Linguistic Motivation in Automatic Sentence Alignment of Parallel Corpora: the Case of Danish-Bulgarian and English-Bulgarian

Angel Genov and Georgi Iliev
Department of Computational Linguistics, Institute for Bulgarian
Bulgarian Academy of Sciences
52 Shipchenski prohod, bl. 17, Sofia 1113, Bulgaria
{angel,georgi}@dcl.bas.bg

Abstract

We report the results from a sentence-alignment experiment on Danish-Bulgarian and English-Bulgarian parallel texts applying a method based in part on linguistic motivations as implemented in the TCA2 aligner. Since the presence of cognates has a bearing on the alignment score of candidate sentences we attempt to bridge the gap between source and target languages by transliteration of the Bulgarian text, written originally in Cyrillic. An improvement in F_1-measure is achieved in both cases.

1 Background

Parallel language resources are fundamental to some of the leading empirical methods in natural language processing today, and machine translation in particular. Due to economic and political considerations until now little attention has been paid to the availability and quality of parallel texts when it comes to so-called medium density languages, as defined in (Varga et al., 2005). A major development in this regard has been the release of the JRC-Acquis Multilingual Parallel Corpus (Steinberger et al., 2006) but a brief investigation of the aligned versions of the JRC-Acquis for the language pairs at hand fails to convince us of the quality of alignment in the corpus in its current state. However, due to the lack of a reference parallel text to perform evaluation, this intuition can neither be confirmed nor rejected at present.

In an attempt to achieve better clarity and provide a basis for further evaluation and comparison, we performed a sentence-alignment experiment on a Danish-Bulgarian and an English-Bulgarian parallel corpus, where translation has taken place in the indicated direction.

2 Alignment Method

Unlike the alignment methods adopted for the purposes of the JRC-Acquis, where a language-independent approach was needed to achieve coverage of most official EU languages, we chose to apply a partially linguistically motivated method to sentence alignment, as implemented in the Translation Corpus Aligner (TCA) 2 (Hofland and Johansson, 2006). The TCA2 is a GUI sentence alignment tool which calculates alignments on the basis of sentence length, a bilingual dictionary of anchor words, and the presence of (near) identical proper names and numbers and cognates in alignment candidates. It is a new implementation of a program which was used, among other things, for the alignment of the English-Norwegian Parallel Corpus (Johansson et al., 1996).

3 Corpora

For English-Bulgarian alignment we used the "1984" parallel corpus developed as part of the MULTEXT-East project (Erjavec, 2010), and provided by the MULTEXT-East Consortium under a research license. It contains a richly annotated hand-aligned parallel text. For our purposes we stripped the English-Bulgarian parallel text of George Orwell's "1984" of all annotation which was not relevant to sentence alignment. Thus the version of the "1984" corpus we used contains only XML tags marking sentence boundaries. The English text contains 6737 sentence units, and the Bulgarian text contains 6707 sentence units.

For Danish-Bulgarian alignment we used the original text of Thomas Rathsack's "Jæger – i krig med eliten" (*Commando – Fighting With the Elite*) published by Politiken's Internet edition on 16 September 2009 (Rathsack, 2009). The Bulgarian translation was provided for research purposes by the respective Bulgarian publisher. The sentence boundaries in the two texts were initially

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 328–331

determined automatically. The resulting sentence boundaries were post edited and sentences were aligned by hand. The parallel corpus thus created contains XML tags marking sentence boundaries only. The Danish text contains 4483 sentence units, and the Bulgarian text contains 4565 sentence units.

We are not aware of previous work in the evaluation of sentence alignment as regards the Danish-Bulgarian language pair.

4 Language-Dependent Input

The TCA2 uses a bilingual dictionary of anchor words whose presence improves the alignment score of candidate sentences. TCA results reported in previous work (Santos and Oksefjell, 2000) have been based on anchor lists of approximately 1000 entries. In the experiment at hand we followed a resource-light strategy, which means we tried to keep the manual input at minimum, while preserving language-dependency.

The heuristics applied in compiling the bilingual dictionaries involved counting the number of occurrences of individual lemmas in the respective Bulgarian texts, disregarding any stop words, and selecting some of the most frequent nominals (that is nouns, adjectives, numerals and pronouns). To them we added some "polar" adverbs (such as "always" and "never"), some time words and the names of the twelve months of the year. The respective anchor lists contain 116 entries each.

One special feature of the TCA2 is the use of multiple variants in one and the same anchor entry, as well as the Kleene star, allowing us to cover a number of morphological and orthographic variations, as otherwise the number of anchor entries would have exploded, in particular in the Danish and the Bulgarian anchor lists.

DA-BG	2*,to*,begge,både/2*,два*,две*,дву*
EN-BG	woman,women/жена*,жени*

Table 1: Sample anchor word entries

It is possible that defining dictionary entries by means of the Kleene star could increase the number of false positives disproportionately, but that is not confirmed by the reported results.

5 Transliteration

An important element of the TCA2 tool is the assignment of a score to alignment candidates based on the presence of cognates (Simard et al., 1992) in them – words that are spelled identically or similarly in the source and target language. Both source languages use the Latin alphabet, whereas Bulgarian is written in Cyrillic. That fact effectively prevents any attempt at basing an alignment score on cognates, which we found to be suboptimal in the case of the TCA2 aligner.

A clear solution lies in the fact that Bulgarian spelling is mostly phonetic. Thus we were able to apply a straightforward transliteration method to convert the entire target text into Latin characters to explore the ability to use cognates in order to improve sentence alignment of the two parallel corpora. The transliteration method so adopted preserves the original length of the Bulgarian text as far as possible and does not take into account the specifics of the source language.

English original:

How often, or on what system, the Thought Police plugged in on any individual wire was guesswork.

Bulgarian original:

Можеше само да се правят предположения колко често и по какъв принцип Полицията на мисълта се включва в индивидуалните системи.

Bulgarian transliterated:

Mojeshe samo da se praviat predpolojeniia kolko chesto i po kakav princip Policiiata na misalta se vkliuchva v individualnite sistemi.

Example 1: Cognates identified upon transliteration of the target text

It can be seen from Example 1 above that cognates become evident by applying even a crude transliteration approach.

6 Evaluation Method

In order to evaluate the performance of the TCA2 we apply the standard metrics of recall, precision and F_1-measure (Jurafsky and Martin, 2008) to the

output of the automatic aligner against the hand-aligned "gold standards", calculated as shown in Example 2 below:

> Precision = number of correct alignments / number of proposed alignments
>
> Recall = number of correct alignments / number of reference alignments
>
> F_1-measure = 2 * Recall * Precision / (Recall + Precision)

Example 2: Calculation of precision, recall and F_1-measure

It should be noted that the manually aligned corpora do not contain any 0-1 or 1-0 alignments, whereas TCA2 allows 0-1 and 1-0 alignments and such alignments are present in the TCA2 automatic output. Other possible alignments are 1-1, 1-2, 2-1.

7 Results

We performed a total of 3 runs of the TCA2 aligner on each language pair, gradually enhancing the linguistic input in the system, as follows:

1st run: sentence-delimited source and target texts, no anchor words, no transliteration of the target text.

2ndrun: sentence-delimited source and target texts, anchor words, no transliteration of the target text.

3rd run: sentence-delimited source and target texts, anchor words, transliteration of the target text.

The results are summarized in Table 2 below:

DA-BG, "Jæger – i krig med eliten"			
	Precision	Recall	F_1-measure
1st run	93.96	93.83	**93.9**
2nd run	97.23	97.12	**97.17**
3rd run	98.39	98.12	**98.26**
EN-BG, "1984"			
	Precision	Recall	F_1-measure
1st run	61.52	63.41	**62.45**
2nd run	87.77	89.17	**88.46**
3rd run	91.52	92.33	**91.92**

Table 2: Evaluation results

8 Discussion of Results

As expected, there was an improvement in all three metrics on each of the first three stages. It must be noted that the weights of different alignment criteria have not been optimized for the specific language pairs at hand, which means that further improvement may be expected. Most notable at present is the 26% improvement in F_1-measure of the English-Bulgarian alignment caused by the enhancement of the TCA2 system with an anchor list, irrespective of the limited size of the list (116 entries in total). It could be argued that the Danish-Bulgarian improvement on the 2nd run is as significant because the higher F_1-measure on the 1st run could be accounted for by the presence of many military terms and English words and expressions which were preserved in the Bulgarian translation of "Jæger – i krig med eliten", which in turn were treated as cognates / proper names by the TCA2 aligner already on the first run.

Most interesting is probably the improvement caused by the transliteration of the Bulgarian text. One disadvantage of phonetic spelling in Bulgarian is that it fails to preserve the original spelling of loan words, thus reducing the number of actual cognates in the text. We expect that by fine-tuning the transliteration rules dependent on the language pair in order to "anglicize" or "danify" the target language, respectively, following the philosophy of (Hana and Feldman, 2004), as applied to Russian and Czech, we could achieve further improvement in sentence alignment based on cognates. However, this will be the subject of a separate detailed study.

9 Conclusion and Future Work

We have shown that by investing a minimum amount of effort we can achieve significant improvement in the partially linguistically motivated approach to sentence alignment of the TCA2 aligner, bringing its performance, as expressed by the F_1-measure, well above 90 per cent in the case of Danish-Bulgarian, and above 90 per cent in the case of English-Bulgarian alignment. By applying a simple transliteration approach to the Bulgarian target we were able to achieve additional improvement of TCA2's performance.

Unlike the evaluation described in (Santos and Oksefjell, 2000), our main purpose was to explore the feasibility of applying a language-dependent method in sentence alignment to parallel texts

where Bulgarian is the target language, not so much as to evaluate the performance of the TCA2 program itself. One difficulty which is not an issue in other alignment tasks, and which we addressed successfully, was the need to bridge the gap between the Cyrillic and the Latin alphabets. Furthermore, while the English-Portuguese evaluation was based on proofreading the results (and only those that were not 1-to-1 correspondencies), the Danish-Bulgarian and the English-Bulgarian alignments were evaluated on the basis of a gold standard which allowed us to calculate the improvement in a traditional manner, by way of the F_1-measure.

Driven by those preliminary results we intend to investigate the systematic differences occurring in the spelling of some loan words in Bulgarian from a morphological perspective. The results of the investigation could then be encoded in the form of language-specific transliteration rules to achieve further improvement in the performance of linguistically motivated sentence-alignment methods.

Acknowledgements

We are grateful to Knut Hofland and the team at the University of Bergen for making the TCA2 program available to us.

This work was supported by the Mathematical Logic and Computational Linguistics: Development and Permeation (2009-2011) Project. The financial support is granted under Contract No. BG051PO001-3.3.04/27 of 28 August 2009.

References

Tomaz Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May, 19-21. European Language Resources Association (ELRA).

Jiri Hana and Anna Feldman. 2004. Portable Language Technology: Russian via Czech. In *Proceedings from the Midwest Computational Linguistics Colloquium, June 25-26, 2004*, Bloomington, Indiana.

Knut Hofland and Stig Johansson. 2006. The Translation Corpus Aligner: A program for alignment of paralell texts. In Stig Johansson and Signe Oksefjell, editors, *Corpora and Cross-linguistic Research. Theory, Method, and Case Studies*, number 24 in Language and Computers: Studies in practical Linguistics, pages 87–100. Rodopi, Amsterdam – Atlanta.

Stig Johansson, Jarle Ebeling, and Humanities Bergen. 1996. Coding and Aligning the English-Norwegian Parallel Corpus. In B. Altenberg K. Aijmer and M. Johansson, editors, *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies (Lund, 4-5 March 1994)*, pages 87–112, Lund. Lund University Press.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 2 edition.

Thomas Rathsack. 2009. Jæger – i krig med eliten. `http://www.politiken.dk`, September, 16.

Diana Santos and Signe Oksefjell. 2000. An evaluation of the translation corpus aligner, with special reference to the language pair English-Portuguese. In *Proceedings of the 12th Nordisk datalingvistikkdager, Trondheim, Department of Linguistics, NTNU*.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR*, abs/cs/0609058.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference. RANLP, Borovets.*, pages 590–596, Bulgaria.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955