Chapter 10

Natural language processing

Krista Lagus, Mikko Kurimo, Timo Honkela, Sami Virpioja, Oskar Kohonen, Mari-Sanna Paukkeri, Mikaela Klami, Ville Turunen, Matti Varjokallio, Matti Pöllä, Ilari Nieminen, Tommi Vatanen

10.1 Introduction

Work in the field of natural language processing involves several research themes that have close connections to work carried out in other groups, especially speech recognition (Chapter 8) and Computational Cognitive Systems groups (Chapter 11). The objective of this research is to develop methods for learning general-purpose representations from text that can be applied to the recognition, understanding and generation of natural language. The results are evaluated in applications such as automatic speech recognition, information retrieval, and statistical machine translation.

During 2008–2009, our research has concentrated on finding suitable units of representations, such as morphemes, constructions, and keyphrases, in an unsupervised and language-independent manner. In addition, we have organized Morpho Challenges, international competitions funded by EU's PASCAL network, where multiple evaluations have been provided for algorithms for unsupervised morpheme analysis.

10.2 Unsupervised learning of morphology

In the theory of linguistic morphology, morphemes are considered to be the smallest meaning-bearing elements of language, and they can be defined in a language-independent manner. It seems that even approximative automated morphological analysis is beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. These applications usually make use of *words* as vocabulary units. However, for highly-inflecting and agglutinative languages, this leads to very sparse data, as the number of possible word forms is very high.

Figure 10.1 shows the very different rates at which the vocabulary grows in various text corpora of the same size. The number of different unique word forms in the Finnish corpus is considerably higher than in the English ones, for example. In addition to the language, the size of the vocabulary is affected by the text type.

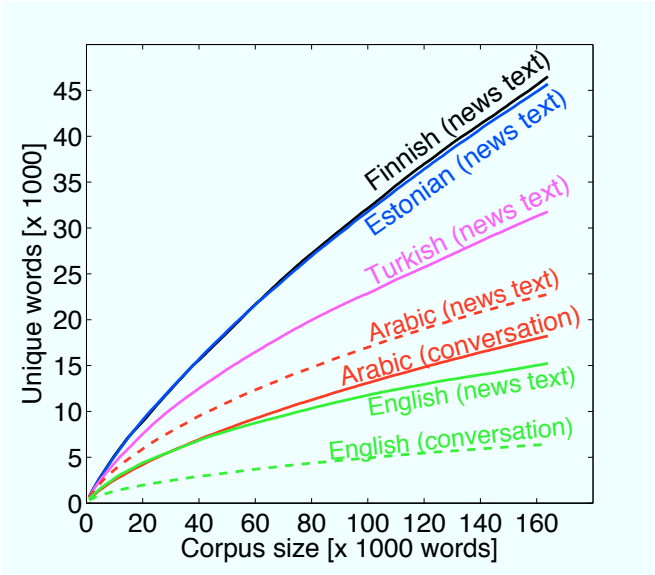

Figure 10.1: The number of different word forms (types) encountered in growing portions of running text (tokens) of various languages and text types.

We have developed language-independent, data-driven methods for the unsupervised discovery of morphemes. *Morfessor* [1] is a family of methods that perform segmentation of words into morpheme-like units. The different versions of Morfessor can be seen as instances of a general model. The model is strongly inspired by the Minimum Description Length (MDL) principle, although the later versions have been expressed in Maximum A Posteriori (MAP) estimation framework [2].

In *Allomorfessor*, Morfessor has been extended to account for the linguistic phenomenon of allomorphy. In allomorphy, an underlying morpheme-level unit has two or more surface realizations (e.g., "day" has an alternative surface form "dai" in "daily"). Recognizing the morpheme-level units should help with applications such as information retrieval and machine translation. In [3], the initial algorithm was evaluated in Morpho Challenge 2008 for English, Finnish, German, and Turkish languages, with moderate but promising results. An improved model performed significantly better in linguistic evaluation [4]. In Morpho Challenge 2009, the new Allomorfessor version performed very well in all languages and tasks, although the amount of allomorphs found by the algorithm was still limited [5].

References

[1] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[2] Krista Lagus, Mathias Creutz, Sami Virpioja, and Oskar Kohonen. Morpheme segmentation by optimizing two-part MDL codes. In *2009 Workshop on Information Theoretic Methods in Science and Engineering (WITMSE)*, Tampere, Finland, August 2009. Extended abstract.

[3] Oskar Kohonen, Sami Virpioja, and Mikaela Klami. Allomorfessor: Towards unsupervised morpheme analysis. In *Working Notes of the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[4] Oskar Kohonen, Sami Virpioja, and Mikaela Klami. Allomorfessor: Towards unsupervised morpheme analysis. In *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008 Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of Lecture Notes in Computer Science, pages 975-982. Springer, 2009.

[5] Sami Virpioja and Oskar Kohonen. Unsupervised morpheme analysis with Allomorfessor. In *Working Notes of the CLEF 2009 Workshop*, Corfu, Greece, 2009.

10.3 Unsupervised discovery of constructions

Construction grammar, originally developed by Charles Fillmore, is a grammatical theory that is beneficial for Natural Language Processing because it provides tools for modeling properties of language that traditional theory ignores (for an overview, see [1]). In particular, the different statistical properties of collocations, multi-word units and idioms are well known in Natural Language Processing.

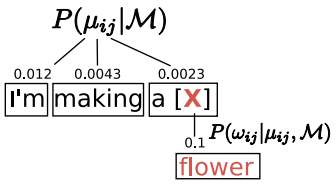

Figure 10.2: An illustration of the construction model used in [2].

In [2] we extended previous work in morphology learning into a method for learning multi-word constructions, as illustrated in figure 10.2. Since the construction grammar framework is a general one, in [3] we developed a framework for construction learning problems that includes both learning syntax and morphology.

References

[1] Adele E. Goldberg. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224, 2003.

[2] Krista Lagus, Oskar Kohonen, and Sami Virpioja. Towards unsupervised learning of constructions from text. In Magnus Sahlgren and Ola Knutsson, editors, *Proceedings of the Workshop on Extracting and Using Constructions in NLP of 17th Nordic Conference on Computational Linguistics, NODALIDA*, May 2009. SICS Technical Report T2009:10.

[3] Oskar Kohonen, Sami Virpioja, and Krista Lagus. A constructionist approach to grammar inference. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, Whistler, Canada, December 2009. Extended abstract.

10.4 Keyphrase extraction

A language-independent keyphrase extraction method, *Likey*, was developed as a follow-on to the earlier language-independent studies. The method utilises statistical analysis of language and comparison to a reference corpus, and it has a light-weight preprocessing phase. Most of the traditional methods for keyphrase extraction are highly dependent on the language used and the need for preprocessing is extensive. On the contrary, *Likey* enables independence from the language being analysed. It is possible to extract keyphrases from text in previously unknown language provided that a suitable reference corpus is available. *Likey* was tested with 11 European languages, including Germanic and Romance languages, Greek and Finnish. The evaluation method was based on Wikipedia articles and their intra-linking. The results were comparable to *tf.idf*, a statistical term weighting method. [1] The keyphrases produced by *Likey* were utilised as features in a web-based interface for collecting and analysis of information on authors and their publications [2].

A web-based demonstration of *Likey* is available at http://cog.hut.fi/likeydemo/. The system highlights keyphrases of a web document written in any of the eleven European languages. Keyphrases extracted from an article in a French online newspaper Le Monde are visualized by the demo in Figure 10.3. For example, the American swimmer "Michael Phelps" and word pair "médailles d'or" (English: *gold medals*) are extracted as keyphrases.

Figure 10.3: Keyphrases extracted by *Likey* from a French online news article.

References

[1] Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä, and Timo Honkela. A language-independent approach to keyphrase extraction and evaluation. In *Coling 2008: Companion volume: Posters*, pages 83–86, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[2] Tommi Vatanen, Mari-Sanna Paukkeri, Ilari T. Nieminen, and Timo Honkela. Analyzing authors and articles using keyword extraction, self-organizing map and graph algorithms. In *Proceedings of the AKRR08*, pages 105-111, 2008.

10.5 Morpho Challenge

Morpho Challenge is a series of scientific competition annually organized by Adaptive Informatics Research Centre for the evaluation of new unsupervised morpheme analysis algorithms. The challenge is part of the EU Network of Excellence PASCAL Challenge Program and in 2008 and 2009 organized in collaboration with Cross-Language Evaluation Forum CLEF.

The objective of the challenge is to design statistical machine learning algorithms that discover which morphemes (smallest individually meaningful units of language) words consist of. Ideally, these are basic vocabulary units suitable for different tasks, such as text understanding, machine translation, information retrieval, and statistical language modeling. The challenge has sofar been organized four times and the results have been published in PASCAL and CLEF workshops in Venice 2006 [1], in Budapest 2007 [2], in Aarhus 2008 [3, 4], and in Corfu 2009 [5].

In the 2009 Morpho Challenge, the evaluation of the submissions have performed by three complementary ways: *Competition 1*: The proposed morpheme analyses were compared to a linguistic morpheme analysis gold standard by matching the morpheme sharing word pairs [5]. *Competition 2*: Information retrieval (IR) experiments were performed, where the words in the documents and queries were replaced by their proposed morpheme representations and the search was based on morphemes instead of words. The IR evaluations were provided for Finnish, German, and English and participants were encouraged to apply their algorithm to all of them. The organizers performed the IR experiments using the queries, texts, and relevance judgments available in CLEF forum and morpheme analysis methods submitted by the challenge participants. The results show that the morpheme analysis has a significant effect in IR performance in all languages, and that the performance of the best unsupervised methods can be superior to the supervised reference methods. *Competition 3*: Statistical machine translation (SMT) experiments were performed, where the words in the source language sentences were replaced by their proposed morpheme representations and the alignment and translation was based on morphemes instead of words [5]. To make te results relevant to the state-of-the-art in SMT, the N-best translation hypotheses of the morpheme-based system were further combined with a conventional word-based system. The word-based system was trained with the same data, but keeping the words unsplit, and the combination was performed by using the minimum Bayes risk combination as in [6]. The experimented language-pairs were Finnish-English and German-English and the results showed that the best unsupervised methods improve the baseline word-based system.

References

[1] Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy and Murat Saraclar. Unsupervised segmentation of words into morphemes - Challenge 2005, An Introduction and Evaluation Report. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*. Venice, Italy, April 12, 2006.

[2] Mikko Kurimo, Mathias Creutz, and Matti Varjokallio. Morpho Challenge evaluation using a linguistic Gold Standard. In *Advances in Multilingual and MultiModal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, Lecture Notes in Computer Science, Vol. 5152, pages 864-873. Springer, 2008.

[3] Mikko Kurimo, Ville Turunen, and Matti Varjokallio. Overview of Morpho Challenge 2008. In *Advances in Multilingual and MultiModal Information Retrieval, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, Lecture Notes in Computer Science. Springer, 2009.

[4] Mikko Kurimo, Mathias Creutz, and Ville Turunen. Morpho Challenge evaluation by information retrieval experiments. In *Advances in Multilingual and MultiModal Information Retrieval, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, Lecture Notes in Computer Science. Springer, 2009.

[5] Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. Overview and results of Morpho Challenge 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September 2009.

[6] Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73-76, Boulder, USA, June 2009. Association for Computational Linguistics.