Improving Natural Language Processing subtasks using statistical machine translation techniques

Alexander Wallin

Lunds Universitet, Faculty of Engineering
alexander@tlth.se

Abstract

In this paper we explore the use of statistical machine translation techniques and bilingual corpora and how this applies to coreference resolution for the Swedish language.
The results are compared with the current method of hand annotation of training corpora and probable sources of errors are discussed.

1. Introduction

Natural Language Processing (NLP) sub-tasks such as coreference resolution and named entity recognition are generally trained using supervised machine learning techniques (Rahman and Ng, 2012; Nothman et al., 2013), though semi-supervised methods achieve comparable results for many subtasks(Lin and Wu, 2009).
Supervised machine learning techniques require large corpora to achieve good results as well as to quantify and compare the model for its intended application.
Dependent on the specific language and NLP sub-task the collection of a corpus may require manually annotating large amounts of texts, which is highly resource intensive(Rahman and Ng, 2012).
Semi-supervised graph based projection is a possible method for generating annotated data for natural languages that lacks a sufficiently large corpora to train various NLP sub-tasks by using a bilingual corpora where the annotated corpora of one language is used to annotate an unlabeled corpora of a different language. This method yields good results in POS-tagging (Das and Petrov, 2011), dependency parsing (McDonald et al., 2011) as well as Named Entity Recognition (Che et al., 2013), among other. In this article we evaluate the graph projection approach for pronomial coreference resolution using the Europarl bilingual corpora (Koehn, 2005) to create a Swedish coreference corpora by projecting from the English corpus annotated by Stanford's CoreNLP (Manning et al., 2014). Annotations for Part-of-Speech tagging and dependency parsing for the Swedish corpora uses the commonly available Stagger (Östling, 2013) and MaltParser (Nivre et al., 2007). The results are contrasted with Språkbanken's SUC-Core, which is a 20 000 word hand annotated corpus for coreference resolution among other uses (Nilsson Björkenstam, 2013).

2. The Alignment

The Europarl corpora consists of 21 language specific corpora and sentence level alignment from one language to that of another using tools based on the Church and Gale algorithm (Gale and Church, 1993). The Swedish corpus consists of 45 665 947 words arrayed into 2 241 386 sentences. The English corpus consists of 53 974 751 words arrayed into 2 218 201 sentences. The overlapping English-Swedish alignment consists of 1 862 234 sentences. The corpus consists of both written text and html. Text containing xml tags were not considered for modeling. The aligned sentences are roughly 10% of the total number of sentences after the exclusion of documents containing xml tags.
The words were aligned using two different algorithms; Giza++ (Och and Ney, 2003) which uses algorithms based on the IBM Alignment model 4 (Och and Ney, 2000) and an algorithm partially based on Part-of-Speech alignment in conjunction with word equivalence, described in Section 2.1.
The coreference resolvers were trained using a pairwise classification algorithm (Soon et al., 2001) with a well balanced lexical feature set (Björkelund and Nugues, 2011) modelled using logistic regression (Fan et al., 2008).

2.1 Alternative alignment algorithm

The IBM Alignment models have well known deficiencies (Knight and Koehn, 2003), which in combination with the lackluster results from Europarl's sentence alignment yielded strange errors in the coreference resolver's model. A simple algorithm was therefore implemented that aligned nouns if there were string equivalence between the two languages and aligned pronouns in accordance to their ordinal numbering. This yielded strongly pessimistic training data for the coreference resolver.

3. Evaluation

Logistic regression were used to create training sets for the two word alignment methods. Confusion matrices were calculated using the whole corpora. The method using the IBM alignment model in Table 3.yielded a F1 score of 55 with a large training corpus, whereas the more pessimistic model in Table 2 yielded a F1 score of 63, but with a very limited training size.
Neither method used in this article would compare favorable with SUC-Core as the generated corpus would either contain systematic errors or be smaller than the well established hand annotated corpus.

		Predicted		
		Positive	Negative	Total
Actual	Positive	170,009	77,256	247,265
	Negative	203,611	1,619,136	1,847,101
	Total	373,620	1,813,499	2,094,366

Table 1: IBM alignment model 4 confusion matrix

		Predicted		
		Positive	Negative	Total
Actual	Positive	505	408	993
	Negative	110	1175	1285
	Total	1680	519	2199

Table 2: Alternative algorithm confusion matrix

4. Conclusions

Using the Europarl corpus resulted in compounded errors from the Gale-Church algorithm used for sentence alignment as well as whichever word alignment method were used. The solution is to either eliminate the errors from the Gale-Church algorithm, i.e. use a corpus where the sentences are more aligned such as in a book, or alternatively use different methods for aligning words and concepts in the two languages.

This methodology shows great promise, especially for languages which lacks hand annotated corpora, though the possible pitfalls and qualitative differences between hand annotated corpora and machine generated equivalents cannot be understated.

Acknowledgements

I would like to thank my thesis advisor Pierre Nugues at the Faculty of Engineering at Lund University for his ongoing support and inspiration.

References

Anders Björkelund and Pierre Nugues. 2011. Exploring lexicalized features for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 45–50. Association for Computational Linguistics.

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *HLT-NAACL*, pages 52–62.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 600–609. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102, March.

Kevin Knight and Philipp Koehn. 2003. What's new in statistical machine translation.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Dekang Lin and Xiaoyun Wu. 2009. Phrase clustering for discriminative learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1030–1038. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristina Nilsson Björkenstam. 2013. Suc-core: A balanced corpus annotated with noun phrase coreference. *Northern European Journal of Language Technology (NEJLT)*, 3:19–39.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151 – 175.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 440–447, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, 3:1–18.

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 720–730, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.