First Steps towards Multi-Engine Machine Translation

Andreas Eisele
Computational Linguistics
Saarland University P.O.Box 151150
D-66041 Saarbrücken, Germany
eisele@coli.uni-saarland.de

Abstract

We motivate our contribution to the shared MT task as a first step towards an integrated architecture that combines advantages of statistical and knowledge-based approaches. Translations were generated using the Pharaoh decoder with tables derived from the provided alignments for all four languages, and for three of them using web-based and locally installed commercial systems. We then applied statistical and heuristic algorithms to select the most promising translation out of each set of candidates obtained from a source sentence. Results and possible refinements are discussed.

1 Motivation and Long-term Perspective

"The problem of robust, efficient and reliable speech-to-speech translation can only be cracked by the combined muscle of deep and shallow processing approaches." (Wahlster, 2001) Although this statement has been coined in the context of VerbMobil, aiming at translation for direct communication, it appears also realistic for many other translation scenarios, where demands on robustness, coverage, or adaptability on the input side and quality on the output side go beyond today's technological possibilities. The increasing availability of MT engines and the need for better quality has motivated considerable efforts to combine multiple engines into one "super-engine" that is hopefully better than any of its ingredients, an idea pionieered in (Frederking and Nirenburg, 1994). So far, the larger group of related publications has focused on the task of selecting, from a set of translation candidates obtained from different engines, one translation that looks most promising (Tidhar and Küssner, 2000; Akiba et al., 2001; Callison-Burch and Flournoy, 2001; Akiba et al., 2002; Nomoto, 2004). But also the more challenging problem of decomposing the candidates and re-assembling from the pieces a new sentence, hopefully better than any of the given inputs, has recently gained considerable attention (Rayner and Carter, 1997; Hogan and Frederking, 1998; Bangalore et al., 2001; Jayaraman and Lavie, 2005).

Although statistical MT approaches currently come out as winners in most comparative evaluations, it is clear that the achievable quality of methods relying purely on lookup of fixed phrases will be limited by the simple fact that for any given combination of topic, application scenario, language pair, and text style there will never be sufficient amounts of pre-existing translations to satisfy the needs of purely data-driven approaches.

Rule-based approaches can exploit the effort that goes into single entries in their knowledge repositories in a broader way, as these entries can be unfolded, via rule applications, into large numbers of possible usages. However, this increased generality comes at significant costs for the acquisition of the required knowledge, which needs to be encoded by specialists in formalisms requiring extensive training to be used. In order to push the limits of today's MT technology, integrative approaches will have to be developed that combine the relative advantages of

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 155–158,
Ann Arbor, June 2005.

both paradigms and use them to compensate for their disadvantages. In particular, it should be possible to turn single instances of words and constructions found in training data into internal representations that allow them to be used in more general ways.

In a first step towards the development of integrated solutions, we need to investigate the relative strengths and weaknesses of competing systems on the level of the target text, i.e. find out which sentences and which constructions are rendered well by which type of engine. In a second step, such an analysis will then make it possible to take the outcomes of various engines apart and re-assemble from the building blocks new translations that avoid errors made by the individual engines, i.e. to find integrated solutions that improve over the best of the candidates they have been built from. Once this can be done, the third and final step will involve feed back of corrections into the individual systems, such that differences between system behaviour can trigger (potentially after manual resolution of unclear cases) system updates and mutual learning.

In the long term, one would hope to achieve a setup where a group of MT engines can converge to a committee that typically disagrees only in truly difficult cases. In such a committee, remaining dissent between the members would be a symptom of unresolved ambiguity, that would warrant the cost of manual intervention by the fact that the system as a whole can actually learn from the additional evidence. We expect this setup to be particularly effective when existing MT engines have to be ported to new application domains. Here, a rule-based engine would be able to profit from its more generic knowledge during the early stages of the transition and could teach unseen correspondences of known words and phrases to the SMT engine, whereas the SMT system would bring in its abilities to apply known phrase pairs in novel contexts and quickly learn new vocabulary from examples.

2 Collecting Translation Candidates

2.1 Setting up Statistical MT

In the general picture laid out in the preceding section, statistical MT plays an important role for several reasons. On one hand, the construction of a relatively well-performing phrase-based SMT system from a given set of parallel corpora is no more overly difficult, especially if — as in the case in this shared task — word alignments and a decoder are provided. Furthermore, once the second task in our chain will have been surmounted, it will be relatively easy to feed back building blocks of improved translations into the phrase table, which constitutes the central resource of the SMT system Therefore, SMT facilitates experiments aiming at dynamic and interactive adaptation, the results of which should then also be applicable to MT engines that represent knowledge in a more condensed form.

In order to collect material for testing these ideas, we constructed phrase tables for all four languages, following roughly the procedure given in (Koehn, 2004) but deviating in one detail related to the treatment of unaligned words at the beginning or end of the phrases[1]. We used the Pharaoh decoder as described on http://www.statmt.org/wpt05/mt-shared-task/ after normalization of all tables to lower case.

2.2 Using Commercial Engines

As our main interest is in the integration of statistical and rule-based MT, we tried to collect results from "conventional" MT systems that had more or less uniform characteristics across the languages involved. We could not find MT engines supporting all four source languages, and therefore decided to drop Finnish for this part of the experiment. We sent the texts of the other three languages through several incarnations of Systran-based MT Web-services[2] and through an installation of Lernout & Hauspie Power Translator Pro, Version 6.43.[3]

[1] We used slightly more restrictive conditions that resulted in a 5.76% reduction of phrase table size

[2] The results were incomplete and different, but sufficiently close to each other so that it did not seem worthwhile to explore the differences systematically. Instead we ranked the services according to errors in an informal comparison and took for each sentence the first available translation in this order.

[3] After having collected or computed all translations, we observed that in the case of French, both systems were quite sensitive to the fact that the apostrophes were formatted as separate tokens in the source texts (l ' homme instead of l'homme). We therefore modified and retranslated the French texts, but did not explore possible effects of similar transformations in the other languages.

3 Heuristic Selection

3.1 Approach

We implemented two different ways to select, out of a set of alternative translations of a given sentence, one that looks most promising. The first approach is purely heuristic and is limited to the case where more than two candidates are given. For each candidate, we collect a set of features, consisting of words and word n-grams ($n \in \{2, 3, 4\}$). Each of these features is weighted by the number of candidates it appears in, and the candidate with the largest feature weight per word is taken. This can be seen as the similarity of each of the candidate to a prototypical version composed as a weighted mixture of the collection, or as being remotely related to a sentence-specific language model derived from the candidates. The heuristic measure was used to select "favorite" from each group of competing translations obtained from the same source sentence, yielding a fourth set of translations for the sentences given in DE, FR, and ES.

A particularity of the shared task is the fact that the source sentences of the development and test sets form a parallel corpus. Therefore, we can not only integrate multiple translations of the same source sentence into a hopefully better version, but we can merge the translations of corresponding parts from different source languages into a target form that combines their advantages. This approach, called triangulation in (Kay, 1997), can be motivated by the fact that most cases of translation for dissemination involve multiple target languages; hence one can assume that, except for the very first of them, renderings in multiple languages exist and can be used as input to the next step[4]. See also (Och and Ney, 2001) for some related empirical evidence. In order to obtain a first impression of the potential of triangulation in the domain of parliament debates, we applied the selection heuristics to a set of four translations, one from Finnish, the other three the result of the selections mentioned above.

3.2 Results and Discussion

The BLEU scores (Papineni et al., 2002) for 10 direct translations and 4 sets of heuristic selections thereof are given in Table 1. These results show that in each group of translations for a given source language, the statistical engine came out best. Furthermore, our heuristic approach for the selection of the best among a small set of candidate translations did not result in an increase of the measured BLEU score, but typically gave a score that was only slightly better than the second best of the ingredients. This somewhat disappointing result can be explained in two ways. Apparently, the selection heuristic does not give effective estimates of translation quality for the candidates. Furthermore, the granularity on which the choices have to bee made is too coarse, i.e. the pieces for which the symbolic engines do produce better translations than the SMT engine are accompanied by too many bad choices so that the net effect is negative.

Source Language	MT Engine	BLEU score
DE	Pharaoh	20.48
	L & H	13.97
	Systran	14.92
	heuristic selection	16.01
	statistical selection	20.55
FR	Pharaoh	26.29
	L & H	17.82
	Systran	20.29
	heuristic selection	21.44
	statistical selection	26.49
ES	Pharaoh	26.69
	L & H	17.28
	Systran	17.38
	heuristic selection	19.16
	statistical selection	26.74
FI	Pharaoh	16.76
all	heuristic selection	22.83
	statistical selection	25.80

Table 1: BLEU scores of various MT engines and combinations

4 Statistical Selection

The other score we used was based on probabilities as computed by the trigram language model for English provided by the organizers of the task, in a representation compatible with the SRI LM toolkit

[4] Admittedly, in typical instances of such chains, English would appear earlier.

(Stolcke, 2002). However, a correct implementation for obtaining these estimates was not available in time, so the selections generated from the statistical language model could not be used for official submissions, but were generated and evaluated after the closing date. The results, also displayed in Table 1, show that this approach can lead to slight improvements of the BLEU score, which however turn out not to be statistically sigificant in then sense of (Zhang et al., 2004).

5 Next Steps

When we started the experiments reported here, the hope was to find relatively simple methods to select the best among a small set of candidate translations and to achieve significant improvements of a hybrid architecture over a purely statistical approach. Although we could indeed measure certain improvements, these are not yet big enough for a conclusive "proof of concept". We have started a refinement of our approach that can not only pick the best among translations of complete sentences, but also judge the quality of the building blocks from which the translations are composed. First informal results look very promising. Once we can replace single phrases that appear in one translation by better alternatives taken from a competing candidate, chances are good that a significant increase of the overall translation quality can be achieved.

6 Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft. We want to thank two anonymous reviewers for numerous pointers to relevant literature, Bogdan Sacaleanu for his help with the collection of translations from on-line MT engines, as well as the organizers of the shared task for making these interesting experiments possible.

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.

Yasuhiro Akiba, Taro Watanabe, and Eiichiro Sumita. 2002. Using language and translation models to select the best among outputs from multiple mt systems. In *COLING*.

Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *ASRU*, Italy.

Chris Callison-Burch and Raymond S. Flournoy. 2001. A program for automatically selecting the best output from multiple machine translation engines. In *Proc. of MT Summit VIII*, Santiago de Compostela, Spain.

Robert E. Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *ANLP*, pages 95–100.

Christopher Hogan and Robert E. Frederking. 1998. An evaluation of the multi-engine mt architecture. In *Proceedings of AMTA*, pages 113–123.

Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proc. of EAMT*, Budapest, Hungary.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12:3–23. First appeared as a Xerox PARC working paper in 1980.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *AMTA*, pages 115–124.

Tadashi Nomoto. 2004. Multi-engine machine translation with voted language model. In *Proc. of ACL*.

Franz-Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain, September.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Manny Rayner and David M. Carter. 1997. Hybrid language processing in the spoken language translator. In *Proc. ICASSP '97*, pages 107–110, Munich, Germany.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*.

Dan Tidhar and Uwe Küssner. 2000. Learning to select a good translation. In *COLING*, pages 843–849.

Wolfgang Wahlster. 2001. Robust translation of spontaneous speech: A multi-engine approach. In *IJCAI*, pages 1484–1493. Invited Talk.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of LREC*, Lisbon, Portugal.