Report on the 9th SaLTMiL Workshop on "Free/open-Source Language Resources for the Machine Translation of Less-Resourced Languages"

by Mikel L. Forcada
Tuesday, 27 May 2014.
Reykjavik (Iceland)
(Proceedings)

The 9th SaLTMiL workshop on Free/open-source language resources for the machine translation of less-resourced languages, held as part of LREC 2014 in Reykjavík on May 27, 2014, from 09:30 to 13:30, was a very well attended event. About 40 people were present, more than the 31 attendees registered as of 22nd May, 2014. After a brief welcoming address by Mikel Forcada, there were two oral sessions, interrupted by the coffee break. Both sessions ran very smoothly, with plenty of questions asked from the audience.

Iñaki Alegria, Unai Cabezon, Unai Fernandez de Betoño, Gorka Labaka, Aingeru Mayor, Kepa Sarasola and Arkaitz Zubiaga
Wikipedia and Machine Translation: killing two birds with one stone

Gideon Kotzé and Friedel Wolff
Experiments with syllable-based English-Zulu alignment

In the second session, chaired by Trond Trosterud, the paper by Matthew Marting and Kevin Unhammer was presented by Francis Tyers as the authors could not make it to Iceland.

Inari Listenmaa and Kaarel Kaljurand
Computational Estonian Grammar in Grammatical Framework

Matthew Marting and Kevin Unhammer
FST Trimming: Ending Dictionary Redundancy in Apertium

Hrvoje Peradin, Filip Petkovski and Francis Tyers
Shallow-transfer rule-based machine translation for the Western group of South Slavic languages

Alex Rudnick, Annette Rios Gonzales and Michael Gasser
Enhancing a Rule-Based MT System with Cross-Lingual WSD

The workshop ended with a 30-minute general discussion, moderated by Francis Tyers. Two main questions were posed to the audience:

1. Is research in minority-language machine translation already mainstream
2. What are the main difficulties in building or putting together free/open-source language resources for small languages, and how should they be addressed? Are we pooling these resources correctly?

The audience was also invited to openly discuss other issues if necessary. Here is a detailed summary of what was discussed.

Question (1) quickly turned into a discussion on research about rule-based systems. Lori Levin said that minority languages are becoming mainstream and researchers are publishing in journal venue, that we need to educate people on the research issues related to rule-based language resources. She added that "It seems that tinkering with statistical models is research whereas tinkering with rules is not".

Robert Frederking: It is hard to publish papers on rule-based machine translation, at least in America. Francis Tyers replied that it may be easier in Europe, perhaps because European and American funding objectives are different.

Someone [not identified in Mikel Forcada's notes: apologies!] said that linguistic research may help facing some issues. Lori Levin said that the problem is partly linguistic and partly not, and the key is where to spend time in rule-based machine translation for maximum impact.

Francis Tyers mentions that the statistical machine translation community is strong partly because they use standardized evaluation measures. Antonio Toral mentioned that most papers in machine translation conferences are on statistical machine translation, and improvements reported are usually less than one BLEU point, but added that, in view of the results of a workshop on machine translation evaluation held the day before, there is very little correlation between BLEU score and productivity gains. Maja Popović added that there is a tendency in statistical machine translation towards morphologically rich, under-resourced languages. Trond is sceptical about research that does not take into account existing morphologies for these languages or does not aim at developing them. Maja Popović adds that it is better to have something than nothing and that all knowledge should be combined.
The role of linguists and their involvement is also discussed. Jonathan Washington explains his experience as a linguist getting involved in morphologies for Turkic languages and the

issues faced. Lori Levin says that we should get more linguists involved, but that rule writing is a skill that not everyone has; many people think they can, but we should educate linguists to be rule writers. Also, on interaction between linguists and prominent statistical machine translation researchers, she says that they are very busy people and that it is quite hard to get into their schedule to discuss these issues.

Laurette Pretorius speaks from a South African context. She says that computational linguistics is not taught in South African Universities, and stresses the importance of collaboration. She says that linguists should not be assigned the boring tasks, such as annotation tasks, but that they should be involved in the whole design. Francis Tyers also talks about the choice between doing tedious annotation or more interesting rule writing and understand that people would rather prefer the second. Mikel Forcada warns about the fact that linguists tend to get carried away by the low-frequency "jewels" of their languages and lose sight of the high-frequency "building blocks" needed for working systems.

Christian Buck returns to the fact that the statistical machine translation community has yearly "shoot-outs" (contests) where they can test their advancement, and that these contests do drive their research.

Jonathan Washington mentions that the computational perspective made him rethink many of the issues relating Turkic languages. Mikel mentions that in fact, computational linguistics descriptions are the best descriptions of language sometimes, and mentions the IXA Group's "computational morphology of Basque" or Elaine Uí Dhonnchadha's Irish morphology as the best description of their languages' morphologies.

Lori Levin stresses the fact that linguists have to be trained to do the linguistic engineering. For instance, lexical-functional grammars may teach aspects such as modularity.

Mikel L. Forcada mentions two problems in rule-based research: one, that rule-based MT as a field is very fragmented after the pervasive irruption of statistical machine translation, and, as a result, we do not speak with one voice and use inconsistent terminologies which make it very difficult to articulate ourselves as a field. Another one is reproducibility: for our rule-based research to be reproducible we have to make it all available, and this naturally leads to free/open-source licensing.

Sjur Moshagen talks about the fact that resources should be reusable in other language technologies and explains that in Tromsø they had no option as they were building the only set of resources for Sámi languages and they had to be reusable, and that linguists had to be trained in engineering issues. In fact, one of the uses was the Apertium machine translation systems for Sámi languages.

Sjur Moshagen opens question (2) about pooling and resource sharing and asks where would be a good place to take all these resources. On reproducibility, it is discussed that commercial rule-based systems often do not even want to be mentioned by name and ask to be called "System A" or "System B" in contests. Some judge this to be unfortunate.

Mikel Forcada mentions that pooling should pay special attention to metadata describing how to use the resources. He says that this is a more difficult problem than licensing.

Friedel Wolff questions whether licensing is really an easy problem and talks about incompatibilities among licenses such as Creative Commons and the General Public License, and says that when it comes to license derived data, decisions may be far from being trivial.

The discussion stops here and moderator Francis Tyers thanks everyone for the rich discussion. Mikel L. Forcada thanks everyone for attending and closes the Workshop.