Improving Machine Translation Between Closely Related Romance Languages

Petr Homola[1] and Vladislav Kuboň[1]

Institute of Formal and Applied Linguistics
Malostranské nám. 25
CZ-118 00 Prague
Czech Republic

Abstract. The paper gives an overview of the shallow-transfer MT system Apertium, describes an experiment with the language pair Portuguese-Spanish and suggests a modification of the system architecture which leads to higher translation quality. Finally, consequences of the architecture improvement for the design of language resources for shallow-transfer based systems are discussed.

1 Introduction

Machine translation (MT) between closely related languages has been paid attention in more than the last twenty years by researchers from many countries. Intuitively, this kind of machine translation should be easier than MT between distant languages since it is possible to exploit the similarity at various linguistic levels. On the other hand, at least one language in the language pair often lacks extensive linguistic resources (such as at least morphologically annotated monolingual corpus or aligned bilingual corpus) that are essential for statistical NLP methods. The most common strategy in MT between related languages, including the open-source system Apertium, is the use of a statistical tagger in order to obtain unambiguous input and a shallow transfer module which adapts source sentences locally so that they conform to grammatical rules of the target language. However, this strategy has severe disadvantages: 1) statistical taggers introduce errors into the morphological annotation of the translated text, and 2) the dictionary is required to contain only one translation for each lemma.

We suggest a modification of the shallow-transfer architecture used in the system Apertium. According to our opinion the tagger should be removed from the system while a statistical ranker should be added after the morphological processing of the target language. Our experiments show that this modification improves the quality of the translation and a pleasant side-effect of the new architecture is that the dictionary can contain more translations for one entry and the transfer rules can be applied non-deterministically, giving possibly more than one result.

This paper is organized as follows: Section 2 contains a brief description of the shallow-transfer MT system Apertium. Section 3 suggests a new system

architecture that leads to higher translation quality. The consequences of the architecture modification for the design of language resources in Apertium are discussed in Section 4 and finally, we conclude in Section 5.

2 The MT System Apertium

There were several experiments in the area of MT between closely related languages, for example [1] for Celtic languages, [2], [3] and [4] for Scandinavian languages, [5] and [6] for Slavic languages or [7] for Turkic languages.

The system Apertium was originally designed for the Romance languages of Spain and it is described in detail in [8]. The authors claim that a word-to-word translation may give an adequate translation of 75% of the text. To improve the translation quality, the architecture given in Figure 1 has been implemented.

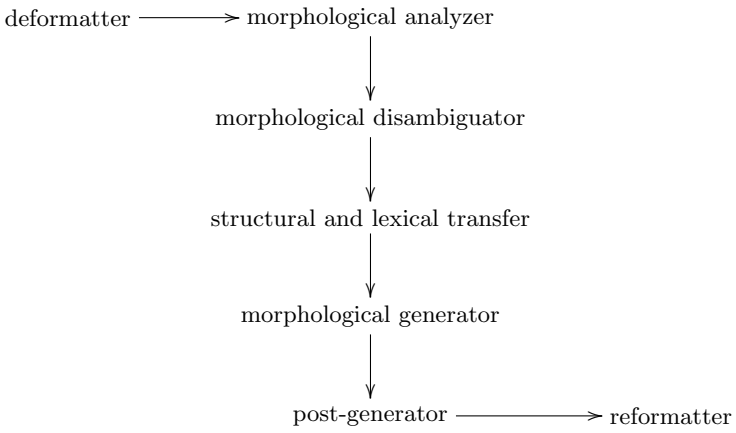

Fig. 1. Architecture of the shallow-transfer MT system Apertium

It is also claimed that this architecture be suitable even for pairs of distant languages, such as Spanish-Basque, which is an intended language pair to be implemented within Apertium.

As the main source of translation errors is morphological ambiguity, a tagger has been prepended before the transfer. The dictionaries contain single equivalents as well as multiword expressions. Transfer rules, which handle, for example, the rearrangement of clitic pronouns, have the form pattern-action and there are approx. 90 of them. The system is capable to process about 5,000 words per second.

Machine translation from Portuguese into Spanish within Apertium is presented in [9]. The system is able to recognize 9,700 Protuguese lemmas and to generate the same amount of Spanish lemmas. The bilingual dictionary contains 9,100 lemma-to-lemma pairs.

3 An Improvement of Apertium's Architecture

According to our experience with an MT system similar to Apertium, the Czech-to-Slovak system Česílko (see [10]), many translation errors, if not most of them, are caused by the comparatively high error rate of the tagger. Since it is rather unrealistic to re-implement the tagger with a significantly better accuracy, the only way to avoid these errors is to omit the whole morphological disambiguation module.

It remains to solve the problem of morphologically ambiguous input. It is well known that a tagger can be partially replaced by a partial (local) parser if the latter is implemented properly. Syntactic analysis is needed in the system anyway to recognize dependencies that are crucial to preserve agreement, for example, between a noun and an adjective which depends on it. Since a simple bottom-up chart parser could do the job effectively, we have made some experiments with an improved implementation of the Q-Systems [11]. We suggest that the remaining ambiguity be resolved by a ranker based on a statistical model for the target language.

The main task of the shallow parser in our system is to deliver an information about the sentence structure to the transfer module so that language specific structural properties could be handled and transferred properly. We have to mention that the parser is not supposed to parse whole sentences. The result is a set of trees which represent only fragments of the sentence. The grammar focuses on the level of noun and prepositional phrases. With this limitation the shallow parser produces highly ambiguous results because it is generally impossible to fully morphologically disambiguate the input sentences on the basis of local context only. The task of selecting the best result is left till the end of the processing chain, to a stochastic ranker of generated target language sentences.

The reason why to rely on statistics is obvious — it would be very complicated (if possible at all) to resolve the rich ambiguity by hand-written rules only. That is why we have implemented a stochastic post-processor which aims to select one particular sentence that is best in the given context.

We use a simple language model based on trigrams (trained on word forms without any morphological annotation) which is intended to sort out "wrong" target sentences (these include grammatically ill-formed sentences as well as inappropriate lexical mapping). The current model has been trained on a corpus of approx. 32 million words which have been randomly chosen from the Spanish Wikipedia.

We have implemented the suggested additional module and evaluated the new architecture on a part of the Europarl corpus [12] using a metrics based on the weighted Levenshtein edit distance.

There are three basic possibilities of the outcome of translation of a segment.

1. The rule-based part of the system has generated a 'perfect'[1] translation (among other hypotheses) and the ranker has chosen this one.

[1] By 'perfect' we mean that the result does not need any human post-processing.

2. The rule-based part of the system has generated a 'perfect' translation but the ranker has chosen a different one.
3. All translations generated by the rule-based part of the system need post-processing.

In the first case, the edit distance is zero, resulting in accuracy equal to 1. In the second case, the accuracy is $1 - d$ with d meaning the edit distance between the segment chosen by the ranker and the correct translation divided by the length of the segment. In the third case, the accuracy is calculated as for (2) except that we use the reference translation to obtain the edit distance.

Given the accuracies for all sentences, we use the arithmetic mean as the translation accuracy of the whole text. The accuracy is negatively influenced by several aspects. If a word is not known to the morphological analyzer, it does not get any morphological information which means that it is practically unusable in the parser. Another possible problem is that a lemma is not found in the dictionary. In such a case, the original source form appears in the translation, which penalizes the score of course. Finally, sometimes the morphological synthesis component is not able to generate the proper word form in the target language (due to partial incompatibility of tagsets for both languages). In such a case, the target lemma appears in the translation.

The results of the evaluation are presented in Table 1.

	original Apertium	new architecture
accuracy (character based)	91.1%	92.4%
accuracy (word based)	87.1%	88.2%

Table 1. Evaluation of the new architecture

For the new architecture to work effectively, the input has to be partially disambiguated. Our experiments show that analyzing noun and prepositional phrases and simple verbal clusters decreases the time consumption of the translation process significantly.

4 Consequences

A big positive side-effect of the new architecture is that the modules are no longer required to generate unambiguous output. Thus a regular non-deterministic parser can be used which produces several syntactic representations of the input, transfer rules can be applied non-deterministically resulting in ambiguous constructions in the target language and, what is particularly important, the dictionary can contain more than one translation for a lemma which is essential for most language pairs.

As for the parser, a bottom-up chart parser has been used for several language pairs, including the Portuguese-Spanish experiment described in this paper. It is

effective enough while offering sufficient formal power. The formalism presented in [11] has turned out to suit the requirements of Apertium very well.

Even more important is the possibility to have more than one translation for a lemma in the dictionary, since even for closely related languages, the translation of a word may depend on the context. The ranker would typically solve at least lexical ambiguities that depend on local context. For example, for the Portuguese word *bola* "ball", there should be at least two Spanish counterparts: *balón* and *pelota*. The first one is the usual translation whereas the other one is used when translating the collocation *bola de tenis*. With an adequate language model for Spanish, the ranker would choose the correct translation even if both translations were in the dictionary independently (i.e., no information about multiword expressions is needed) which further simplifies the design of language resources for Apertium.

5 Conclusions and future work

We present an MT experiment with the language pair Portuguese-Spanish within the open-source shallow-transfer MT system Apertium and suggest a modification of this system's architecture that leads to higher translation quality. Furthermore, we discuss the consequences and side-effects of the suggested architecture modification for the design of language resources (lexicons, dictionaries, rules) and claim that the system could be significantly improved if non-deterministic parser and transfer would be used.

It seems that, after a good set of syntactic rules for the source language has been developed, the way to further improvements of the translation quality could be achieved by using a better model of the target language. In our experiment, we have used a trigram language model trained on word forms. It remains subject to further research whether a language model trained on lemmas, tags or a combination of these would give better results.

The improvement of translation quality has been achieved by changing the architecture of the system — removing the disambiguation of the input text at the very beginning of the translating process and adding a stochastic ranker as the last module of the system. Since this module does not depend on the language pair (requiring only target language-specific data), we assume that the quality increase should be even higher for languages with rich inflection. Our further research will also focus on MT between distant languages, for example the language pair English-Catalan.

Acknowledgments

The presented research has been supported by the grant No. 1ET100300517 of the GAAV ČR.

References

1. Scannell, K.P.: Machine translation for closely related language pairs. In: Proceedings of the Workshop Strategies for developing machine translation for minority languages, Genoa, Italy (2006)
2. Dyvik, H.: Exploiting Structural Similarities in Machine Translation. Computers and Humanities **28** (1995) 225–245
3. Bick, E., Nygaard, L.: Using Danish as a CG Interlingua: A Wide-Coverage Norwegian-English Machine Translation System. In: Proceedings of NODALIDA, Tartu, Estonia (2007)
4. Ahrenberg, L., Holmqvist, M.: Back to the Future? The Case for English-Swedish Direct Machine Translation. In: Proceedings of Recent Advances in Scandinavian Machine Translation, Uppsala, Sweden (2005)
5. Hajič, J.: An MT System Between Closely Related Languages. In: Proceedings of the third conference of the European Chapter of the Association for Computational Linguistics, Copenhagen, Denmark (1987) 113–117
6. Hajič, J., Hric, J., Kuboň, V.: Machine translation of very close languages. In: Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, Washington, USA, April 2000, pp. 7–12. (2000)
7. Altintas, K., Cicekli, I.: A Machine Translation System between a Pair of Closely Related Languages. In: Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002), Orlando, Florida (2002) 192–196
8. Corbi-Bellot, A., Forcada, M., Prtiz-Rojas, S., Perez/Ortiz, J.A., Remirez-Sanchez, G., Martinez, F.S., Alegria, I., Mayor, A., Sarasola, K.: An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain. In: Proceedings of the 10th Conference of the European Association for Machine Translation, Budapest (2005)
9. Armentano-Oller, C., Carrasco, R.C., Corbí-Bellot, A.M., Forcada, M.L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M.A.: Open-source Portuguese-Spanish machine translation. In: Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, Rio de Janeiro, Brasil (2006)
10. Hajič, J., Homola, P., Kuboň, V.: A simple multilingual machine translation system. In: Proceedings of the MT Summit IX, New Orleans (2003)
11. Colmerauer, A.: Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. Technical report, Mimeo, Montréal (1969)
12. Koehn, P.: Europarl: A Multilingual Corpus for Evaluation of Machine Translation (2002) http://people.csail.mit.edu/ koehn/publications/europarl.ps.