(S) a team from(S) almost 1000 police(I) officer(I) himself(S) is using(S) the(S) of security .	71.43
hyp(t4)	from(S) the previous(I) day(I) a team of almost 1000 police(S) is occupying(S) the(S) of security .	50.00
hyp(t5)	since the day(I) before(I) a team of almost 1(S) policemen is *(D) *(D) pursuing(S) security .	42.86
output	since the previous day a team of almost 1000 policemen is in charge of security .	

Table 2: Example from the 782th sentence from the testset. First we choose the first input as the backbone. Second, we make the confusion network measuring the performance by TER. Then, the consensus translation of ”since the previous day a team of almost 1000 policemen is in charge of security .” is obtained as an output.

fusion network, as in (3):

$$\hat{E}_k \quad = \quad \arg\max_{e \in \mathcal{E}} p_k(e|F) \qquad (3)$$

Note that this word posterior probability can be used as a measure how confident the model is about this particular word translation (Koehn, 2010), as defined in (4):

$$p_i(e|F) \quad = \quad \sum_j \delta(e, e_{j,i}) p(e_j|F) \qquad (4)$$

where $e_{j,i}$ denotes the i-th word and $\delta(e, e_{j,i})$ denotes the indicator function which is 1 if the i-th word is e, otherwise 0. However, in practice as is shown by (Du et al., 2009; Leusch et al., 2009), the incorporation of a language model in this voting process will improve the quality further. Hence, we use the following features in this voting process: word posterior probability, 4-gram and 5-gram target language model, word length penalty, and NULL word length penalty. Note that Minimum Error-Rate Training (MERT) is used to tune the weights of the confusion network. In Table 2, “since the previous day a team of almost 1000 policemen is in charge of security .” is selected in this voting process. In the final step, we remove the ϵ-arcs if existed.

3 Experiments

We use MERT (Och, 2003) internally to tune the weights and language modeling is provided by SRILM (Stolcke, 2002). We did not use any external language data resources.

Our results as obtained by the system described in Section 2 (which automatically selects and discards translations provided by the component MT systems) are shown in the results line in Table 3. Although the organizers provide the reference set for the testset, the decision that we make in the following is based on the results obtained on the development set performance since we cannot access the reference set in “real life” situations. Due to the performance on the development set, we tuned the parameters in our system as is described in Section 2.

The improvement in BLEU was 2.16 points absolute and 9.2% relative compared to the performance of system t2, the single best performing system (we optimized according to BLEU). Except for METEOR, we achieved the best performance in NIST (0.14 points absolute and 2.1% relative), WER (0.71 points absolute and 1.1% relative) and PER (0.64 points absolute and 1.3% relative) as well.

In order to shed further light on the intermediate results, we sampled three combinations of single best translation outputs, which are shown in Table 3 as well. Combination 1 includes all of the five single best translation outputs. Combination 2 includes t1, t2, t4, and t5 which eliminates system t2 which performed worst in terms of development set perfor-

	NIST	BLEU	METEOR	WER	PER
system t1	6.3934	0.1968/0.1289*	0.5022487	62.3685	47.3074
system t2	6.3818	0.2337/0.1498*	**0.5732194**	64.7816	49.2348
system t3	4.5648	0.1262/0.0837*	0.4073446	77.6184	63.0546
system t4	6.2136	0.2230/0.1343*	0.5544878	64.9050	50.2139
system t5	6.7082	0.2315/0.1453*	0.5412563	60.6646	45.1949
results	**6.8419**	**0.2553**	0.5683086	**59.9591**	**44.5357**
combination 1 (1,2,3,4,5)	6.7151	0.2505	0.5701207	60.6993	45.5148
combination 2 (1,2,4,5)	6.8419	0.2553	0.5683086	59.9591	44.5357
combination 3 (2,4,5)	6.7722	0.2498	0.5687383	60.6723	45.2257

Table 3: We do experiments and obtained the results as above (See the results line). All the scores are on testset except those marked * (which are on devset). On comparison, we did sampling of three combinations of the single systems, which shows that our results are equivalent to the combination 2. These exeprimental results validate our motivating results: it is often the case that some radically bad translation output may harm the final output by system combination. In this case, system t3 whose BLEU score is 12.62 has a negative effect on the results of system combination. The best performance was achieved by removing this system, i.e. the combination of systems t1, t2, t4, and t5.

mance. Combination 3 includes t2, t4, and t5 which eliminates the two worst systems in terms of the development set performance.

It is evident that our overall result is equivalent to Combination 2. Combination 2 achieved the best performance among these three combinations in NIST (0.13 points absolute and 2% relative), WER (0.70 points absolute and 1.1% relative) and PER (0.66 points absolute and 1.4% relative) as well. Combination 1 is second best in terms of BLEU scores. The improvement in BLEU was 1.68 points absolute and 7.1% relative. Combination 3 achieves 1.61 points improvement absolute and 6.9% relative.

4 Discussion

In Statistical Machine Learning (Vapnik, 1998), the term Bayes risk refers to the minimum risk over all possible measurable functions. This strategy leads to find the best hypothesis under the worst case analysis which is called agnostic learning (Kearns et al., 1994). In agnostic learning, with probability 1-δ, the number of training samples sufficient to ensure that every hypothesis H having zero training error will have a true error m of at most ϵ, is investigated as is shown in (5):

$$m \quad \geq \quad \frac{1}{\epsilon}(\ln|H| + \ln|\frac{1}{\delta}|) \tag{5}$$

In Support Vector Machines (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000), this strategy is called the empirical risk minimization or the structural risk minimization. For example, in the case of an (independent) regression problem,[2] Bayes risk is defined as in (6):

$$R(t) \quad = \quad \inf_g R(g) \tag{6}$$

where t is a target function and g is a true function. Bayes risk can be further rewritten as in (7):

$$R(g) \quad = \quad P(g(X) \neq Y) = \mathbb{E}(\mathbf{1}_{g(X)\neq Y}) \tag{7}$$

where **1** denotes an indicator function. As we cannot measure this risk since P is unknown, we use the following empirical risk (8) to measure the performance:

$$R_n(g) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{g(X_i)\neq Y_i} \tag{8}$$

This leads to the theory of worst case analysis taken by Support Vector Machines. To seek minimal risk is equivalent to seeking high probability mass in the hypothesis space since Eq (8) counts how many $g(X_i)$ and Y_i disagree with each other. We seek high counts of disagreement.

[2] Let us consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$. We assume that a set of n IID pairs (X_i, Y_i) sampled according to an unknown but fixed distribution P. Suppose that our task is to predict a function $g : \mathcal{X} \rightarrow \mathcal{Y}$ where we call g a true function. Now, let t be a target function $t(x) = \text{sgn}\eta(x)$ where $\eta(x) = \mathbb{E}[Y|X = x] = 2\mathbb{P}[Y = 1|X = x] - 1$.

In the case of Machine Translation, this analogy can be extended. As is shown in Eq (1), MBR decoding seeks to obtain the translations whose probability mass are concentrated (Koehn, 2010) where each word is split as in Eq (4) if we take the confusion network-based approach of system combination. Hence, if the same words appear in the same word position, such words may occupy the high probability mass in Eq (4). If we include incorrect translation output among candidate translation outputs in the same word position, incorrect words may occupy the high probability mass. Then, the resulting output may include such bad words, causing the overall BLEU score to be low. Although this is not a conclusive explanation, this explains the possibility in a qualitative way why our combination 1 can be worse than our combination 2 in Table 3.

5 Conclusion and Further Studies

This paper describes the system combination module in the MT system MaTrEx developed at Dublin City University. We deployed the system combination module to this evaluation campaign. In this paper, we introduce a new input selection mechanism which removes some radically bad systems for the sake of achieving final better overall performance. Although this phenomenon was observed between JP-EN (Okita et al., 2010b), we implemented this mechanism in the procedure in this paper and showed the same to hold between ES-EN. Improvement was 2.16 BLEU points absolute and 9.2% relative compared to the best single system.

Further study will investigate the effect of bad translation inputs in system combination. Currently our implementation of Eq (2) is somewhat naive, in that the approach considers all subsets of translations contributed by the individual MT systems. We will work on a strategy how to select translation inputs optimally. In particular such a discussion will be fruitful if our inputs are the 1000-best list as in the case of Tromble et al. (Tromble et al., 2008) and DeNero et al. (DeNero et al., 2009). Their improvements are in general quite small compared to the confusion network-based approach. As is shown in Figure 1, the 100-best list and the 1000-best list produced by Moses (Koehn et al., 2007) tend not to be sufficiently different and do not produce meaningful translation alternatives. As a result, their BLEU score tends to be low compared to the (nearly best) single systems. This means that in our strategy those MT inputs may be better removed rather than employed as a useful source in system combination.

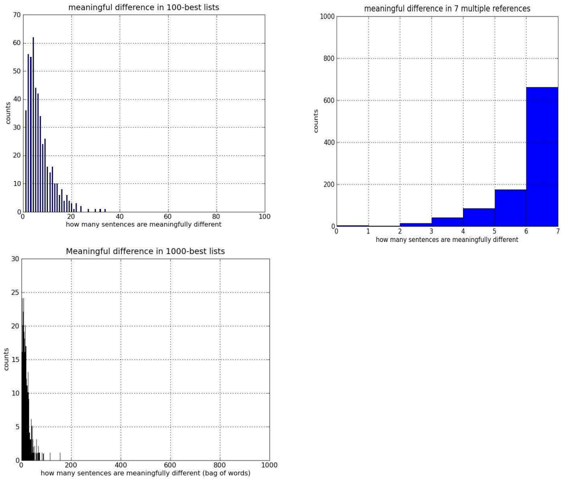


Figure 1: The upper left figure shows the count of exact matches among the translation outputs of Moses as a 100-best list after stop-word removal and sorting; We project each sentence in a 100-best list onto vector space model and count the number of points. The lower left figure shows the same quantity for a 1000-best list. The upper right figure shows the same quantity for a 7-multiple reference (human translation). We use the parallel data of IWSLT 07 JP-EN where we use devset5 (500 sentence pairs) as a development set and devset4 (489 sentence pairs) as a test set; 7-multiple references consist of devset4 and devset5 (989 sentence pairs). For example, the upper left figure shows that 7% of sentences produce only one meaningful sentence in a 100-best list and the other 99 sentences in a 100-best list is just a reordered version. In contrast, the upper right figure of human translation shows that more than 70% of sentences in 7 multiple references are meaningfully different.

Yet another avenue for further study is to provide prior knowledge into the system combination module. In word alignment, one successful strategy is to embed prior knowledge about alignment links (Okita et al., 2010a; Okita, 2011; Okita and Way, 2011), which work as the link between statistical learning and linguistic resources. We have shown that the selection of MT input sentences is an effective strategy in this paper. Similarly, it would be interesting to incorporate some prior knowledge about system combination, for example, (in)correct words or phrases in some particular translation output.

6 Acknowledgements

This research is supported by the the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the T4ME project (grant agreement no. 249119).

References

Srinivas Bangalore, G. Bordel, and G. Riccardi. 2001. Computing consensus translation from multiple machine translation systems. *In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 350–354.

Nello Cristianini and John Shawe-Taylor. 2000. Introduction to Support Vector Machines. *Cambridge University Press*.

John DeNero, David Chiang, and Kevin Knight. 2009. Fast consensus decoding over translation forests. *In proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 567–575.

Jinhua Du, Yifan He, Sergio Penkale, and Andy Way. 2009. MaTrEx: the DCU MT System for WMT 2009. *In Proceedings of the Third EACL Workshop on Statistical Machine Translation*, pages 95–99.

Michael J. Kearns, Robert Schapire, and Linda Sellie. 1994. Towards efficient agnostic learning. *Machine Learning*, 17:115–141.

Philipp Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for Statistical Machine Translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Philipp Koehn. 2010. Statistical machine translation. *Cambridge University Press*.

Sanjiv Kumar and William Byrne. 2002. Minimum Bayes-Risk word alignment of bilingual texts. *In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 140–147.

Gregor Leusch, Eugene Matusov, and Hermann Ney. 2009. The rwth system combination system for wmt 2009. *In Fourth EACL Workshop on Statistical Machine Translation (WMT 2009)*, pages 56–60.

Eugene Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. *In Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–40.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Tsuyoshi Okita and Andy Way. 2011. Given bilingual terminology in statistical machine translation: Mwe-sensitve word alignment and hierarchical pitman-yor process-based translation model smoothing. *In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*.

Tsuyoshi Okita, Alfredo Maldonado Guerra, Yvette Graham, and Andy Way. 2010a. Multi-Word Expression sensitive word alignment. *In Proceedings of the Fourth International Workshop On Cross Lingual Information Access (CLIA2010, collocated with COLING2010), Beijing, China.*, pages 1–8.

Tsuyoshi Okita, Jie Jiang, Rejwanul Haque, Hala Al-Maghout, Jinhua Du, Sudip Kumar Naskar, and Andy Way. 2010b. MaTrEx: the DCU MT System for NTCIR-8. *In Proceedings of the MII Test Collection for IR Systems-8 Meeting (NTCIR-8), Tokyo.*, pages 377–383.

Tsuyoshi Okita. 2011. Word alignment and smoothing method in statistical machine translation: Noise, prior knowledge and overfitting. *PhD thesis. Dublin City University*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method For Automatic Evaluation of Machine Translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. *In Proceedings of the International Conference on Spoken Language Processing*, pages 901–904.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice minimum bayes-risk decoding for statistical machine translation. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629.

Vladimir Vapnik. 1998. Statistical learning theory. *Wiley and Sons*.

DFKI System Combination with Sentence Ranking at ML4HMT-2011

Eleftherios Avramidis
German Research Center for Artificial Intelligence (DFKI)
Language Technology Group (LT)
Berlin, Germany
eleftherios.avramidis@dfki.de

Abstract

We present a pilot study on a Hybrid Machine Translation system that takes advantag e of multilateral system-specific metadata provided as part of the shared task. The proposed solution offers a machine learning approach, resulting into a selection mechanism able to learn and rank system outputs on the sentence level, based on their quality. For training, due to the lack of human annotations, word-level Levenshtein distance has been used as a quality indicator, whereas a rich set of sentence features was extracted and selected from the dataset. Three classification algorithms (Naïve Bayes, SVM and Linear Regression) were trained and tested on pairwise featured sentence comparisons. The approaches yielded high correlation with original rankings (tau = 0.52) and selected the best translation in 54% of the cases.

1 Introduction

Optimizing Machine Translation (MT) performance through Hybrid Machine Translation has been a long standing goal, given the possible benefit from combining systems of different theoretical backgrounds (Habash, 2003). So far, research has adopted several approaches to MT system combinations. A vast majority of them treat the participating MT systems as black boxes, aiming to combine them based on some universal measure of quality (Callison-Burch and Flournoy, 2001). This has also allowed combinations of different outputs to take place on a word or phrase level (Matusov et al., 2006; Rosti et al., 2007; Hillard et al., 2007).

Meanwhile, there have been many suggestions that information derived from the translation process can contain useful hints for the quality of the produced output. Positive results have been shown on the development of Confidence Estimation metrics, in most of the cases complementing other universal features (Quirk, 2004; Rosti et al., 2007; Specia et al., 2009). Though, the best way to take advantage of such information, deriving from systems of different origin, remains still an open question.

Here we demonstrate a pilot study which tries to take advantage of the multi-dimensional and heterogeneous annotation annotation over MT output, provided in the frame of the ML4HMT-2011 shared task. In Section 2, we try to re-formulate the problem in a way which is easier to approach using Machine Learning (ML). In Section 3, we show how a suitable feature set has been extracted. In Section 4 we show the performance of Machine Learning algorithms and in Section 5 we provide a discussion of the results.

2 Re-formulation of the problem

2.1 Focus on a sentence level

The ML4HMT-2011 corpus provides a development and a test set of approximately 1,000 sentences each, translated by 5 different systems. Each translation output is accompanied with metadata referring to parts of the process each system performed. Although the annotation is rich, the main difficulty of the task relies on the fact that each system provides a different set of metadata, which are scattered over different derivation steps, that are not comparable through with other. For example, statistical systems provide statistics on the decoding steps and their search algorithm, while rule-based systems yield several derivation steps within their tree analyses. For this reason, a simplified approach would be to restrict the granularity of the combination on the sentence level. This allows for a better picture on the compilation of the feature vectors that are required in a ML approach. It could also be applied for selecting the backbone translation in other MT combination approaches.

2.2 Pairwise decisions and ranking

Working on a sentence level leads to the goal of building an empirical selection mechanism, which would be able to estimate the quality of the generated sentence alternatives on the fly and choose accordingly. A draft learning approach on this direction would use a classification method, where the id of the best system serves as the class, and metadata from all alternative outputs forms the feature vector for the classification. This approach, however, would result in a really difficult problem to solve, given also the size of the data, which would probably lead into sparseness problems.

Instead, we consider the tactic of breaking the quality judgement into pairwise comparisons, between all the 5 translation outputs per source sentence. This gives a total of about 17,000 training instances with binary classes, which makes the training of a classifier more plausible. Additionally, the classifier now has to "learn" and provide a binary answer to the much simpler question "*which of these two sentences is better?*", given the meta-data from the two systems themselves. The pairwise (positive or negative) judgments are then summed up, so as to order the 5 outputs based on their predicted quality. We have therefore reformulated the problem into modelling a *quality ranking* of the sentences. Coming back to the system combination requirement, the best ranked sentence can then be selected for the combined output.

2.3 Supervised learning

It would make sense to try to learn such a mechanism, given a training set with relatively reliable quality indicators, for example, results of human evaluation. Unfortunately, although a development set has been provided, it does not include an objective measurement of quality within each set of 5 alternatives. The only relevant information can be derived from the reference translation, which, in a way, forms the gold translation that that the MT systems should reach.

As an answer to this question, we examined the so-called segment-level metrics that could provide this information. In the end, word-level Levenshtein distance seemed to adhere better to our needs. So, thereafter we consider this as a quality indication and we will develop and evaluate the ML outcome based on it. This would provide us with an intuition for the learning capabilities of the approach and allow a potential shift to gold human judgments, when these are available.

3 Extracting and selecting features

Given the decisions described above, the various multilateral and overlapping annotations on several levels of the translation process have to be converted to a shallow set of sentence-level features.

3.1 Defining sentence-level features

Based on our intuition given the knowledge about the functioning of each system, we extracted the following features:

- **Joshua**: overall translation probability, tuned weights, count of phrases. Decoding search features included the number of pre-pruned, added, merged nodes, and of fuzzy matches. Three sentence-level statistics were derived from the sequence of feature scores for every decoding step leading to the dominant output: average, standard deviation and variance.

- **MaTrEx**: overall translation probability, tuned weights, count of phrases. As done above, three sentence-level statistics were derived from the sequence of phrase scores and future cost estimates: average, standard deviation and variance.

- **Lucy**: indication that the system performed phrasal analysis and segment combination in the transfer phase (Federmann and Hunsicker, 2011), counts of all nodes appearing in the derivation trees.

- **All**: Scores provided by external linguistic analysis tools, including language model probability (bi-gram, tri-gram, 5-gram), PCFG parsing score (ratio of target to source), number of tokens, number of unknown words. This information was needed for the systems which had no other features easily extractable.

feature	Inf. gain	Gain ratio	Gini
Lucy phrasal analysis	0.181	0.092	0.059
Joshua total probability	0.100	0.050	0.030
External 5gram score	0.000	0.037	0.000
MaTrEx std deviation of future cost	0.058	0.029	0.019
MaTrEx std deviation of probabilities	0.058	0.029	0.019
Joshua/MaTrEx phrase count	0.012	0.005	0.004

Table 1: Results of feature selection by Information Gain, Gain Ratio and Gini Index

feature	ReliefF
Joshua total probability	0.064
Lucy phrasal analysis	0.023
MaTrEx total probability	0.012
Joshua merged nodes	0.011
Joshua word penalty variance	0.010

Table 2: Results of ReliefF feature selection

classifier	p.ac.	τ	b.ac
SVM	0.52	0.52	0.53
Bayes	0.63	0.43	0.54
Linear	0.51	0.25	0.50

Table 3: Results of the classification process

N-gram features have been generated with the SRILM toolkit (Stolcke, 2002) using a language model trained over all monolingual training sets for the WMT 2011 Shared Task (Callison-Burch et al., 2011), interpolated on the 2007 test set. PCFG parsing was done with the Berkeley Parser (Petrov and Klein, 2007), trained over an English and a Spanish treebank (Mariona Taulé and Recasens, 2008). The feature selection algorithms (as well as the learning algorithms below) were implemented with the Orange toolkit (Demšar et al., 2004).

3.2 Feature selection

The whole extraction process, despite the fact that many other annotations were ignored, resulted in a set of more than 50 features per sentence (particularly due to the counts of tree tags). Many machine learning algorithms perform better when they are provided rather smaller sets of uncorrelated features. Even for the algorithms that perform sentence selection themselves, big sets increase the complexity and required runtime.

Three feature selection algorithms were examined as a first step. We computed scores for all attributes based on ReliefF (Kononenko, 1994), Information Gain (Kullback and Leibler, 1951), Gain Ratio and Gini index (Ceriani and Verme, 2011), which can be seen in Tables 1 and 2. We chose the features that have a score higher than 0.01 in either of the metrics.

4 Machine learning algorithms

For the actual task of learning the pairwise comparisons, we trained a SVM, a Naïve Bayes (Cleveland, 1979) and a linear classifier. Feature selection was applied for the latter two, as well as imputation for the missing feature values. Due to implementation issues SVM was lacking the features of category "all" (Section 3). We computed the *pairwise accuracy* (p.ac.) of the classification, the *segment-level tau coefficient* (τ), which indicates the correlation with the rankings produced with word-level Levenshtein distance and the accuracy when focusing only on whether the *best rank* was predicted (b.ac), all measured over the test set. The results can be seen in table 3

5 Discussion

Best-rank accuracy indicates that the classifiers managed to provide the best solution right away, in 50-54% of the cases. This is relatively low, but it can be still considered a small success, given the fact that the probability of random selection out of the five alternatives would be 20%. With some manual evaluation look-up in the classification performed by SVM, we were able to draw the conclusion that this has mostly to do with the fact that the classifier comes to a level of uncertainty concerning the two best ranked sentences. So, most of the times, contradictory judgments would lead to a tie for the two best scored systems, although only one of them needs to be selected. We believe that further processing needs to take place, so that ties as a result of uncertain classification, particularly for the first rank, can be eliminated.

The classifier built with SVM gives the best average sentence-level correlation. This means that it predicted the ranking of the systems better than the other systems, although there were mistakes. Though, the reproduced ranking was rarely too bad, since only 6% of the sentences had a negative tau coefficient. We can also note that the tau correlation given in this task is much higher than the ones achieved by evaluation metrics in WMT Shared Tasks (Callison-Burch et al., 2011), which go up to $\tau = 0.35$. Though, human rankings are not comparable with Levenshtein distance rankings, therefore no clear comparison can be done.

6 Conclusion

We presented an effort to reduce Hybrid Machine Translation selection into sentence-level ranking. Features extracted from the sentence level have been used to train three classification algorithms. SVM shows high sentence-level correlation with the original quality score, whereas Naïve Bayes succeeds slightly better into choosing the best translation per sentence. The potential for further improvement, with more sophisticated feature extraction should be examined.

Acknowledgments

This work has been developed within the TaraXŰ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development. Many thanks to Lukas Poustka for preprocessing and training the Spanish grammar.

References

Callison-Burch, C. and Flournoy, R. S. (2001). A program for automatically selecting the best output from multiple machine translation engines. In *Proceedings of the Machine Translation Summit VIII*, pages 63–66.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Ceriani, L. and Verme, P. (2011). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality*, pages 1–23. 10.1007/s10888-011-9188-x.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Demšar, J., Zupan, B., Leban, G., and Curk, T. (2004). Orange: From experimental machine learning to interactive data mining. In *Principles of Data Mining and Knowledge Discovery*, pages 537–539.

Federmann, C. and Hunsicker, S. (2011). Stochastic parse tree selection for an existing rbmt system. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 351–357, Edinburgh, Scotland. Association for Computational Linguistics.

Habash, N. Y. (2003). *Generation-heavy hybrid machine translation*. PhD thesis, University of Maryland at College Park, College Park, MD, USA. AAI3094491.

Hillard, D., Hoffmeister, B., Ostendorf, M., Schlueter, R., and Ney, H. (2007). iROVER: Improving system combination with classification. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 65–68, Rochester, New York. Association for Computational Linguistics.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. In *Proceedings of the European conference on machine learning on Machine Learning*, pages 171–182, Secaucus, NJ, USA. Springer-Verlag New York, Inc.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86.

Mariona Taulé, M. A. M. and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Matusov, E., Ueffing, N., and Ney, H. (2006). *Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment*, pages 33–40. Association for Computational Linguistics.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York. Association for Computational Linguistics.

Quirk, C. B. (2004). Training a sentence-level machine translation confidence measure. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Protugal. European Language Resources Association (ELRA).

Rosti, A.-V., Ayan, N. F., Xiang, B., Matsoukas, S., Schwartz, R., and Dorr, B. (2007). Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York. Association for Computational Linguistics.

Specia, L., Cancedda, N., Dymetman, M., Turchi, M., and Cristianini, N. (2009). Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the EAMT*, pages 28–35. European Association for Machine Translation.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*, pages 901–904, Denver, Colorado, USA. ISCA Archive.

DFKI System Combination using Syntactic Information at ML4HMT-2011

Christian Federmann, Sabine Hunsicker, Yu Chen, Rui Wang
Language Technology Lab
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
{cfedermann,sabine.hunsicker,yuchen,wang.rui}@dfki.de

Abstract

We present a substitution approach for the combination of machine translation outputs. Using a translation template derived from the output obtained from a rule-based translation engine, we identify parts of the template that could possibly be improved by adding in segments from other MT output. Substitution candidates are determined based on their part-of-speech. Alternative translations from the additional engines are retrieved by using word alignment. Substitution is based on several decision factors, such as part-of-speech, local left-/right-context, and language model probabilities. Our approach differs from other methods as it puts its main focus on preserving the syntactic structure inherited from the rule-based translation template. For the language pair Spanish-English an improvement in BLEU score can be observed.

1 Introduction

Statistical machine translation (SMT) systems have seen a lot of research progress during the last decade. They have effectively outperformed many existing, rule-based machine translation approaches due to their data-driven nature: SMT systems can be trained on large, parallel data sets and they can be tuned according to automated scoring metrics. This is often impossible for rule-based MT (RBMT) engines, in particular if they rely on hand-crafted rules and if they do not involve an overall probability model. This clearly indicates that such systems can profit from further research to catch up with and perhaps even beat current state-of-the-art statistical systems.

Rule-based translation output can have certain advantages over statistically translated content: the syntactic structure of the output is usually correct and complete and the word forms are properly generated. While this is often not fully reflected by standard automatic evaluation metrics such as BLEU (Papineni et al., 2001), it sometimes shows in manual evaluation where human evaluators notice the syntactic quality (i.e., grammaticality) of the output and rank RBMT output better than the automatic scores.

It is interesting to note that recently rule-based systems were able to outperform their statistical opponents in several open evaluation events (Callison-Burch et al., 2009; Callison-Burch et al., 2011). Furthermore, different machine translation paradigms seem to produce output containing complementary errors (Thurmair, 2009). Hence, it makes sense to search for effective ways of combining different systems in order to benefit from the respective advantages of different paradigms while trying to avoid their individual shortcomings. Therefore, we are more focusing on integrating systems of different types instead of applying general system combination techniques because previous results showing correlations between systems suggest that combining them has a great impact on the performance of the combined results (Macherey and Och, 2007).

Previous approaches on system combination include, among others, direct selection from the candidate translations (Callison-Burch and Flournoy, 2001; Akiba et al., 2001), combining word lattices or n-best lists (Frederking and Nirenburg, 1994), hypothesis regeneration with an SMT decoder (Chen et al., 2007; Eisele et al., 2008; Chen et al., 2009) and ROVER-like voting schemes on confusion networks (Jayaraman and Lavie, 2005; Matusov et al., 2006; Rosti et al., 2007; He et al., 2008; Leusch et al., 2009). The last approach constructs a confusion network based on pairwise word alignments of the translation hypothesis, which might be re-ordered. The voting module selects the best consensus translation from the confusion network based on several statistical models. The target language model plays an important role in the voting procedure. It is very likely that the final translation does not resemble any of the hypotheses from the individual systems.

In this shared task, we follow the constituent substitution approach for system combination proposed by (Federmann et al., 2009). The substitution method is similar to voting on a confusion network that has a fixed backbone, however taking more linguistic information into account. Similar work has been reported in (Habash et al., 2009; Espana-Bonet et al., 2011). We choose the translations from an RBMT system as our fixed backbones, or "translation templates" in the hope of retaining the better syntactic structures created by such a system. The consensus translation is then produced by replacing complete constituents in the translation template rather than isolated words. Corresponding phrases in the other candidate translations are identified through word alignments back to the original source sentences. Our substitution algorithm is guided by several decision factors, including part-of-speech, local context, and a language model.

The remainder of this paper is structured as follows. In Section 2, we describe our system combination approach for the ML4HMT shared task and explain our substitution algorithm. Our experiments and results with the resulting combination system are presented in Section 3. Finally, we conclude and provide an outlook on future work in Section 4.

2 System Combination Approach

Our system combination approach is based on previous work on constituent substitution for system combination. One system is chosen as providing the translation template while the remaining systems provide alternative translation variants (on a segment level) which maybe substituted into the template according to a set of decision factors that are derived from syntactic features.

2.1 Finding the right translation template

The organisers of the ML4HMT shared task provided us a data set containing a development set of 1,025 sentences and a test set including 1,026 sentences. For each of these sentences, the source text, the corresponding reference translation, and the translation output as well as various annotations from five machine translation systems were available as source data. Depending on the MT system, the level of annotation details varied greatly and the overall annotation was very heterogeneous which, in our view, made it difficult to make equal use of all annotations/systems. This might be something that could be improved for future work on this data.

We chose the translations by the Lucy RBMT system (Alonso and Thurmair, 2003) as our translation backbone. There are two reasons for this:

1. As a rule-based system, Lucy creates structurally sound sentences. The drawbacks of missing vocabulary coverage and incorrect lexical choice can be made up by mining other translations for better translation variants.

2. Additionally, of all five systems included in the workshop data, only the Lucy system provides analysis trees of the source sentence. Other systems only include trees for the target side of the translation, with many of the systems providing no syntactic information at all.

As our substitution approach is based on identifying *interesting*[1] phrases in the source sentence which are then linked to target language translations via word alignment, we decided to use the translations from the Lucy system as our translation template.

2.2 Reconstructing Lucy parse trees

The organisers of the workshop provided a flattened representation of the Lucy parse trees. Using some heuristics derived from the development set, we designed an algorithm to approximate the original deep

[1] Where *interesting* means *suitable for substitution within our system combination experiments*, e.g., noun phrases.

tree structures. For example, the XML fragment shown in Figure 1 describes the Spanish phrase *la inflación europea*. The noun phrase consists of:

– the determiner (*la*),
– the noun (*inflación*), and
– an adjective phrase (*europea*).

Our heuristics include a mapping which children a node is allowed to have: a node of the category **NO**, e.g., can either be a normal noun (**NST**) or a pronoun (**PRN**). Other part-of-speech categories are not legal wrt. the training data available from the ML4HMT development set.

With those heuristics, we built an XML parser which traverses the flattened XML tree representations and generates corresponding, *approximated tree structures* with a deeper structure. Figure 2 shows the syntactic tree we create from the XML fragment depicted in Figure 1. This deep tree is only an approximation of the original tree and does not contain all information that would be contained within parse trees generated from the original Lucy RBMY system, but it is nevertheless suitable to be used in our approach as we only consider substituting single words inside the candidate phrases we find in the source text parse trees.

2.3 Substitution algorithm

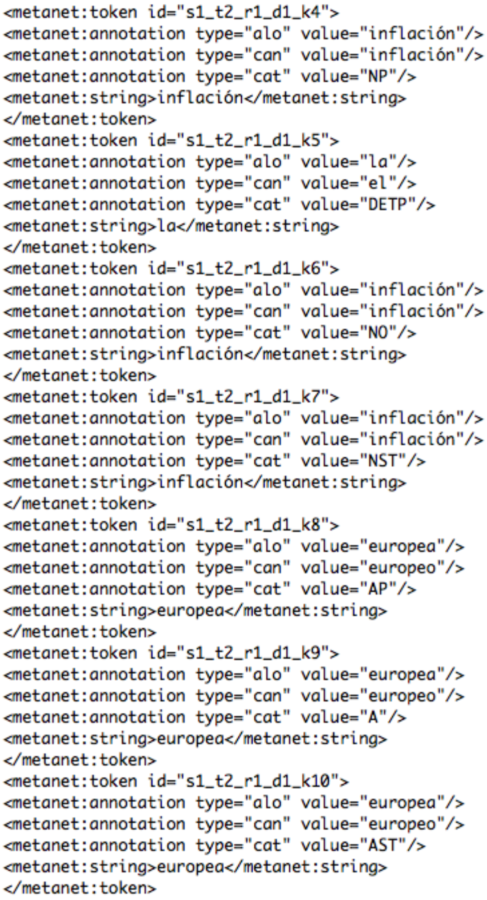
Previously, we have presented a language-independent substitution approach to system combination. Although this work also used rule-based machine translations as backbone[2], we exclusively relied on SMT systems to obtain alternative translation fragments. In this workshop we have access to translation output from systems that follow a variety of paradigms, however.

Lucy is an example for a rule-based MT system.

Apertium is also rule-based, whereas

Metis follows a hybrid approach and translates using a bilingual dictionary and a monolingual target language corpus.

MaTrEx includes several translation modules, but for this workshop a standard phrase-based



```
<metanet:token id="s1_t2_r1_d1_k4">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NP"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k5">
<metanet:annotation type="alo" value="la"/>
<metanet:annotation type="can" value="el"/>
<metanet:annotation type="cat" value="DETP"/>
<metanet:string>la</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k6">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NO"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k7">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NST"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k8">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="AP"/>
<metanet:string>europea</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k9">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="A"/>
<metanet:string>europea</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k10">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="AST"/>
<metanet:string>europea</metanet:string>
</metanet:token>
```

Figure 1: Flattened representation of a Lucy parse tree.

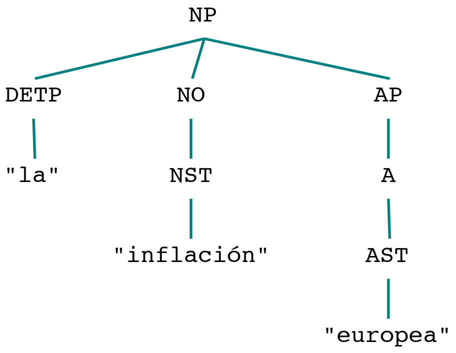


Figure 2: Approximated tree structure.

[2] We also used Lucy RBMT translation output in this previous work, but worked on original parse trees, not approximated tree structures.

SMT model (Moses (Koehn et al., 2007)) was used.

Joshua provides output from a hierarchical phrase-based SMT model.

Using the approximated parse trees, we identify *interesting phrases* suitable for substitution: we consider noun, verb and adjective phrases. These are derived from the trees structures, while the potential substitution fragments from the other systems' output are linked using word alignment. Word alignment is computed using GIZA++ (Och and Ney, 2000). Each candidate translation by the four additional systems is evaluated according to the following features:

Matching POS? We only substitute if the part-of-speech of the candidate matches the reference, i.e., the translation template. This way we will not destroy the syntactic structure.

Majority vote Two or more systems may offer the same candidate translation. We prefer more frequent candidate fragments.

Context We take into account the part-of-speech of the surrounding tokens, left and right, to ensure that the fragment will fit into the context.

Language Model The candidate fragments as well as their -1 left and -1 right context are scored using a language model trained on EuroParl (Koehn, 2005).

3 Experiments

We tried out several combinations of features in our substitution system. In this section, we report on our experiments with the ML4HMT data set and provide results from comparing our system combination results to the baseline Lucy RBMT translation output. In our experiments, we translated from Spanish→English.

In our evaluation of the approach, we focus on the comparison to the Lucy baseline as our approach cannot be tuned with automated scoring metrics. Hence, it cannot be meaningfully compared to other systems in terms of BLEU scores.

3.1 Data sets

The WMT 2008 news test set of 2,051 sentences had been split into a development set of 1,025 sentences and a test set of 1,026 sentences. We used the development set data for the creation of the XML parser that approximates Lucy tree structures. We examined different combinations of features used in our substitution algorithm on the development data set.

3.2 Experimental results

In Table 1, we show the different feature configurations we tried. It is worth noting that each configuration performed better than the baseline, which was the Lucy RBMT system; this means that fragments from other systems actually did improve it. Table 2 presents results obtained from automated scoring metrics for the different system configurations applied on the development set data. Finally, in Table 3 we give the total number of substitutions that have been performed for each of the system configurations during our work on the development set.

Configuration	Matching POS?	Context
strict	yes	yes
pos	yes	no
context	no	yes
relaxed	no	no

Table 1: Feature configurations for experiments

Configuration	NIST	BLEU
baseline	5.0568	0.1516
strict	5.0937	0.1532
pos	5.0962	0.1534
context	**5.0984**	**0.1535**
relaxed	5.0932	0.1535

Table 2: Automated scoring results for development set.

The results shown in Table 2 indicate a possible improvement over the Lucy baseline. However, as the differences in BLEU between the configurations are not conclusive, we performed a manual evaluation of development set results. For example, the *context* feature disallows the substitution of *it is* ***saved*** by *it is* ***saves***. Removing this feature leads to

Configuration	# of substitutions
strict	412
pos	1,121
context	458
relaxed	1,317

Table 3: Substitution statistics for development set.

many more substitutions, which largely do not impact translation quality.

Based on our findings from the manual evaluation of development set results, we decided to use the *context* configuration in our final submission to the workshop. The context restriction includes part-of-speech matching implicitly, so adding this feature to the context restriction does not lead to any further improvements.

4 Conclusion

Whereas in previous work we only used translations generated by purely statistical MT systems as additional input, our system for the ML4HMT shared task could exploit output from systems of different paradigms. It remains to be investigated how this change affected resulting translation quality. The substitution approach showed improvements, although it was restricted to only single-word substitutions. In this hybrid setup we could retain the good syntactic structure of the RBMT output (which we used as translation template), while improving the lexical semantics by integrating translation fragments from other systems within the ML4HMT data set.

Future work includes expanding the substitution range to entire phrases and multi-word expressions. Restricting ourselves to single words has shown to help in retaining the good syntactic structure, but it also limits the impact of the additional systems on the baseline. By relaxing this restriction, we will open up our system to more extensive changes in the syntactic structure, which we will have to monitor closely to make sure we will not introduce translation candidates that will break the structure. Also, our features used for controlling the substitution algorithm are handcrafted at the moment; here we can see benefits from applying machine learning tools to actually learn helpful features from the given data. This will be an interesting extension of the system and would hopefully improve the substitution.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.: 249119). Part of this work was supported by the EuroMatrixPlus project (IST-231720) which is also funded by the European Community. We are grateful to two anonymous reviewers from LIHMT 2011 for their comments.

References

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain.

Juan A. Alonso and Gregor Thurmair. 2003. The Comprendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA.

Chris Callison-Burch and Raymond Flournoy. 2001. A program for automatically selecting the best output from multiple machine translation engines. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source SMT decoder. In *Proceedings of WMT07*, pages 193–196, Prague, Czech Republic, June. Association for Computational Linguistics.

Yu Chen, Michael Jellinghaus, Andreas Eisele, Yi Zhang, Sabine Hunsicker, Silke Theison, Christian Federmann, and Hans Uszkoreit. 2009. Combining multi-engine translations with moses. In *Proceedings of the*

Fourth Workshop on Statistical Machine Translation, pages 42–46. Association for Computational Linguistics, 3.

Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio, June. Association for Computational Linguistics.

Cristina Espana-Bonet, Gorka Labaka, Arantza Diaz de Ilarraza, Llus Màrquez, and Kepa Sarasola. 2011. Hybrid machine translation guided by a rule-based system. In *Proceedings of the 13th Machine Translation Summit*, pages 554–561, September.

Christian Federmann, Silke Theison, Andreas Eisele, Hans Uszkoreit, Yu Chen, Michael Jellinghaus, and Sabine Hunsicker. 2009. Translation combination using factored word substitution. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 70–74. Association for Computational Linguistics, 3.

Robert E. Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *ANLP*, pages 95–100.

Nizar Habash, Bonnie J. Dorr, and Christof Monz. 2009. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation*, 23(1):23–63.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 98–107, Honolulu, Hawaii, October. Association for Computational Linguistics.

Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proc. of EAMT*, Budapest, Hungary.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of Annual meeting of the Association for Computation Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech, June.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Gregor Leusch, Evgeny Matusov, and Hermann Ney. 2009. The RWTH system combination system for WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 51–55, Athens, Greece, March. Association for Computational Linguistics.

Wolfgang Macherey and Franz J. Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 986–995, Prague, Czech Republic, June. Association for Computational Linguistics.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–40, Trento, Italy, April.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. pages 440–447, Hongkong, China, October.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM.

Antti-Veikko I. Rosti, Spyridon Matsoukas, and Richard M. Schwartz. 2007. Improved word-level system combination for machine translation. In *ACL*.

Gregor Thurmair. 2009. Comparing different architectures of hybrid Machine Translation systems. In *MT Summit XII 2009*, August.

Results from the ML4HMT Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT

Christian Federmann
Language Technology Lab
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
cfedermann@dfki.de

Abstract

We describe the ML4HMT shared task which aims to foster research on improved system combination approaches for MT. Participants of the challenge are requested to build hybrid translations by combining the output of several MT systems of different types. We describe the ML4HMT corpus and the annotation format we have designed for it and briefly summarize the participating systems. Using automated metrics scores and extensive manual evaluation, we discuss the performance of the various systems. An interesting result from the shared task is the fact that we observed different systems winning according to the automated metrics and according to the manual evaluation. We conclude by summarising the first edition of the challenge and give an outlook to future work.

1 Introduction

The "Shared Task on Applying Machine Learning techniques to optimise the division of labour in Hybrid MT" is an effort to trigger systematic investigation on improving state-of-the-art Hybrid MT, using advanced machine-learning (ML) methodologies. Participants of the challenge are requested to build Hybrid/System Combination systems by combining the output of several MT systems of different types and with very heterogeneous types of metadata information, as provided by the organizers.

The main focus of the shared task is trying to answer the following question: *Could Hybrid/System Combination MT techniques benefit from extra information (linguistically motivated, decoding and runtime) from the different systems involved?*

Our research in work package 2 of the META-NET project focuses on the design and development of such advanced combination methods, building bridges to the machine learning community to foster joint and systematic exploration of novel system combination techniques; for this, we have collected translation output from various machine translation systems, including information such as part-of-speech, word alignment, or language model scores. The collected data has been released as a multilingual corpus[1]. Furthermore, we have organised a workshop including a challenge exploiting the ML4HMT corpus[2].

The remainder of this paper is structured as follows: in Section 2 we describe the data given to the shared task participants and give a detailed description of the challenge. Section 3 presents the systems taking part in the challenge before we present and discuss evaluation results in Section 4. We conclude by giving a summary of the ML4HMT shared task and an outlook to future work in Section 5.

2 Challenge Description

The participants are given a bilingual development set, aligned at a sentence level. For each sentence, the corresponding *bilingual data set* contains:

– the source sentence,

– the target (reference) sentence, and

[1] Data package available from http://www.dfki.de/~cfedermann/ML4HMT-data-1.0.tgz

[2] See http://www.dfki.de/ml4hmt/

– the corresponding multiple output translations from 5 different systems, based on different MT approaches.

For the ML4HMT data set we decided to use the following systems: Apertium (Ramírez-Sánchez et al., 2006), Joshua (Li et al., 2009), Lucy (Alonso and Thurmair, 2003), MaTrEx (Penkale et al., 2010), and Metis (Vincent Vandeghinste and Schmidt, 2008)). The output has been annotated with system-internal metadata information derived from the translation process of each of the systems.

2.1 Annotated Data Format

We have developed a new dedicated format derived from XLIFF (XML Localisation Interchange File Format) to represent and store the corpus data. XLIFF is an XML-based format created to standardize localization. It was standardized by OASIS in 2002 and its current specification is v1.2 released on Feb-1-2008 (`http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html`).

An XLIFF document is composed of one or more `<file>` elements, each corresponding to an original file or source. Each `<file>` element contains the source of the data to be localized and the corresponding localized (translated) data for one locale only. The localizable texts are stored in `<trans-unit>` elements each having a `<source>` element to store the source text and a `<target>` (not mandatory) element to store the translation.

We introduced new elements into the basic XLIFF format (in the "`metanet`" namespace) allowing a wide variety of meta-data annotation of the translated texts by different MT systems (tools). The tool information is included in the `<tool>` element appearing in the header of the file. Each tool can have several parameters (model weights) which are described in the `<metanet:weight>`.

Annotation is stored in `<alt-trans>` element within the `<trans-unit>` elements. The `<source>` and `<target>` elements in the `<trans-unit>` elements refer to the source sentence and its reference translation, respectively. The `<source>` and `<target>` elements in the `<alt-trans>` elements specifies the input and output of a particular MT system (tool). Tool-specific scores assigned to the translated sentence are listed in the `<metanet:scores>` element and the derivation of the translation is specified in the `<metanet:derivation>` element. Its content is tool-specific.

The full format specification is available as an XML schema. An example annotation from the ML4HMT data set is depicted in Figure 1.

2.2 Development and Test Sets

We decided to use the WMT 2008 (Callison-Burch et al., 2008) news test set as a source for the annotated corpus. This is a set of 2,051 sentences from the news domain translated to several languages, including English and Spanish but also others. The data was provided by the organizers of the Third Workshop on Machine Translation (WMT) in 2008. This data set was split into our own development set (containing 1,025 sentence pairs) and test set (containing 1,026 sentence pairs).

3 Participating Systems

3.1 DCU

The system described in Okita and van Genabith (2011) presents a system combination module in the MT system MaTrEx (Machine Translation using Examples) developed at Dublin City University. A system combination module deployed by them achieved an improvement of 2.16 BLEU (Papineni et al., 2001) points absolute and 9.2% relative compared to the best single system, which did not use any external language resources. Their system is based on system combination techniques which use a confusion network on top of a Minimum Bayes Risk (MBR) decoder (Kumar and Byrne, 2002).

One interesting, novel point in their submission is that for the given single best translation outputs, they tried to identify which inputs they will consider for the system combination, possibly discarding the worst performing system(s) from the combination input. As a result of this selection process, their BLEU score, from the combination of the four single best systems, achieved 0.48 BLEU points absolute higher than the combination of the five single best systems.

3.2 DFKI-A

A system combination approach with a sentence ranking component is presented in Avramidis (2011). The paper reports on a pilot study on a Hybrid Machine Translation that takes advantage of multilateral system-specific metadata provided as part of the shared task. The proposed solution offers a machine learning approach, resulting in a selection mechanism able to learn and rank and select systems' translation output on the complete sentence level, based on their respective quality.

For training, due to the lack of human annotations, word-level Levenshtein distance has been used as a (minimal) quality indicator, whereas a rich set of sentence features was extracted and selected from the dataset. Three classification algorithms (Naive Bayes, SVM and Linear Regression) were trained and tested on pairwise featured sentence comparisons. The approaches yielded high correlation with original rankings (tau=0.52) and selected the best translation on up to 54% of the cases.

3.3 DFKI-B

The authors of Federmann et al. (2011) report on experiments that are focused on word substitution using syntactic knowledge. From the data provided by the workshop organisers, they choose one system to provide the "translation backbone". The Lucy MT system was suited best for this task, as it offers parse trees of both the source and target side, which allows the authors to identify interesting phrases, such as noun phrases, in the source and replace them in the target language output. The remaining four systems are mined for alternate translations on the word level that are potentially substituted into the aforementioned template translation if the system finds enough evidence that the candidate translation is better. Each of these substitution candidates is evaluated concerning a number of factors:

– the part-of-speech of the original translation must match the candidate fragment.

– Additionally they may consider the 1-left and 1-right context.

– Besides the part-of-speech, all translations plus their context are scored with a language model trained on EuroParl (Koehn, 2005).

– Additionally, the different systems may turn up with the same translation, in that case the authors select the candidate with the highest count ("majority voting").

The authors reported improvements in terms of BLEU score when comparing to the translations from the Lucy RBMT system.

3.4 LIUM

Barrault and Lambert submitted results from applying the open-source MANY (Barrault, 2010) system on our data set. The MANY system can be decomposed into two main modules.

1. The first one is the alignment module which actually is a modified version of TERp (Snover et al., 2009). Its role is to incrementally align the hypotheses against a backbone in order to create a confusion network. Each hypothesis acts as backbone, yielding each the corresponding confusion network. Those confusion networks are then connected together to create a lattice.

2. The second module is the decoder. This decoder is based on the token pass algorithm and it accepts as input the lattice previously created. The costs computed in the decoder can be expressed as a weighted sum of the logarithm of feature functions. The following features are considered in decoding:

   – the language model probability, given by a 4-gram language model
   – a word penalty, which depends on the number of words in the hypothesis
   – a null-arc penalty, which depends on the number of null arcs crossed in the lattice to obtain the hypothesis
   – the system weights: each word receives a weight corresponding to the sum of the weights of all systems which proposed it.

4 Evaluation Results

To evaluate the performance of the participating sytems, we computed automated scores, namely BLEU, NIST, METEOR (Banerjee and Lavie, 2005), PER, Word error rate (WER) and Translation Error Rate (TER) and also performed an extensive, manual evaluation with 3 annotators ranking system combination results for a total of 904 sentences.

System	BLEU	NIST	METEOR	PER	WER	TER
DCU	**25.32**	**6.74**	56.82	**60.43**	**45.24**	**0.65**
DFKI-A	23.54	6.59	54.30	61.31	46.13	0.67
DFKI-B	23.36	6.31	**57.41**	65.22	50.09	0.70
LIUM	24.96	6.64	55.77	61.23	46.17	0.65

Table 1: Automated scores for ML4HMT test set.

4.1 Automated Scores

Results from running automated scoring tools on the submitted translations are reported in Table 1. The overall best value for each of the scoring metrics is print in **bold face**. Table 2 presents automated metric scores for the individual systems in the ML4HMT corpus, also computed on the test set. These scores give an indicative baseline for comparison with the system combination results.

4.2 Manual Ranking

The manual evaluation is undertaken using the Appraise (Federmann, 2010) system; a screenshot of the evaluation interface is shown in Figure 2. Users are shown a reference sentence and the translation output from all four participating systems and have to decide on a ranking in *best-to-worst order*. Table 3 shows the average ranks per system from the manual evaluation, again the best value per column is printed in **bold face**. Table 4 gives the statistical mode per system which is the value that occurs most frequently in a data set.

4.3 Inter-annotator Agreement

Next to computing the average rank per system and the statistical mode, we follow Carletta (1996) and compute κ scores to estimate the inter-annotator agreement. In our manual evaluation campaign, we had $n = 3$ annotators so computing basic, pairwise annotator agreement is not sufficient—instead, we apply Fleiss (1971) who extends Scott (1955) for computing inter-annotator agreement for $n > 2$.

Annotation Setup As we have mentioned before, we had $n = 3$ annotators assign ranks to our four participating systems. As ties were not allowed, this means there exist $4! = 24$ possible rankings per sentence (e.g., *ABCD, ABDC, etc.*)[3]. In a second evaluation scenario, we only collected the *1-best* ranking system per sentence, resulting in a total of four categories (A: ”system A ranked 1st”, etc.). In this second scenario, we can expect a higher annotator agreement due to the reduced number categories. Overall, we collected 904 sentences with an overlap of $N = 146$ sentences for which all annotators assigned ranks.

Scott's π allows to measure the pairwise annotator agreement for a classification task. It is defined as

$$\pi = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ represents the fraction of rankings on which the annotators agree, and $P(E)$ is the probability that they agree by chance. Table 5 lists the pairwise agreement of annotators for all four participating systems. Assuming $P(E) = 0.5$ we obtain an overall agreement π score of

$$\pi = \frac{0.673 - 0.5}{1 - 0.5} = 0.346 \tag{2}$$

which can be interpreted as *fair agreement* following Landis and Koch (1977). WMT shared tasks have shown this level of agreement is common for language pairs, where the performance of all systems is rather close to each other, which in our case is indicated by the small difference measured by automatic metrics on the test set (Table 1). The lack of ties, in this case might have meant an extra reason for disagreement, as annotators were forced to distinguish a quality difference which otherwise might have been annotated as “equal”. We have decided to compute Scott's π scores to be comparable to WMT11 (Bojar et al., 2011).

Fleiss κ Next to the π scores, there also exists the so-called κ score. Its basic equation is strikingly

[3] Given this huge number of possible categories, we were already expecting resulting κ scores to be low.

System	BLEU	NIST	METEOR	PER	WER
Joshua	19.68	6.39	50.22	47.31	62.37
Lucy	**23.37**	6.38	**57.32**	49.23	64.78
Metis	12.62	4.56	40.73	63.05	77.62
Apertium	22.30	6.21	55.45	50.21	64.91
MaTrEx	23.15	**6.71**	54.13	**45.19**	**60.66**

Table 2: Automated scores for baseline systems on ML4HMT test set.

System	Annotator #1	Annotator #2	Annotator #3	Overall
DCU	2.44	2.61	2.51	2.52
DFKI-A	2.50	2.47	2.48	2.48
DFKI-B	**2.06**	**2.13**	**1.97**	**2.05**
LIUM	2.89	2.79	2.93	2.87

Table 3: Average rank per system per annotator from manual ranking of 904 (overlap=146) translations.

similar to (1)

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3}$$

with the main difference being the κ score's support for $n > 2$ annotators. We compute κ for two configurations. Both are based on $n = 3$ annotators and $N = 146$ sentences. They differ in the number of categories that a sentence can be assigned to (k)

1. *complete* scenario: $k = 24$ categories. For this, we obtained an overall κ score of

$$\kappa_{complete} = \frac{0.1 - 0.054}{1 - 0.054} = 0.049 \tag{4}$$

2. *1-best* scenario: $k = 4$ categories. For the reduced number of categories, κ improved to

$$\kappa_{1-best} = \frac{0.368 - 0.302}{1 - 0.302} = 0.093 \tag{5}$$

It seems that the large number of categories of the *complete* scenario has indeed had an effect on the resulting $\kappa_{complete}$ score. This is a rather expected outcome, still we report the κ scores for future reference. The *1-best* scenario supports an improved κ_{1-best} score but does not reach the level of agreement observed for the π score.

It seems that DFKI-B was underestimated by BLEU scores, potentially due to its rule-based characteristics. This is a possible reason for the relatively higher inter-annotator agreement when compared with other systems. Also, DCU and LIUM may have low inter-annotator agreement as their background is similar.

Due to the fact that κ is not really defined for *ordinal data* (such as rankings in our case), we will investigate other measures for inter-annotator agreement. It might be a worthwhile idea to compute α scores, as described in Krippendorff (2004). Given the average rank information, statistical mode, π and κ scores, we still think that we have derived enough information from our manual evaluation to support for future discussion.

5 Conclusion

We have developed an Annotated Hybrid Sample MT Corpus which is a set of 2,051 sentences translated by five different MT systems[4] (Joshua, Lucy, Metis, Apertium, and MaTrEx). Using this resource we have launched the Shared Task on Applying Machine Learning techniques to optimise the division of labour in Hybrid MT (ML4HMT-2011), asking participants to create combined, hybrid translations using machine learning algorithms or other, novel ideas for making best use of the provided ML4HMT corpus data. Four participating combination systems, each following a different solution strategy, have been submitted to the shared task. We computed automated metric scores and conducted an extensive manual evaluation campaign to assess the quality of the hybrid translations. Interestingly,

[4] Not all systems available for all language pairs.

System	Ranked 1st	Ranked 2nd	Ranked 3rd	Ranked 4th	Mode
DCU	62	79	**97**	62	3rd
DFKI-A	73	65	**82**	80	3rd
DFKI-B	**127**	84	47	42	1st
LIUM	38	72	74	**116**	4th

Table 4: Statistical mode per system from manual ranking of 904 (overlap=146) translations.

Systems	π-Score	Systems	π-Score	Annotators	π-Score
DCU, DFKI-A	0.296	DCU, DFKI-B	0.352	#1,#2	0.331
DCU, LIUM	0.250	DFKI-A, DFKI-B	0.389	#1,#3	0.338
DFKI-A, LIUM	0.352	DFKI-B, LIUM	0.435	#2,#3	0.347

Table 5: Pairwise agreement (using Scott's π) for all pairs of systems/annotators.

the system winning nearly all the automatic scores (DCU) only reached a third place in the manual evaluation. Vice versa, the winning system according to manual rankings (DFKI-B) ranked last place in the automatic metric scores based evaluation. This clearly indicates that more systematic investigation of hybrid system combination approaches, both on a system level and wrt. the evaluation of such systems' output, needs to be undertaken. We have learned from the participants that our ML4HMT corpus is too heterogeneous to be used easily in system combination approaches; hence we will work on an updated version for the next edition of this shared task. Also, we will further focus on the integration of advanced machine learning techniques as these are expected to support better exploitation of our corpus' data properties. We are looking forward to an interesting workshop and thank the participants for their efforts during the ML4HMT-2011 Shared Task.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.: 249119). We thank the organisers of LIHMT 2011 for their support.

References

Juan A. Alonso and Gregor Thurmair. 2003. The Comprendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA.

Eleftherios Avramidis. 2011. DFKI System Combination with Sentence Ranking at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November. META-NET. to appear.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Loïc Barrault. 2010. MANY : Open source machine translation system combination. *Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for Machine Translation*, 93:147–155.

Ondrej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine*

Translation, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22:249–254, June.

Christian Federmann, Yu Chen, Sabine Hunsicker, and Rui Wang. 2011. DFKI System Combination using Syntactic Information at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November. META-NET. to appear.

Christian Federmann. 2010. Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *LREC*.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Klaus Krippendorff. 2004. Reliability in content analysis. some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Shankar Kumar and William Byrne. 2002. Minimum bayes-risk word alignments of bilingual texts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 140–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

J.R. Landis and G.G. Koch. 1977. Measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.

Tsuyoshi Okita and Josef van Genabith. 2011. DCU Confusion Network-based System Combination for ML4HMT. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November. META-NET. to appear.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM.

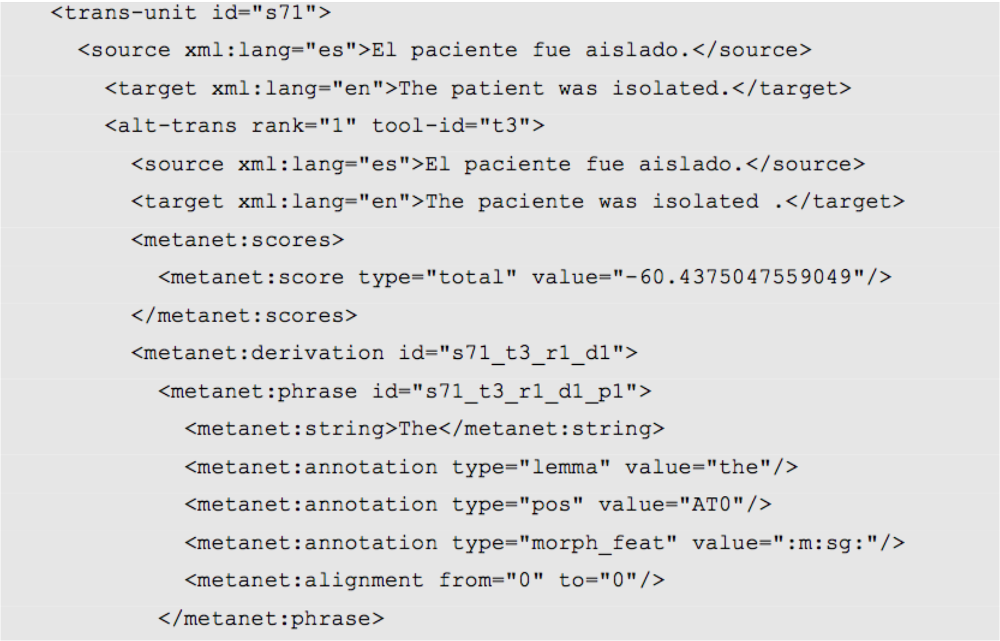
Sergio Penkale, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, and Andy Way. 2010. Matrex: the dcu mt system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 143–148, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. 2006. Opentrad apertium open-source machine translation system: an opportunity for business and research. In *Proceeding of Translating and the Computer 28 Conference*, November.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, 19(3):321–325.

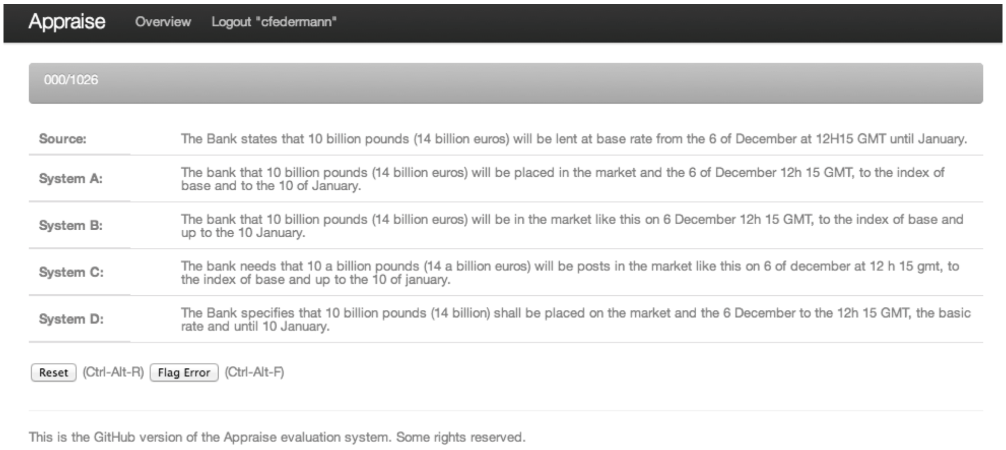
Matthew G. Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23:117–127, September.

Ineke Schuurman Stella Markantonatou Sokratis Sofianopoulos Marina Vassiliou Olga Yannoutsou Toni Badia Maite Melero Gemma Boleda Michael Carl Vincent Vandeghinste, Peter Dirix and Paul Schmidt. 2008. Evaluation of a machine translation system for low resource languages: Metis-ii. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.



```
<trans-unit id="s71">
  <source xml:lang="es">El paciente fue aislado.</source>
    <target xml:lang="en">The patient was isolated.</target>
    <alt-trans rank="1" tool-id="t3">
      <source xml:lang="es">El paciente fue aislado.</source>
      <target xml:lang="en">The paciente was isolated .</target>
      <metanet:scores>
        <metanet:score type="total" value="-60.4375047559049"/>
      </metanet:scores>
      <metanet:derivation id="s71_t3_r1_d1">
        <metanet:phrase id="s71_t3_r1_d1_p1">
          <metanet:string>The</metanet:string>
          <metanet:annotation type="lemma" value="the"/>
          <metanet:annotation type="pos" value="AT0"/>
          <metanet:annotation type="morph_feat" value=":m:sg:"/>
          <metanet:alignment from="0" to="0"/>
        </metanet:phrase>
```

Figure 1: Example of annotation from the ML4HMT corpus.



Source:	The Bank states that 10 billion pounds (14 billion euros) will be lent at base rate from the 6 of December at 12H15 GMT until January.
System A:	The bank that 10 billion pounds (14 billion euros) will be placed in the market and the 6 of December 12h 15 GMT, to the index of base and to the 10 of January.
System B:	The bank that 10 billion pounds (14 billion euros) will be in the market like this on 6 December 12h 15 GMT, to the index of base and up to the 10 January.
System C:	The bank needs that 10 a billion pounds (14 a billion euros) will be posts in the market like this on 6 of december at 12 h 15 gmt, to the index of base and up to the 10 of january.
System D:	The Bank specifies that 10 billion pounds (14 billion) shall be placed on the market and the 6 December to the 12h 15 GMT, the basic rate and until 10 January.

This is the GitHub version of the Appraise evaluation system. Some rights reserved.

Figure 2: Screenshot of the Appraise interface for human evaluation.

Author Index