Parallel Corpora and Phrase-Based Statistical Machine Translation for New Language Pairs via Multiple Intermediaries

Andreas Eisele

Computational Linguistics
Saarland University
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
eisele@coli.uni-saarland.de

Abstract

We present a large parallel corpus of texts published by the United Nations Organization, which we exploit for the creation of phrase-based statistical machine translation (SMT) systems for new language pairs. We present a setup where phrase tables for these language pairs are used for translation between languages for which parallel corpora of sufficient size are so far not available. We give some preliminary results for this novel application of SMT and discuss further refinements.

1. Motivation

As recent DARPA MT evaluations (Lee & Przybocki, 2005) have shown, MT systems with interesting performance levels can be built from large repositories of parallel texts available from international organizations, using statistical techniques to turn these corpora into collections of aligned substrings. However, these evaluations focused on the translation of Chinese and Arabic to English and exploited only a small part of the rapidly growing set of available parallel texts (Erjavec e.a. 2005). Issues related to morphologically richer target languages have not been explored in this setting. In contrast to this, Köhn (2002) did not only make a sizeable collection of records of the European Parliament in 11 languages freely available to the research community; the same author has also shown in (Köhn 2005) how to build MT systems for all 110 possible language pairs (LPs) from these resources, providing interesting insights into the dependence of BLEU-score estimates of translation quality on morphological properties of the involved languages, mainly the target language.

The work described in the present paper builds upon this groundwork and aims at the investigation of three interrelated questions:

- How to turn large text collections in six languages available from UN publications into parallel corpora and SMT systems that potentially cover 30 instead of only 2 LPs
- How to extend parallel text collections that are available for new European languages (Erjavec et al. 2005) into SMT systems that cover much larger parts of the language matrix than has so far been possible
- How to bridge the gap between "big" languages like Chinese, Arabic, and Russian on one side and a larger number of "smaller" European languages on the other, using a combination of English, French, and Spanish as an intermediary.

The main motivation for this work is the fact that, whereas machine translation was among the first proposed uses of digital computers in the middle of the 20th century, and since then has been the target of intensive world-wide research and development efforts, the translation directions for which MT systems have been developed have been limited to a rather small number of language pairs for which at least one of the languages has a wide distribution. An enumeration of commercial MT systems given in (Hutchins et al. 2005) shows a core set of 10 languages (English, German, French, Japanese, Spanish, Italian, Russian, Portuguese, Polish, and Ukrainian) that is "fully connected" by these systems. These 90 language pairs make up almost 50% of the 183 pairs covered by all enumerated systems, and none of the systems translate between languages outside this set, with the exception of Chinese ↔ Korean. Fig. 1 visualizes the number of MT systems that cover the various language pairs according to the data in the compendium[1]. These statistics emphasize the (perhaps obvious) fact that almost all languages for which MT systems exist at all have a possibility to automatically translate into and from English. It may therefore look simple to enhance the reach of MT to new language pairs by going via English as an intermediary. However, given the limited quality level of today's MT systems, adding an intermediate step will significantly reduce the overall quality, so that this approach has so far not been practically relevant.

In this paper, we propose a generalized approach that allows improving the resulting quality by going via multiple languages instead of only one. Especially, we show how this setup can help to translate between widespread languages like Chinese, Russian, and Arabic on one hand, and many of the European languages for which large parallel corpora with these source languages do not exist.

2. The United Nations Document Collection

Publications by the UNO have been the basis for the creation of large parallel corpora at least twice in the past. In 1994 LDC published a set of parallel corpora created from UNO publications (*LDC Catalog No.:* LDC94T4A),

[1] The source language is given on the left and the target language on the top. Only languages with more than one "partner language" are included in our selection, whereas the total numbers of MT systems/language pairs specified in the last two rows and columns refer to all the systems given in (Hutchins et al. 2005)

which contained parallel documents in English, French, and Spanish. In the framework of the DARPA/NIST MT evaluation campaign, new parallel corpora were collected, which were then distributed by LDC to the registered participants of the campaign. However, the distribution was limited to Arabic-English and Chinese-English parts of the publications, which restricts applicability for work aiming at European languages other than English. Given that UNO documents are typically published in 6 languages, restrictions to subsets of these languages seem somewhat unmotivated, especially as the effect on the number of covered language pairs is more than linear. In order to exploit this document repository in full generality, we collected a multilingual set of documents from http://documents.un.org/, covering all six official UN-languages (AR, EN, ES, FR, RU, ZH) and also containing a smaller fraction of German translations. So far, we have gathered 37013 documents that exist in all 6 official languages, 2107 of which also have a German translation. The German documents that can be aligned with all 6 other languages contain about 2.7 million tokens in 88000 alignment units. From these documents, parallel text was extracted and sentence alignments across all languages were computed, resulting in a collection of sentences or sentence fragments for which at least one aligned version is available. We plan to make these data sets available in a manner similar to the way the Europarl corpus has been distributed. Should this form of distribution conflict with legal restrictions of any kind, it is also possible to apply the approach chosen by the "hunglish" project (Varga et al., 2005), who were able to get rid of legal restrictions for copyrighted material by shuffling the sentence pairs of their parallel corpus The UNO collection is currently being reworked and extended, and details of the distribution scheme will be announced at a later time. From these resources, nine SMT systems translating from (AR, RU, ZH) to (EN, ES, FR) are currently being constructed, following essentially the approach proposed by (Köhn 2005), including the tools made available by him.

3. An architecture for indirect SMT

A first motivation for the collection of the corpus in 6 UNO languages was the possibility to use it directly to train SMT systems for 30 language pairs. For more than a third of the language pairs covered by this resource, no commercial MT engine has so far been reported in (Hutchins 2005). However, the fact that four of these languages belong to the "inner circle" of MT languages, for which many MT systems exist, and the fact that three of them are contained in the Europarl corpus open up even more possibilities. In particular, it is possible to build up a setup as depicted in Fig. 2, where more than one intermediate language is used to go from Russian, Chinese, or Arabic to any of the Europarl languages or any other official EU language, as soon as large parallel corpora spanning all these languages become available. There are several independent reasons why adding alternative intermediate languages can be advantageous. This may on one hand reduce the risk to obtain bad translations due to missing coverage in any of the resources that are being used, as in such cases the system may fall back to an alternative path. On the other hand, if details are lost because one of the languages is not able to express a given notion in the appropriate specificity[2], constraints from the other languages can help to compensate for this. Similar ideas not only hold for lexical selection, but also for the choice of appropriate syntactic constructions in the intermediate languages. This can be seen as a generalization of the idea of using constraints from renderings of a given text in multiple languages, which has been around since a long time (Kay 1997) and for which practical progress has been reported in (Och & Ney 2001). Obviously, these two potential advantages of adding possibilities and adding constraints stand in a certain tension; hence the combination of both aspects will most likely require such a stochastic framework in order to facilitate the integration of weak evidence from multiple sources.

4. State of the implementation

In order to obtain a first impression of the potential of the setup sketched in the last section, we did some preliminary experiments, but unfortunately, given the large number of possible variations and refinements and some lack of time we could so far only scratch the surface of what can or needs to be done. From the set of documents that exist in all 6 languages, we selected about 400000 sentence pairs for three language pairs (RU-EN, RU-FR, RU-ES) for which both lengths remained below 100 tokens and do not differ by a factor of more than 2. From this subset, we generated phrase tables with the help of GIZA++ and the training scripts by Philipp Köhn. Using these phrase tables, the SRILM language models for English, Spanish, and French available from http://www.statmt.org/wmt06/shared-task/, and the Pharaoh decoder from ISI (http://www.isi.edu/licensed-sw/pharaoh/), it is already possible to obtain rough translations from unseen Russian text to these three languages. It would be possible to use the vast amount of existing translations to fine-tune the parameters of the models, but this process will take some additional time, and so far, we only informally inspected the results for mismatches in vocabulary, in capitalization, in tokenization conventions etc. We also computed sentence and word alignments of the Europarl corpus for the pairs consisting of our intermediate languages and a possible target language (in this case German, for which a language model is also distributed for the WMT06 shared task). Whereas it would be easily possible to connect the six parts into the setup sketched in Fig. 2, this has so far not yet been tried out due to lack of time.

5. Next Steps

Given the so far preliminary state of the system, many of the next steps seem obvious. Not only do we need to connect the parts that already exist into an end-to-end system that translates from Russian, Arabic, or Chinese into any of the Europarl languages, but we also need to optimize various implementation details of the system (such as tokenization conventions and the treatment of capitalization) and make sure the conventions agree between the parts of the system that are to be connected. Once such an end-to-end baseline system is in place, the next step will then be the calibration of the relative

[2] Translators often have to make delicate choices between expressions that are either too broad or too specific.

weights of the contributions from the three intermediate languages. In the simplest case, we can select the most probable translation candidate according to the statistical model of the target language. However, since the sentences in the intermediate languages are implicitly aligned with the source sentence, it is particularly easy to identify corresponding parts of the intermediate sentences and to use this alignment for finding good combinations of candidate substrings in the target language generated via different intermediaries.

6. Acknowledgements

The work described in this paper was made possible by the DFG in the framework of the Ptolemaios project at Saarland University, headed by Jonas Kuhn. I also want to thank Sascha Osherenko and Greg Gulrajani for interesting discussions and for practical help with crawling the UNO web site.

7. References

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen & Ralf Steinberger (2005). Massive multilingual corpus compilation; Acquis Communautaire and totale. In: 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (L&T'05). Poznań, Poland, 21-23 April 2005. Available from www.jrc.cec.eu.int/langtech/

J. Hutchins, W. Hartmann,and E. Ito (2005): Compendium of translation software: commercial machine translation systems and computer-based translation support tools. 11th edition, July 2005. Available from www.eamt.org.

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 Köhn (2002), Europarl: A Multilingual Corpus for Evaluation of Machine Translation, unpublished draft, December 2002, available from www.iccs.informatics.ed.ac.uk/~pkoehn/publications/.

Philipp Köhn, Franz Josef Och, and Daniel Marcu (2003), Statistical Phrase-Based Translation, in proceedings of HLT/NAACL 2003.

Philipp Köhn (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit X, Phuket.

Audrey Lee and Mark Przybocki (2005). NIST 2005 machine translation evaluation official results. Official release of automatic evaluation scores for all submissions, August.

Franz Josef Och and Hermann Ney (2001): Statistical Multi-Source Translation. MT Summit VIII

D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh, and V. Trón (2005). Parallel corpora for medium density languages. Proceedings of RANLP 2005

	English	German	French	Japanese	Spanish	Italian	Russian	Portuguese	Ukrainian	Polish	Korean	Czech	Chinese	Dutch	Swedish	Hungarian	Greek	Croatian	Catalan	Arabic		
English	*	47	42	44	45	31	19	31	6	9	16	1	23	8	1	4	3	1	2	15	**386**	**36**
German	48	*	25	3	8	9	13	3	2	4	1	1	-	1	1	2	-	1	-	-	**127**	**20**
French	41	24	*	3	10	11	5	6	1	2	1	1	1	3	-	-	3	-	-	2	**114**	**15**
Japanese	42	3	-	*	3	3	1	3	1	1	16	-	12	-	-	-	-	-	-	-	**88**	**11**
Spanish	42	7	10	3	*	8	4	7	1	1	1	1	-	-	-	-	-	-	3	-	**88**	**12**
Italian	29	9	11	3	8	*	4	3	1	1	1	1	-	-	-	-	-	-	-	-	**71**	**11**
Russian	23	13	5	1	4	2	*	1	2	1	-	1	-	-	-	-	-	-	-	-	**53**	**10**
Portuguese	29	4	5	4	7	3	1	*	1	1	2	-	1	-	-	-	-	-	-	-	**58**	**11**
Ukrainian	6	2	1	1	1	1	2	1	*	1	-	-	-	-	-	-	-	-	-	-	**16**	**9**
Polish	8	3	2	1	1	1	1	2	1	*	-	-	-	-	-	-	-	-	-	-	**20**	**9**
Korean	15	-	-	17	-	-	-	-	-	-	*	-	1	-	-	-	-	-	-	-	**33**	**3**
Czech	1	1	1	-	-	1	1	-	-	-	-	*	-	-	-	-	-	-	-	-	**5**	**5**
Chinese	21	-	-	12	-	-	-	-	-	-	1	-	*	-	-	-	-	-	-	-	**34**	**3**
Dutch	8	1	3	-	-	-	-	-	-	-	-	-	-	*	-	-	-	-	-	-	**12**	**3**
Swedish	2	1	-	-	-	-	-	-	-	-	-	-	-	-	*	-	-	-	-	-	**3**	**2**
Hungarian	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	*	-	-	-	-	**4**	**2**
Greek	2	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	*	-	-	-	**5**	**2**
Croatian	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*	-	-	**2**	**2**
Catalan	2	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	*	-	**5**	**2**
Arabic	14	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*	**16**	**2**
Latin	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**2**	**2**
Finnish	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**3**	**2**
Esperanto	1	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**2**	**2**
	363	**122**	**113**	**93**	**90**	**70**	**51**	**57**	**16**	**21**	**39**	**6**	**39**	**12**	**2**	**6**	**6**	**2**	**5**	**17**		
	33	**18**	**13**	**12**	**10**	**10**	**10**	**9**	**9**	**9**	**8**	**6**	**6**	**3**	**2**	**2**	**2**	**2**	**2**	**2**		

Figure1: Number of commercial MT systems per language pair, according to (Hutchins 2005)

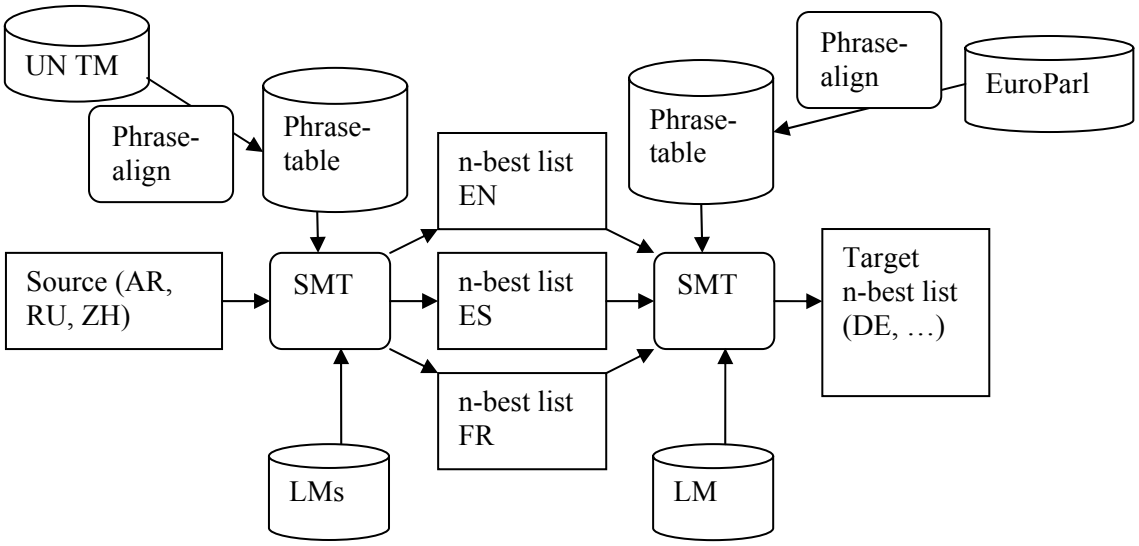

Figure2: Architecture of an integrated SMT system for language pairs for which large parallel corpora are not available