Chapter 10

Natural language processing

Krista Lagus, Mikko Kurimo, Timo Honkela, Sami Virpioja, Oskar Kohonen, Mari-Sanna Paukkeri, Tiina Lindh-Knuutila, Ville T. Turunen, Ilkka Kivimäki, Laura Leppänen, Sini Pessala, Santosh Tirunagari

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 **??**). 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 2010–2011, our research has concentrated on minimally supervised and language-independent methods for morphology induction, keyphrase extraction, and creation and evaluation of vector space models.

10.2 Unsupervised and semi-supervised morphology induction

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 morphological analysis is beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. These applications usually use *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. For example, the number of different unique word forms in the Finnish corpus is considerably higher than in the English corpus. In addition to the language, the size of the vocabulary is affected by the genre.

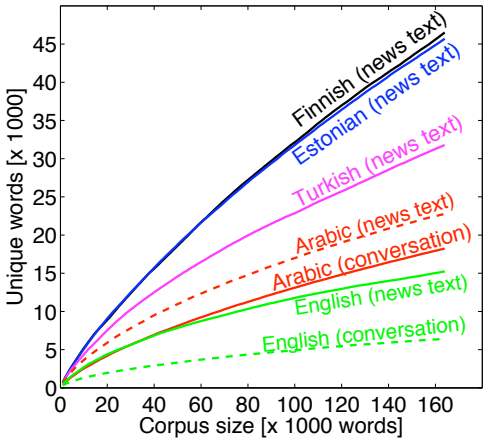

Figure 10.1: The number of different word forms (types) encountered in growing portions of running text (tokens) of various languages and text types.

Apart from practical use in various natural language processing applications, learning the phenomena underlying word construction in natural languages is an important question in psycholinguistics. Psycholinguistic questions regarding morphology include, for example, how the different word forms are learned, constructed, and stored in our mind in the so-called mental lexicon.

In 2010, we continued the series of Morpho Challenge competitions previously organized in 2005, 2007, 2008, and 2009. The objective of Morpho Challanges is to design statistical machine learning algorithms that discover discover the set of morphemes from which words are constructed [1]. The Morpho Challenge 2010 was funded by the EU Network of Excellence PASCAL2 Challenge Program. The evaluations included four languages and three evaluations: comparison to a linguistic gold standard, evaluation in an information retrieval task, and evaluation in a machine translation task. As a new task we introduced semi-supervised learning, in which a small set of linguistic gold standard morpheme analyses are provided as a training set. Four international groups participated in the Challenge, and the results and algorithms were published in a technical report [2].

Based on the Morpho Challenge results collected over five years, we have performed an extensive meta-evaluation of various evaluation methods for unsupervised learning of morphology [3]. Apart from comparing existing methods, we have further developed the evaluation methods and published evaluation software for the research community.

We have also continued to develop Morfessor [4], an unsupervised method for morphology induction. In *Allomorfessor* [5], 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"). Allomorfessor has performed well in Morpho Challenge evaluations, although the amount of allomorphs found by the algorithm was limited.

In order to enable Morfessor to model complex morphological phonomena such as allomorphy, as well as to provide a reasonable baseline for the semi-supervised learning evaluated in Morpho Challenge 2010, we have also developed a semi-supervised learning algorithms for Morfessor [6]. The linguisic evaluation of Morpho Challenge shows that the accuracy of Morfessor improves rapidly already with small amounts of labeled data, surpassing the state-of-the-art unsupervised methods at 1000 labeled words for English and at 100 labeled words for Finnish. A further extension of the method has achieved the best published results for the semi-supervised learning setup of Morpho Challenge [7]. We have also studied the effect of word frequencies learning in generative models of morphology such as Morfessor, and found that using logaritmically dampened frequencies seem to provide better results than learning on word tokens and at least as good results than learning on word types [8].

Finally, in collaboration with the Brain Research Unit of the O.V. Lounasmaa laboratory at Aalto University, we have developed psycholinguistic framework for evaluating machine learning of morphology [9]. We use reaction times in a word recognition task as a proxy that provides an indirect measure of the underlying mental processing. In general, longer reaction times reflect more effortful cognitive processing. In comparison of several statistical models revealed that Morfessor Categories-MAP [4] provides an accurate and compact model for the reaction time data. Moreover, we observed a strong effect for the type and amount of the training data to the correlations. Figure 10.2 shows how Morfessor Categories-MAP predicts too high reaction times for abstract words such as *knowledge* and too low reactions times for concrete words such as *mother*.

References

[1] Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87-95, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[2] Mikko Kurimo, Sami Virpioja, and Ville T. Turunen (Eds.). *Proceedings of the Morpho Challenge 2010 workshop*. Technical Report TKK-ICS-R37, Aalto University School of Science and Technology, Department of Information and Computer Science, Espoo, Finland, September 2010.

[3] Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2), 2011.

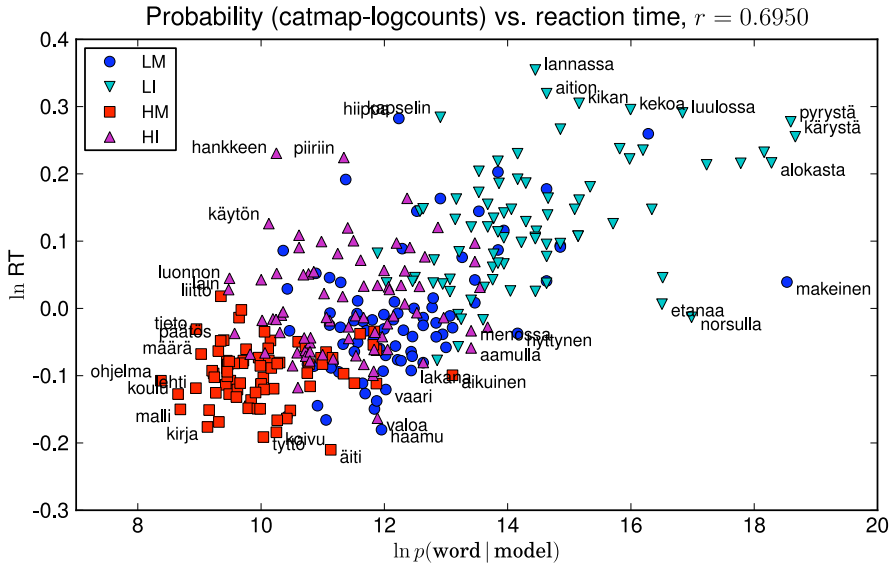

Figure 10.2: Scatter plot of reaction times and log-probabilities from Morfessor Categories-MAP [9]. The words are divided into four groups: low-frequency monomorphemic (LM), low-frequency inflected (LI), high-frequency monomorphemic (HM), and high-frequency inflected (HI). Words that have faster reaction times than predicted are often very concrete and related to family, nature, or stories: *tyttö* (girl), *äiti* (mother), *haamu* (ghost), *etanaa* (snail + partitive case), *norsulla* (elephant + adessive case). Words that have slower reaction times than predicted are often more abstract or professional: *ohjelma* (program), *tieto* (knowledge), *hankkeen* (project + genitive case), *käytön* (usage + genitive case), *hiippa* (miter), *kapselin* (capsule + genitive case).

[4] 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.

[5] Sami Virpioja, Oskar Kohonen, and Krista Lagus. Unsupervised morpheme analysis with Allomorfessor. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, volume 6241 of Lecture Notes in Computer Science, pages 609-616. Springer Berlin / Heidelberg, September 2010.

[6] Oskar Kohonen, Sami Virpioja, and Krista Lagus. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78-86, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[7] Oskar Kohonen, Sami Virpioja, Laura Leppänen, and Krista Lagus. Semi-supervised extensions to Morfessor Baseline. In Mikko Kurimo, Sami Virpioja, and Ville T. Turunen, editors, *Proceedings of the Morpho Challenge 2010 Workshop*, pages 30–34, Espoo, Finland, September 2010. Aalto University School of Science and Technology, Department of Information and Computer Science. Technical Report TKK-ICS-R37. Extended abstract.

[8] Sami Virpioja, Oskar Kohonen, and Krista Lagus. Evaluating the effect of word frequencies in a probabilistic generative model of morphology. In Bolette Sandford Pedersen, Gunta NeÅ¡pore, and Inguna Skadina, editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, volume 11 of NEALT Proceedings Series, pages 230-237. Northern European Association for Language Technology, Riga, Latvia, May 2011.

[9] Sami Virpioja, Minna Lehtonen, Annika Hultén, Riitta Salmelin, and Krista Lagus. Predicting reaction times in word recognition by unsupervised learning of morphology. In Timo Honkela, Wlodzislaw Duch, Mark Girolami, and Samuel Kaski, editors, *Artificial Neural Networks and Machine Learning — ICANN 2011*, volume 6791 of Lecture Notes in Computer Science, pages 275–282. Springer Berlin / Heidelberg, June 2011.

10.3 Keyphrase extraction

A language-independent keyphrase extraction method, *Likey*, extracts keyphrases from a document using phrase frequency ranks and comparison to a reference corpus. It has a light-weight preprocessing phase, whereas most of the other methods for keyphrase extraction are highly dependent on the language used and the need for preprocessing is extensive. Many of them need also a training corpus. On the contrary, *Likey* enables independence from the language used. It is possible to extract keyphrases from text in previously unknown language, provided that a suitable reference corpus is available. The method was further developed and applied for a set of scientific articles [1]. The evaluation was conducted against both author-provided and manually extracted keyphrases in the articles.

Learning taxonomic relations

As an application for the *Likey* keyphrase extraction method, a method for learning taxonomic relations from a set of text documents was developed [2]. *Likey* and two other methods for feature extraction were used to create document vectors for Wikipedia articles about animals in English and Finnish. The vectors were clustered hierarchically using the Self-Organizing Map (SOM). The resulting taxonomy were compared to a scientific classification of the animals, that can be seen in Figure 10.3.

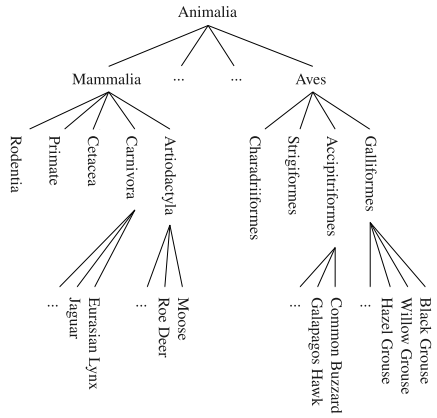

Figure 10.3: Part of the reference taxonomy.

References

[1] Mari-Sanna Paukkeri and Timo Honkela (2010) Likey: Unsupervised Language-Independent Keyphrase Extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics. Uppsala, Sweden, July 2010.

[2] Mari-Sanna Paukkeri, Alberto Pérez García-Plaza, Sini Pessala, and Timo Honkela (2010) Learning taxonomic relations from a set of text documents. In *Proceedings of 5th International Symposium Advances in Artificial Intelligence and Applications (AAIA'10)*. Wisla, Poland, October 2010.

10.4 Vector space models of language

Vector space models are a standard way to represent documents or words as vectors of features. The model provides a solution to the problem of representing symbolic information (words) in numerical form for computational processing. In a vector space, similar items are close to each other, and the closeness can be measured using vector similarity measures. Vector space models are applied, for example, in various information retrieval tasks and text categorization tasks.

Dimensionality reduction in document clustering

In document clustering, semantically similar documents are grouped together. The dimensionality of document collections is often very large, thousands or tens of thousands of terms. Thus, it is common to reduce the original dimensionality before clustering. Cosine distance is widely seen as the best choice for measuring the distances between documents in k-means clustering. The effect of dimensionality reduction on different distance measures in document clustering was analysed in [1]. The results show that after dimensionality reduction into small target dimensionalities, such as 10 or below, the superiority of cosine measure does not hold. Also, for small dimensionalities, PCA dimensionality reduction method performs better than SVD. Further, the effect of l_2 normalization for different distance measures was studied. The experiments are run for three document sets in English and one in Hindi.

Analysis of adjectives in a word vector space

Large number of studies indicate that methods using co-occurrence data provide useful information on the relationships between the words, as words with similar or related meaning will tend to occur in similar contexts. This intuition has been carefully assessed, in particular, for nouns and verbs. In [2], we study how well the co-occurrence statistics provide a basis for automatically creating a representation for a group of adjectives as well. In this study, a the text collection used was extracted from English Wikipedia, and the evaluation was carried out with 72 adjectives which formed 36 antonym pairs (i.e. good-bad). Further, we compare three dimension reduction methods and their effect on the quality of final representation: The Principal Component Analysis, the Self-Organizing Map and Neighbor Retrieval Visualizer (NeRV). Figure 10.4 visualizes the adjectives and their neighbors after dimension reduction with the NeRV.

Vector space evaluation using CCA

The vector spaces are generated using different feature extraction methods for text data. However, evaluation of the feature extraction methods may be difficult. Indirect evaluation in an application is often time-consuming and the results may not generalize to other applications, whereas direct evaluations that measure the amount of captured semantic information usually require human evaluators or annotated data sets.

We have developed a novel direct evaluation method for vector space models of documents based on canonical correlation analysis (CCA) [3]. The evaluation method is based on

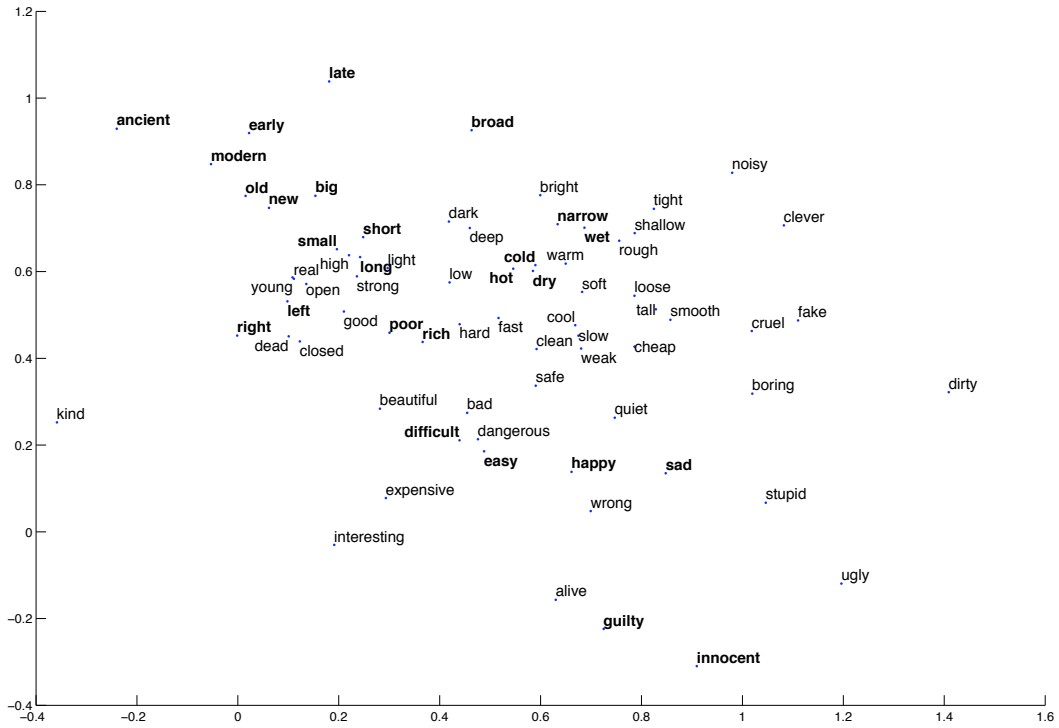

Figure 10.4: The set of adjectives used in the study projected into a 2-dimensional space using the Neighbor Retrieval Visualizer (NeRV) method. The words in bold have the antonym in their local neighborhood.

unsupervised learning, it is language and domain independent, and it does not require additional resources besides a parallel corpus.

CCA is a classical method for finding linear relationship between two data sets. In our setting, the two sets are parallel text documents in two languages. A good feature extraction method should provide representations that reflect the semantic contents of the documents. We assume that the underlying semantic contents is independent of the language, illustrated by the generation model on the left part of Figure 10.5. Then we can study which feature extraction methods capture the contents best by measuring the dependence between the representations of a document and its translation, illustrated on the right part of Figure 10.5.

In the case of CCA, the applied measure of dependence is correlation, which means that it can only find linear dependence. In a related study [4], we have shown that kernelized version of CCA outperforms linear CCA in a sentence matching task. Unfortunately, choosing the kernel and its parameters would require additional optimization step and held-out data for vector space evaluation.

We have demonstrated the proposed evaluation method on a sentence-aligned parallel corpus. The method was validated in three ways: (1) showing that the obtained results with bag-of-words representations are intuitive and agree well with the previous findings, (2) examining the performance of the proposed evaluation method with indirect evaluation methods in simple sentence matching tasks, and (3) in a quantitative manual evaluation of word translations. The results of the evaluation method correlate well with the results

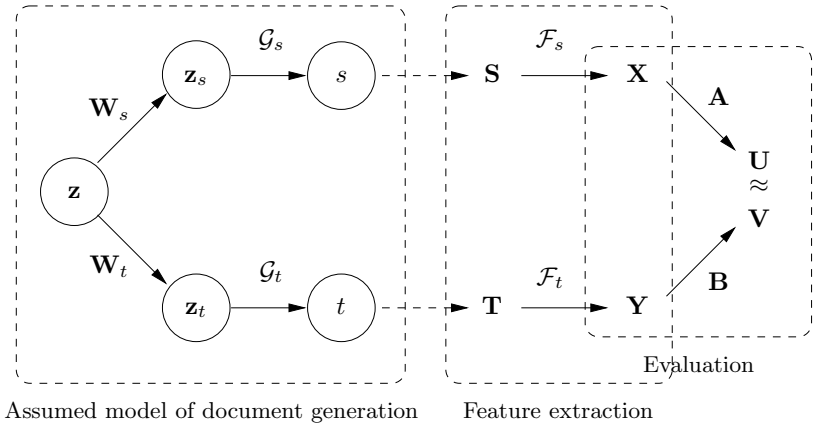

Figure 10.5: On the left: Assumed model for generation of documents s and t. Vector $\mathbf{z}$ in the language-independent semantic space $\mathcal{Z}$ is projected onto vectors $\mathbf{z}_s$ and $\mathbf{z}_t$ in the language-specific subspaces $\mathcal{Z}_s$ and $\mathcal{Z}_t$. Processes $\mathcal{G}_s$ and $\mathcal{G}_t$ generate document pairs from the respective subspaces. On the right: The process of evaluating feature extraction method $\mathcal{F}$ with CCA. The aligned document collections $\mathbf{S}$ and $\mathbf{T}$ are reduced to matrices $\mathbf{X}$ and $\mathbf{Y}$ of feature vectors using $\mathcal{F}$. Then $\mathbf{X}$ and $\mathbf{Y}$ are projected onto a common vector space using CCA.

of the indirect and manual evaluations.

References

[1] Mari-Sanna Paukkeri, Ilkka Kivimäki, Santosh Tirunagari, Erkki Oja, and Timo Honkela (2011) Effect of Dimensionality Reduction on Different Distance Measures in Document Clustering. In B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part III, LNCS 7064, pp. 167-176. Springer-Verlag Berlin Heidelberg.

[2] Timo Honkela and Tiina Lindh-Knuutila and Krista Lagus (2010) Measuring Adjective Spaces. In K. Diamantaras, W. Duch, L. S. Iliadis (Eds.): Proceedings of ICANN 2010, Artificial Neural Networks, pp. 368-373. Springer Verlag Berlin Heidelberg.

[3] Sami Virpioja, Mari-Sanna Paukkeri, Abhishek Tripathi, Tiina Lindh-Knuutila, and Krista Lagus. Evaluating vector space models with canonical correlation analysis. *Natural Language Engineering*, to appear. Available on CJO 2011.

[4] Abhishek Tripathi, Arto Klami, and Sami Virpioja. Bilingual sentence matching using kernel CCA. In *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, pages 130–135, Kittilä, Finland, August 2010. IEEE.