English-Slovenian Statistical Machine Translation: from a Lower- to a Highly-Inflected Language

Jerneja Žganec Gros, Stanislav Gruden

Alpineon R&D
Ulica Iga Grudna 15, SI-1000 Ljubljana
Slovenia
jerneja@alpineon.si

Abstract

Freely available tools and language resources were used to build the VoiceTRAN statistical machine translation (SMT) system. Various configuration variations of the system are presented and evaluated. The VoiceTRAN SMT system outperformed the baseline conventional rule-based MT system in both English-Slovenian in-domain test setups. To further increase the generalization capability of the translation model for lower-coverage out-of-domain test sentences, an "MSD-recombination" approach was proposed. This approach not only allows a better exploitation of conventional translation models, but also performs well in the more demanding translation direction; that is, into a highly inflectional language. Using this approach in the out-of-domain setup of the English-Slovenian JRC-ACQUIS task, we have achieved significant improvements in translation quality.

Introduction

Machine translation (MT) systems automatically convert text strings from a source language (SL) into text strings in the target language (TL). They often allow for customization by application domain (e.g., weather), which improves the output by limiting the scope of allowable substitutions. This technique is particularly effective in domains in which formal or formulaic language is used, and therefore machine translation of government and legal documents more readily produces usable output than translation of less standardized texts or even spoken language.

Some initial machine translation attempts have been reported for translation from Slovenian into English (Vičič, 2002; Romih & Holozan, 2002; Žganec Gros et al., 2006; Sepesy Maučec et al., 2006). However, very little has been done for the opposite translation direction, from English into Slovenian (Žganec Gros et al., 2006). We have performed experiments in both translation directions, in which especially the latter proved to be a demanding task due to the highly inflectional nature of Slovenian.

This paper continues with a description of the VoiceTRAN statistical machine translation (SMT) experiment. The goal was to evaluate the performance of various SMT configuration variations against the performance of a baseline conventional rule-based MT system in order to find out, which MT system should be used in the VoiceTRAN speech-to-speech translation system.

Language Resources

The following language resources were used in the experiments. The bilingual language resources always refer to the English-Slovenian language pair:

· bilingual text corpora: the VoiceTRAN application-specific corpus and two freely available corpora: the JRC-ACQUIS corpus (Steinberger et al. 2006) and the IJS-ELAN corpus (Erjavec, 2002);

· the monolingual FDV-IJS Slovenian corpus, collected at the University of Ljubljana, and annotated within the VoiceTRAN project;

· English-Slovenian conventional dictionaries: an in-domain dictionary of military terminology and a conventional general dictionary (Korošec, 2002).

The monolingual and bilingual corpora were automatically annotated with context-disambiguated lemmas and morphosyntactic descriptions (MSDs), which included part-of-speech (POS) information (Erjavec & Džeroski, 2004).

The VoiceTRAN SMT System

The usual freely available tools were used to build the VoiceTRAN SMT system:

· the GIZA++ toolkit (Och & Ney, 2003) was used for training the VoiceTRAN translation model;

· the CMU-SLM toolkit (Rosenfeld, 1994) was used for building the language model, and

· the ISI ReWrite Decoder (Germann, 2003) was applied for translating the test sentences.

The "MSD recombination" SMT Approach

Bilingual training data are needed to train a SMT translation model that is then able to generalize and translate new sentences. Due to the statistical nature of system training words that appear more frequently in the training corpora are more likely to develop a suitable statistical model, whereas rare words tend to be overlooked. If the training data contain too many words of the latter type, the resulting models do not perform well due to data sparsity.

This effect is even more pronounced when translating from less-inflected into more highly-inflected languages (e.g., from English to Slovenian), when a word in the source language can be translated by many words in the target language depending on the context.

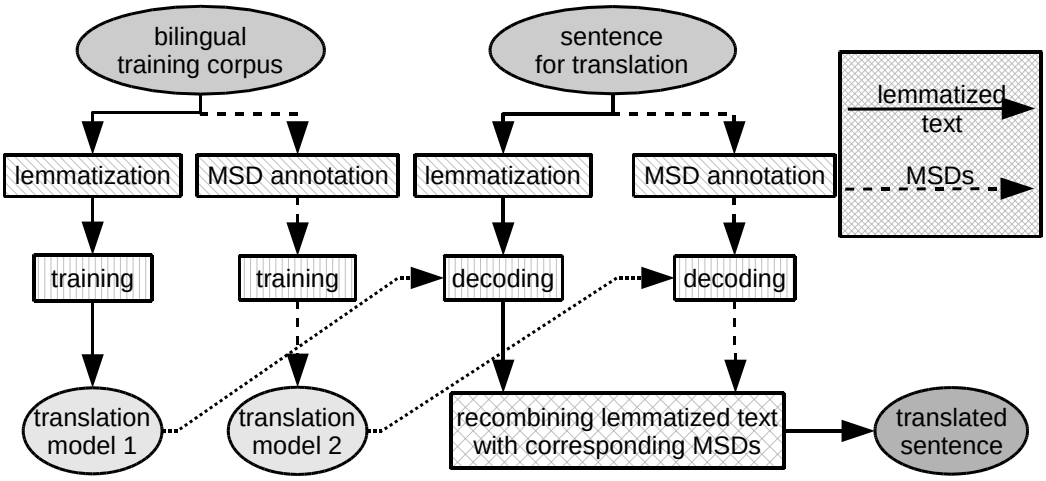

Figure 1: Description of the "MSD recombination" statistical machine translation approach.

There are several ways to tackle this problem. For example, we can build translation models based on lemmas only. This can work to a certain extent when translating in the opposite direction: from a highly-inflected language to a less-inflected one (e.g., from Slovenian into English) (Vičič, 2002; Sepesy Maučec et al., 2006). The lemmatized English target output often matches the correct reference translation. In contrast, when translating from English into Slovenian, translation accuracy through relying on lemmas only is rather poor.

The monolingual and bilingual corpora used in our experiments are equipped with MSD annotations (including POS information) and can be exploited to further improve the translations. Some attempts have been reported regarding the use of larger tokens in the training phase, consisting of MSDs (and POS), concatenated to their corresponding lemmas (Vičič, 2002; Sepesy Maučec et al., 2006). However, the data sparsity did not decrease significantly.

We used a different approach, which we call "MSD recombination." First, two corpora were derived from the initial training corpus.

In the first corpus, the sentences in both languages were preprocessed so that all word forms were replaced by their lemmas, using the lemmatization information provided in the source corpus. In order to derive the second corpus, all original word forms were replaced by their corresponding MSDs. These two corpora were then separately fed into the SMT training system.

The decoding was performed as shown in Figure 1: every test sentence was preprocessed into two sentences, in which words were replaced by lemmas in the first sentence, and by relevant MSDs in the second sentence.

Then we traced how each "lemma + MSD" pair in the source language transformed into a corresponding "lemma + MSD" pair in the target language. The resulting "lemma + MSD" pair was ultimately recombined to construct the final word form in the target language. The optimal number of MSD classes has been determined experimentally.

Experiment Setup

We describe two SMT experiments with different test and training setups: the ACQUIS corpus experiment and the VoiceTRAN corpus experiment.

In both experiments, the performance of the simple translation method, in which the sentences used for training the SMT system were taken directly from the original corpus without any prior modifications, was compared against the performance of the "MSD recombination" approach, and a baseline rule-based translation system, Presis (Romih & Holozan, 2002). Presis is a commercial conventional bidirectional rule-based translation system for the language pair Slovenian-English. It had been adapted to the tested application domain by upgrading the lexicon with application-specific terms.

JRC-ACQUIS Experiment

The first SMT experiment was performed on the English-Slovenian part of the JRC-ACQUIS corpus (Steinberger et al., 2006), which is equivalent to the SVEZ-IJS corpus. All tokens contain automatically assigned context-disambiguated lemmas and MSD (including POS) annotations. Sentences longer than 25 words were discarded from the training corpus as were the test sentences. The final bilingual training corpus contained 127,475 sentences with approximately 1.04 million tokens in the Slovenian part and 1.23 million tokens in the English part. A conventional general dictionary with 140,000 translation pairs was added as an option. The Slovenian language model was trained on the Slovenian part of the corpus.

All test sentences were excluded from the training material for the translation and language models. The first set of test sentences (the "in-domain test set") was selected directly from the corpus at regular intervals, resulting in 1,000 test sentences containing 8,113 tokens in the Slovenian part and 9,508 tokens in the English part.

For the second test we used test sentences selected from one of the components of the IJS-ELAN corpus, the ORWL file (Orwell's 1984), which is a text type significantly different from the rest of the ACQUIS corpus; hence we refer to it as the "out-of-domain test set." We randomly selected 1,000 sentences, containing 10,622 tokens in the Slovenian part and 12,363 tokens in the English part.
This setup enabled us to evaluate the system performance with test sentences that were highly correlated to the training data ("in-domain setup"), as well as with those that had low correlation to the training material ("out-of-domain setup").

VoiceTRAN Experiment

The second SMT experiment was performed on the first version of the restricted-domain VoiceTRAN parallel corpus, collected within the VoiceTRAN project and limited to government texts of the Slovenian Ministry of Defense. In comparison to the ACQUIS corpus, the VoiceTRAN sentences are more homogeneous and cover a more compact domain. Again, all tokens are annotated with automatically assigned context-disambiguated lemmas and MSD (and POS) information.
The translation model was trained with 2,508 sentences, containing 23,100 tokens in the Slovenian part and 26,900 tokens in the English part. A conventional bilingual dictionary of military terminology with 18,423 entries – including multiword expressions – was added as an option.
The Slovenian language model was trained on the Slovenian part of the VoiceTRAN corpus, with the optional addition of the FDV-IJS Slovenian monolingual corpus from the same application domain, containing 302,000 sentences and 3.19 million tokens.
The test sentences for the "in-domain" VoiceTRAN corpus experiment were selected from the VoiceTRAN corpus at regular intervals, resulting in 278 test sentences with 2,554 tokens in the Slovenian part and 2,951 tokens in the English part. Due to the modest size of the VoiceTRAN corpus, no "out-of-domain" tests were performed.

Performance Evaluation

To measure the "closeness" between the MT-generated hypothesis and human reference translations, standard objective MT metrics were used:

· GTM: General Text Matcher (Turian et al., 2005),

· NIST: a metric proposed by NIST (Doddington, 2002), and

· METEOR: Metric for Evaluation of Translation with Explicit ORdering (Banrjee & Lavie, 2005).

The scores obtained for the BLEU (Papineni, 2001) metric were too small to be considered reliable and they are not presented in tise paper.
The SMT evaluation efforts were centered on three system variation impacts:

1. the impact of the choice of the SMT approach: "simple" vs. "MSD recombination,"

2. the impact of conventional dictionaries: a general dictionary in the JRC-ACQUIS experiment and an in-domain terminology dictionary in the VoiceTRAN experiment,

3. the impact of the application domain to which the test sentences belong: in-domain, out-of-domain.

In both tests, the performance of different configurations of the VoiceTRAN SMT system was compared to the efficiency of the baseline Presis rule-based system.

JRC-ACQUIS Test Results

Table 1 presents the average evaluation scores (GTM, NIST, and METEOR) of the tested SMT system configuration versus the baseline system, both for the "in-domain" as well as for the "out-of-domain" setup.

"In-domain" setup	GTM	NIST	METEOR
SMT: "simple"	0.36	0.91	0.29
SMT: "MSD recombination"	0.33	0.78	0.26
baseline: Presis	0.29	0.71	0.23
"Out-of-domain" setup	GTM	NIST	METEOR
SMT: "simple"	0.17	0.52	0.11
SMT: "MSD recombination"	0.18	0.54	0.12
baseline: Presis	0.32	0.97	0.23

Table 1: JRC-ACQUIS evaluation scores.

In the "in-domain setup," the simple SMT approach performed slightly better than the "MSD recombination" method. Both SMT configurations outperformed the baseline rule-based MT system.
As expected, the baseline rule-based system was the best in the "out-of-domain setup," whereas here the "MSD recombination" SMT system performed better than the simple SMT system. Because the sources of the training data had been automatically tagged with lemmas and MSDs, the resulting imperfections in the training material had negative effects, especially on the "MSD recombination" translation method results. Therefore, we intend to retag the source corpora in the continuation of the project.

VoiceTRAN Test Results

Table 2 presents the average evaluation scores of the tested SMT system and training set configuration versus the rule-based baseline system for the VoiceTRAN test setup.

VoiceTRAN setup	GTM	NIST	METEOR
SMT: no dictionary	0.27	0.65	0.27
SMT: in-domain dictionary	0.40	0.86	0.35
baseline: domain-adapted Presis	0.31	0.69	0.26

Table 2: VoiceTRAN evaluation scores.

Both tested versions of the VoiceTRAN SMT system performed better than the baseline rule-based system, which had also adapted to the VoiceTRAN application domain prior to the experiments.

The addition of the in-domain terminology dictionary resulted in a substantial drop in the WER score and a rise in the other metric scores.
Due to the modest size of the VoiceTRAN corpus, only in-domain tests were performed and only the simple translation method was evaluated. Clearly, the first version of the VoiceTRAN corpus does not contain enough training material, and therefore an upgrade of the corpus is planned.

Discussion

The values obtained for measuring translation quality using the standard metrics mentioned were generally low. However, they served well enough to indicate whether we have improved the system by introducing changes into individual components of the SMT system.

Conventional Bilingual Dictionaries

In contrast to the first experiment, in which the inclusion of a complete general dictionary into the training material did not result in any significant translation performance improvement, in the second experiment the inclusion of an application-specific terminology dictionary considerably improved the translation accuracy and even outperformed the rule-based baseline system (Table 2), which had also been adapted to the application domain.
This can be explained in the following way. First, a large general dictionary introduces many new words that must be incorporated into the statistical model, thus relatively weakening the relations between words in phrases derived from the bilingual corpus. Often frequent words in a general dictionary have several translation candidates. Which candidate is the correct one depends on the context, and the proper context often has not been encountered in the training material and is not available from the lexical entries.
On the other hand, an application-specific terminological dictionary has a much broader coverage of the target area. Further, the number of multiple translation choices for a single lexical entry is usually much lower than in a general dictionary and, most importantly, the translation choices are all relevant to the application domain.
Therefore, we may draw the following conclusions. The use of a conventional bilingual dictionary, if available, can help bootstrap the training of the statistical alignment models and also helps to cover vocabulary that does not occur in the training corpus. Ideally, this dictionary is an application-specific terminology dictionary and would thus include entries relevant for the specific domain.
However, the available dictionary is often not domain-specific, which leads to the problem that out-of-domain lexicon entries may overshadow the in-domain lexicon entries learned from the training corpus. A possible solution is to extract those lexicon entries from the general purpose lexicon that are relevant in the domain – for example, by extracting all lexicon entries that really co-occur in the bilingual training corpus, e.g., by using the data structure of suffix arrays as proposed in Och (2002).

Choice of Translation Method

In comparison to the simple translation method, the "MSD recombination" translation method did not perform well for test sentences extracted from the unprocessed corpus in the first experiment. Similar results were obtained in the VoiceTRAN experiment on VoiceTRAN corpus test sentences. In both cases, the test sentences were in-domain.
The "MSD recombination" method performed better when ORWL test sentences were used, proving its potential for translation of out-of-domain sentences.
The simple translation method apparently adapted well to inflected Slovenian words, some of which were frequent enough in the training material to allow for sufficient training of the statistical model. As a consequence, when testing in-domain test sentences well correlated to the training corpus, the test set translations were translated rather well.
As expected, the "MSD recombination" translation method performed better when translating texts that were very different from the training sentence set, as was the case with the ORWL test corpus.

General Observations

In both in-domain test setups, the VoiceTRAN SMT system outperformed the baseline conventional rule-based MT system. On the other hand, in the out-of-domain test setup the performance of the "MSD recombination" SMT system was also not far behind the baseline system. Therefore we may conclude the following: when lexical coverage of a SMT system has dropped sufficiently (approaching the out-of-domain scenario), the simple translation model can be replaced by the "MSD recombination" translation model, especially when translating into a highly inflected language.
Finally, we would like to mention that we found that the applied evaluation metrics are not suitable for evaluating translations from English into Slovenian, i.e. from a low-inflected source language into a highly-inflected target language.
These metrics are all based on an exact comparison of entire words, which works well for English. Due to the rich inflectional paradigms in Slovenian, words that are semantically correctly translated but have the wrong ending receive a calculated score of zero.
For example, a method that attributes score points for finding a correct word stem in the target language would provide a much better translation quality estimation. Therefore, we will try to implement a language-adapted version of the METEOR metric by plugging in a stemmer and a stop list.
Nevertheless, the evaluation methods used were suitable for the purposes of our research because we were only looking for an indicator showing improvement or deterioration when using various MT systems and training set configurations.

Conclusions

Various configuration variations of the VoiceTRAN SMT system were presented and evaluated. In both in-domain test setups, the VoiceTRAN SMT system outperformed the baseline conventional rule-based MT system.
To increase the generalization capability of the translation model for lower-coverage out-of-domain test sentences, an "MSD-recombination" approach was proposed. This approach not only allows a better exploitation of conventional translation models, but also performs well in the more demanding translation direction; that is, into a highly inflected language. Using this approach in the out-

of-domain setup of the English-Slovenian JRC-ACQUIS task, we have achieved significant improvements in translation quality.

Acknowledgements

The work presented in this paper was performed as part of the VoiceTRAN project, contract number M2-0019, supported by the Slovenian Ministry of Defence and the Slovenian Research Agency.

References

Banerjee, S. and Lavie, A. (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 at the 43rd Annual Meeting of the Association of Computational Linguistics. Ann Arbor, Michigan.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality using N-gram Cooccurrence Statistics. Proceedings of the 2nd Human Language Technologies Conference. San Diego.

Erjavec, T. (2002). The IJS-ELAN Slovene-English parallel corpus. International Journal on Corpus Linguistics (pp. 1-20), Vol. 7. No. 1.

Erjavec, T. & Džeroski, S. (2004). Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. Applied Artificial Intelligence (pp. 17-41). Vol. 18. No. 1.

Germann, U. (2003). Greedy Decoding for Statistical Machine Translation in Almost Linear Time. Proceedings of the HLT-NAACL-2003. available at http://www.isi.edu/ licensed-sw/rewrite-decoder/.

Korošec, T. (2002). Opravljeno je bilo pomembno slovarsko delo o vojaškem jeziku. Slovenska vojska (pp. 12-13). Vol. 10, No. 10. (in Slovenian)

Och, F.J. (2002). Statistical Machine Translation: From Single-Word Models to Alignment Templates. Doctoral dissertation, RWTH Aachen, Germany.

Och, F.J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Vol. 29, No. 1, pp. 19–51, available at http://www.fjoch.com/GIZA++.html.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: A Method for Automatic Evaluation of Machine Translation. RC 22176(W0109-022), IBM Research.

Romih, M. & Holozan, P. (2002). Slovensko-angleški prevajalni sistem (A Slovene-English Translation System). In Proceedings of the 3rd Language Technologies Conference 3 (p. 167). Ljubljana, Slovenia. (in Slovenian)

Rosenfeld, R. (1994). The CMU Statistical Language Modeling Toolkit, and Its Use in the 1994 ARPA CSR Evaluation. Proceedings of the ARPA SLT Workshop, available at http://www.speech.cs.cmu.edu/SLM/ toolkit.html.

Sepesy Maučec, M., Brest, J., and Kačič, Z. (2006). Slovenian to English machine translation using corpora of different sizes and morpho-syntactic information. Proceedings of the IS-LTC 2006, Ljubljana, pp. 222–225.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, ELRA, Paris, pp. 2142–2147.

Turian, J.P., Shen, L., and Dan Melamed, I., Proteus Technical Report #03-005: Evaluation of Machine Translation and its Evaluation. available at http://nlp.cs.nyu.edu/eval/.

Vičič, J. (2002). Avtomatsko prevajanje iz slovenskega v angleški jezik na osnovi statističnega strojnega prevajanja. (Automatic SMT: Slovenian-English), masters' thesis, University of Ljubljana, Slovenia.

Žganec Gros, J., Gruden, S., Mihelič, F., Erjavec, T., Vintar, Š., Holozan, P., Mihelič, A., Dobrišek, S., Žibert, J., Logar, N., Korošec., T. (2006). The VoiceTRAN Speech Translation Demonstrator, Proceedings of the IS-LTC 2006, Ljubljana, Slovenia, pp. 234–239.