5th SALTMIL Workshop on Minority Languages
Genoa, 23/5/2006

Statistical Machine Translation with a Small Amount of Bilingual Training Data

Maja Popović, Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany

Overview

- Introduction to Statistical Machine Translation (SMT)
- Motivation
- SMT with Sparse Training Data
- Recent Results
  - Spanish-English
  - Serbian-English
- Conclusions

Statistical Machine Translation (SMT)

- finding a target language sequence $\hat{e}_1^I$ given a source language sequence f_1^J:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}}\{Pr(e_1^I|f_1^J)\}$$

- log-linear model:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}}\{\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\}$$

  – language model

$$h_m(e_1^I, f_1^J) = \log Pr(e_1^I)$$

  – translation model

$$h_m(e_1^I, f_1^J) = \log Pr(f_1^J|e_1^I)$$

  – ...

Motivation

- **translation model probabilities are extracted from a bilingual parallel text - training corpus**
- **the quality of a translation system usually depends on the size of this corpus**

  - a large bilingual parallel corpus is often not available

  $\Rightarrow$ strategies for exploiting limited amounts of bilingual data for statistical machine translation

SMT with Sparse Bilingual Training Data

- **Al-Onaizan & Germann[+], 2000**
  - **comparing different translation methods on a small bilingual Tetun-English corpus**
  - **found out that the human mind is very capable of deriving dependencies such as morphology, cognates, proper names, etc.**

  $\Rightarrow$ **a crucial reason for better performance of human translation**

- **Callison-Burch & Osborne, 2003**
  - **co-training method for extension of a training corpus**
  - **new sentence pairs are produced by multiple translation models trained on different language pairs**

  $\Rightarrow$ **the best improvements achieved after two or three re-training rounds**

SMT with Sparse Bilingual Training Data

- **Niessen & Ney, 2004**
  - **investigating impact of the corpus size for translation from German into English**
  - **morpho-syntactic information and conventional dictionary are used for improving the performance**

  $\Rightarrow$ **acceptable translation quality even with a very small corpus**

- **Matusov & Popović$^+$, 2004**
  - **translation of spontaneous speech**
  - **acquiring additional training data using an n-gram coverage measure**
  - **morpho-syntactic information**

- **Popović & Ney, 2005**
  - **translation of Spanish-English and Catalan-English pair**
  - **phrasal lexicon and morpho-syntactic information for Spanish and Catalan verbs**

  $\Rightarrow$ **acceptable translation quality with only one thousand task-specific sentence pairs for training**

SMT with Sparse Bilingual Training Data

- **Popović & Vilar$^+$, 2005**
  - **translation of Serbian-English pair with a very small corpus**
  - **morpho-syntactic information and phrasal book**
  - **$\Rightarrow$ translation results comparable with results for other language pairs**

- **Goldwater & McClosky, 2005**
  - **translation of Czech-English pair**
  - **morphological information**

- **Lopez & Resnik, 2005; Martin & Mihalcea$^+$, 2005**
  - **word alignments for languages with scarce resources**
    - **Romanian-English**
    - **Inuktikut-English**
    - **Hindi-English**

Recent Results
- Experimental Settings -

Spanish-English and Serbian-English language pair

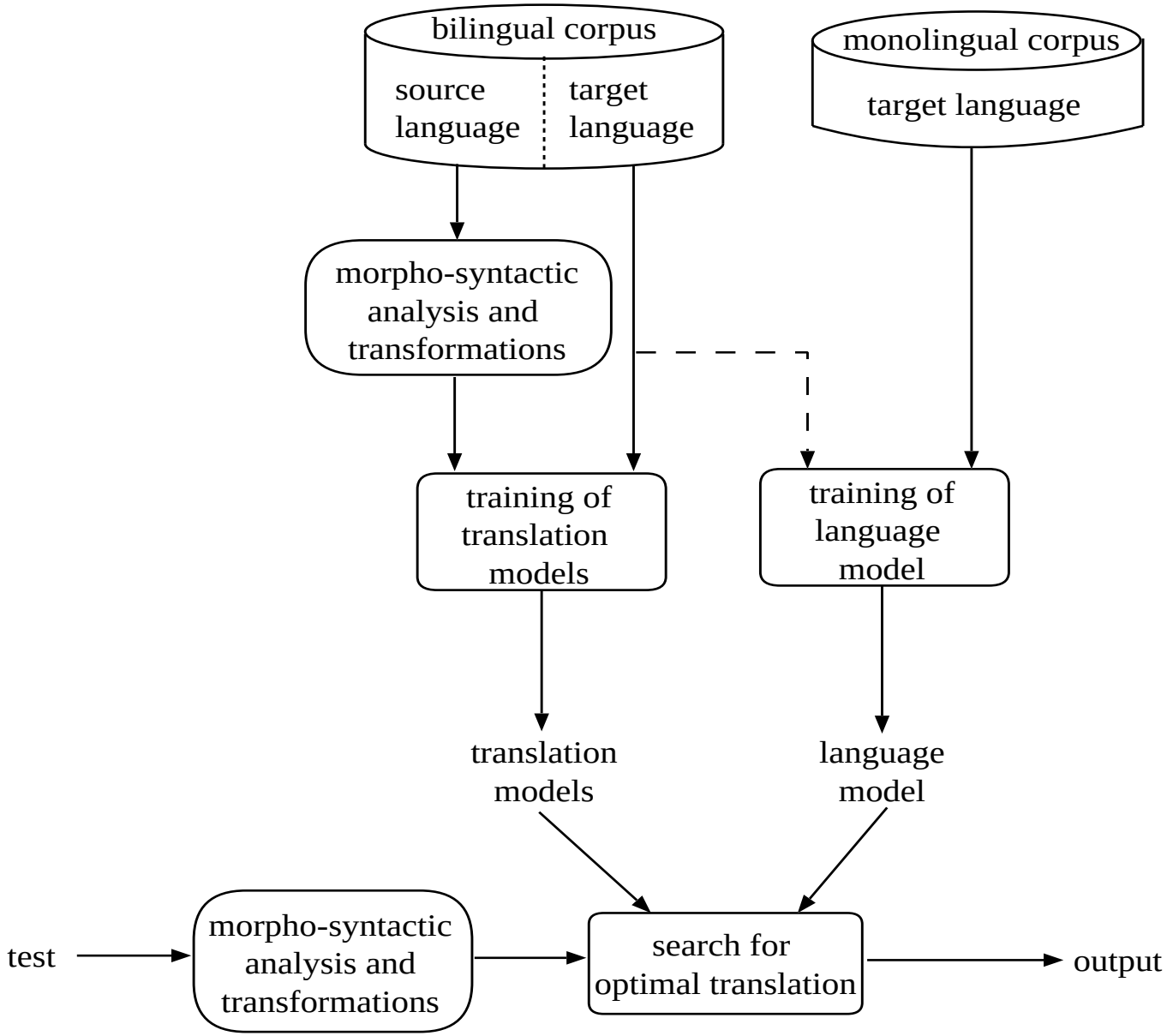
- **Spanish-English**
  - **European Parliament Plenary Sessions (EPPS) corpus**
  - **conventional dictionary**
- **Serbian-English**
  - **Assimil language course corpus**
  - **small phrasal book**

- **various sizes of bilingual training corpus**
- **appropriate morpho-syntactic transformations**
- **language model trained on the largest target language text**

Translation System

- state-of-the-art translation system
- log-linear combination of seven models:
  - two phrase-based models
    (source to target and target to source)
  - two single word based models at phrase level
    (source to target and target to source)
  - language model
  - phrase penalty and word penalty

Training and Test with Sparse Bilingual Resources



Spanish↔English
- Training Corpora -

Training		Spanish	English
1.3M	Sentences	1281427	
	Running Words+PM	36578514	34918192
	Vocabulary	153124	106496
	Singletons [%]	35.2	36.2
13k	Sentences	13360	
	Running Words+PM	385198	366055
	Vocabulary	22425	16326
	Singletons [%]	47.6	43.7
1k	Sentences	1113	
	Running Words+PM	31022	29497
	Vocabulary	5809	4749
	Singletons [%]	60.8	55.3
dict.	Entries	52566	
	Running Words+PM	60964	62011
	Vocabulary	31126	30761
	Singletons [%]	67.7	67.4

Spanish↔English
- Test Corpus -

Test	Spanish	English
Sentences	840	1094
Running Words+PM	22774	26917
Distinct Words	4081	3958
OOVs (1.3M) [%]	0.14	0.25
OOVs (13k) [%]	2.8	2.6
OOVs (1k) [%]	10.6	9.4
OOVs (dict.) [%]	19.1	16.2

Spanish↔English
- Morpho-syntactic transformations -

- **local reorderings of nouns and adjectives**
- **replacing Spanish adjectives with their base forms**

Spanish	**original:**	**motivos económicos y políticos**
	reordered:	**económicos y políticos motivos**
	+adjective base	**económico y político motivos**
English	**original:**	**economic and political reasons**
	reordered:	**reasons economic and political**

Spanish→English
- Translation Results -

Spanish→English		WER	PER	BLEU
dict	baseline	60.4	49.3	19.4
	+adjective treatment	56.4	46.8	23.8
1k	baseline	52.4	40.7	30.0
	+dictionary	48.0	36.5	36.0
	+adjective treatment	44.5	34.8	40.9
13k	baseline	41.8	30.7	43.2
	+dictionary	40.6	29.6	46.3
	+adjective treatment	38.3	29.0	49.6
1.3M	baseline	34.5	25.5	54.7
	+reorder adjective	33.5	25.2	56.4

- dictionary alone might be used for multilingual information retrieval
- reasonable translation quality with small corpora
  - dictionary and morpho-syntactic information are very important
- 1000 times larger corpus ⇔ 12-25% relative decrease of error rates

English→Spanish
- Translation Results -

English→Spanish		WER	PER	BLEU
dict	baseline	67.6	55.9	14.1
	adjective treatment	65.7	54.5	16.5
1k	baseline	60.1	47.4	23.9
	+dictionary	56.0	43.2	28.3
	adjective treatment	53.9	42.0	30.6
13k	baseline	49.6	37.4	36.2
	+dictionary	48.6	36.3	37.2
	adjective treatment	47.3	35.7	39.1
1.3M	baseline	39.7	30.6	47.8
	+reorder adjective	39.6	30.5	48.3

- similar effect as for the other translation direction
- improvements through dictionary and morpho-syntactic information are slightly smaller
  - translation into a more inflected language is more difficult
  - Spanish has a rather free word order

Serbian↔English
- Training Corpora -

Training		Serbian	English
2.6k	Sentences	2632	
	Running Words+PM	22227	24808
	Vocabulary	4546	2645
	Singletons [%]	60.0	45.8
0.2k	Sentences	200	
	Running Words+PM	1666	1878
	Vocabulary	778	603
	Singletons [%]	79.4	65.5
phrases	Entries	351	
	Running Words+PM	617	730
	Vocabulary	335	315
	Singletons [%]	71.3	66.3

Serbian↔English
- Test Corpus -

Test	Serbian	English
Sentences	260	
Running Words+PM	2100	2336
Distinct Words	891	674
OOVs (2.6k) [%]	11.7	4.9
OOVs (0.2k) [%]	35.2	21.8

Serbian↔English
- Morpho-syntactic transformations -

- converting Serbian words into base forms

mali mala malog malu malom ...	small	⇒	mali	small

- additonal treatment of Serbian verbs

idemo	⇒	PL1 ići	we	go
idem	⇒	SG1 ići	I	go

- removing English articles

when I have the flu, I keep a supply of paper handkerchiefs.

⇓

when I have flu, I keep supply of paper handkerchiefs.

Serbian↔English
- Translation Results -

Serbian→English		WER	PER	BLEU
0.2k	baseline	65.5	60.8	8.3
	+phrases	65.0	59.8	10.3
	+base forms	59.2	54.8	13.9
	+verb POS+neg	60.0	52.6	14.8
2.6k	baseline	44.5	37.9	32.1
	+base forms	42.9	37.4	35.4
	+verb POS+neg	41.9	34.7	34.6

- results for extremely small corpus comparable with results for a dictionary
- phrases are helpful to some extent
- morphological information is very important
- acceptable performance with less than three thousand sentence pairs

English↔Serbian
- Translation Results -

English→Serbian		WER	PER	BLEU
0.2k	baseline	73.4	68.4	6.8
	+phrases	71.9	67.5	9.3
	+remove article	66.7	62.2	9.4
2.6k	baseline	51.8	45.8	23.1
	+remove article	50.4	44.6	24.6

- higher error rates due to the rich morphology and free word order
- phrases are more important for this translation direction
- removing English articles is helpful

Conclusions

- **an acceptable translation quality can be achieved with a very small amount of task-specific parallel text, especially if**
  - **– conventional dictionaries and/or phrasal books**
  - **– morpho-syntactic knowledge**

  are available

- **translation systems built only on**
  - **– conventional dictionary**
  - **– phrasal book**
  - **– extremely small parallel corpus**

  might be useful for document classification or multilingual information retrieval

References

- **Al-Onaizan & Germann[+], 2000.**
  Translation with scarce resources.
  ***17th National Conference on Artificial Intelligence (AAAI)*,**
  pages 672–678, Austin, TX, August.
- **Callison-Burch & Osborne, 2003.**
  Co-training for statistical machine translation.
  ***6th Annual CLUK Research Colloquium*, Edinburgh, UK, January.**
- **Goldwater & McClosky, 2005**
  Improving statistical machine translation through morphological analysis.
  ***Conf. on Empirical Methods for Natural Language Processing (EMNLP)*,**
  Vancouver, Canada, October.
- **Lopez & Resnik, 2005.**
  Improved HMM alignment for languages with scarce resources.
  ***ACL Workshop on Builiding and Using Parallel Texts: Data-Driven Machine Translation and Beyond*,**
  pages 83–86, Ann Arbor, MI, June.

References

- **Martin & Mihalcea$^+$, 2005.**
  Word alignments for languages with scarce resources.
  ***ACL Workshop on Builiding and Using Parallel Texts: Data-Driven Machine Translation and Beyond*,**
  pages 65–74, Ann Arbor, MI, June.
- **Matusov & Popović$^+$, 2004.**
  Statistical machine translation of spontaneous speech with scarce resources.
  ***Int. Workshop on Spoken Language Translation (IWSLT)*,**
  pages 139–146, Kyoto, Japan, September.
- **Niessen & Ney, 2004.**
  Statistical machine translation with scarce resources using morpho-syntactic information.
  ***Computational Linguistics*, 30(2):181–204**

References

- **Popović & Ney, 2005.**
  Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data.
  ***10th Annual Conf. on the European Asssociation for Machine Translation (EAMT)*,**
  pages 212–218, Budapest, Hungary, May.
- **Popović & Vilar+, 2005.**
  Augmenting a small parallel text with morpho-syntactic language resources for Serbian-English statistical machine translation.
  ***ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*,**
  pages 41–48, Ann Arbor, MI, June.
- **Popović & Ney, 2006.**
  POS-based word reorderings for statistical machine translation.
  ***5th Int. Conf. on Language Resources and Evaluation (LREC)*,**
  Genoa, Italy, May.