Chapter 10

Natural language processing

Krista Lagus, Timo Honkela, Mathias Creutz, Mikko Kurimo, Sami Virpioja, Jaakko Väyrynen, Oskar Kohonen, and Krister Lindén

10.1 Unsupervised segmentation of words into morphs

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. Many existing applications make use of *words* as vocabulary units. However, for some languages, e.g., Finnish and Turkish, 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 Finnish and English text corpora of the same size. The number of different unique word forms in the Finnish corpus is considerably higher than in the English one.

We have developed *Morfessor*, a language-independent, data-driven method for the unsupervised segmentation of words into morpheme-like units. There are different versions of Morfessor, which correspond to consecutive steps in the development of the model [1, 2, 3, 4]. All versions can be seen as instances of a general model, as described in [5].

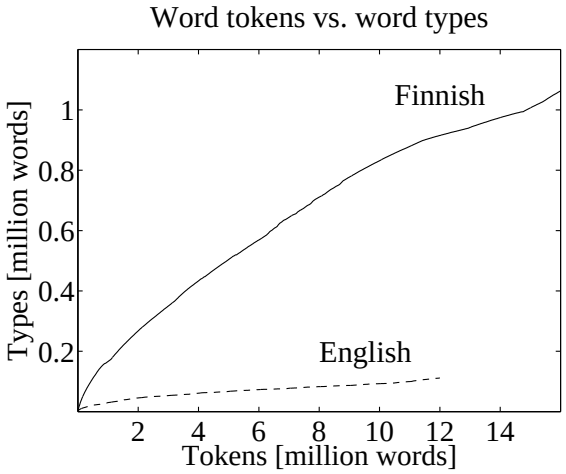

Figure 10.1 The number of different word forms (types) encountered in growing portions of running text (tokens) of Finnish and English.

The general idea behind the Morfessor model is to discover as compact a description of the data as possible. Substrings occurring frequently enough in several different word forms are proposed as *morphs* and the words are then represented as a concatenation of morphs, e.g., "hand, hand+s, left+hand+ed, hand+ful".

An optimal balance is sought between compactness of the *morph lexicon* versus the compactness of the representation of the *corpus*. The morph lexicon is a list of all distinct morphs (e.g., "hand, s, left, ed, ful") together with some stored properties of these morphs. The representation of the corpus can be seen as a sequence of pointers to entries in the morph lexicon; e.g. the word "lefthanded" is represented as three pointers to morphs in the lexicon.

Among others, de Marcken [6], Brent [7], and Goldsmith [8] have shown that the above type of model produces segmentations that resemble linguistic morpheme segmentations, when formulated mathematically in a probabilistic framework or equivalently using the Minimum Description Length (MDL) principle [9].

A shortcoming of previous splitting methods is that they either do not model *context-dependency* or they *limit the number of splits* per word to two or three. For instance, failure to incorporate context-dependency in the model may produce splits like "s+wing, ed+ward, s+urge+on" on English data, since the morphs "-s" and "-ed" are frequently occurring suffixes in the English language, but the algorithm does not make this distinction and thus suggests them in word-initial position as prefixes. By limiting the number of allowed segments per word the search task is alleviated and context-dependency can be modeled. However, this makes it impossible to correctly segment compound words with several affixes (pre- or suffixes), such as the Finnish word "aka+n+kanto+kiso+i+ssa"

aarre + **kammio** + *i* + *ssa*, **aarre** + **kammio** + *nsa*, **bahama** + **saar** + *et*,
bahama + **saari** + *lla*, **bahama** + **saar** + *ten*, **edes** + **autta** + *isi* + *vat*,
edes + **autta** + *ma* + *ssa*, nais + **auto** + *ili* + *ja* + *a*, pää + **aihe** + *e* + *sta*,
pää + **aihe** + *i* + *sta*, **pää** + *hän*, taka + **penkki** + *lä* + *in*+ *en*, **voi** + *mme* + *ko*
abandon + *ed*, **abandon** + *ing*, **abandon** + *ment*, **beauti** + *ful*,
beauty + *'s*, **calculat** + *ed*, **calculat** + *ion* + *s*, **express** + *ion* + *ist*,
micro + **organ** + *ism* + *s*, **long** + **fellow** + *'s*, **master** + **piece** + *s*,
near + *ly*, **photograph** + *er* + *s*, **phrase** + *d*, un + **expect** + *ed* + *ly*
ansvar + *ade*, **ansvar** + *ig*, **ansvar** + *iga*, **ansvar** + *s* + för + **säkring** + *ar*,
blixt + ned + **slag**, **dröm** + *de*, **dröm** + *des*, **drömma** + *nde*, in + **lopp** + *et* + *s*,
in + **lägg** + *n* + *ing* + *ar*, **målar** + *e*, **målar** + **yrke** + *t* + *s*, o + ut + **nyttja** + *t*,
poli + *s* + **förening** + *ar* + *na* + *s*, **trafik** + **säker** + *het*, över + **fyll** + *d* + *a*

Figure 10.2: Examples of segmentations learned from data sets of Finnish, English, and Swedish text. Suggested prefixes are underlined, stems are rendered in **boldface**, and suffixes are *slanted*.

(transl. "in the wife-carrying contests").

We have focused our efforts on developing a segmentation model that incorporates context-dependency without restricting the number of allowed segments per word. This has resulted in two model variants [3, 4]. The former is based on Maximum Likelihood (ML) optimization, in combination with some heuristics. The latter constitutes an attempt to more elegant model formulation, within the Maximum a Posteriori (MAP) framework.

Evaluation

Morfessor has been evaluated in two complementary ways: directly by comparing to linguistic morpheme segmentations of Finnish and English words, and indirectly as a component of a large (or virtually unlimited) vocabulary Finnish speech recognition system. In both cases, Morfessor outperforms state-of-the-art solutions. The speech recognition experiments are described in Section 9.3.

In order to carry out the direct evaluation, linguistic reference segmentations needed to be produced as part of the current project, since no available resources were applicable as such. This work has resulted in a morphological "gold standard", called *Hutmegs* (Helsinki University of Technology Morphological Evaluation Gold Standard) [10, 11]. Hutmegs contains analyses for 1.4 million Finnish and 120 000 English word forms, which have been produced by further processing the contents of the Finnish Two-Level Morphological Analyzer from Lingsoft, Inc. and the English CELEX database from the Linguistic Data Consortium (LDC). Hutmegs is publicly available for research; inexpensive one-time license fees need to be paid to Lingsoft and the LDC, for access to the Finnish and English analyses, respectively.

When the latest context-sensitive Morfessor versions [3, 4] are evaluated against the Hutmegs gold standard, they clearly outperform a frequently used benchmark algorithm [8] on Finnish data, and perform as well or better than the benchmark on English data (depending on the size of the data sets used).

Some sample segmentations of Finnish, English, as well as Swedish words, are shown in Figure 10.2. These include correctly segmented words, where each boundary coincides with a linguistic morpheme boundary (e.g., "aarre+kammio+i+ssa, edes+autta+isi+vat, abandon+ed, long+fellow+'s, in+lopp+et+s"). In addition, some words are over-segmented,

with boundaries inserted at incorrect locations (e.g., "in+lägg+n+ing+ar" instead of "in+lägg+ning+ar"), as well as under-segmented words, where some boundary is missing (e.g., "bahama+saari+lla" instead of "bahama+saar+i+lla").

In addition to segmenting words, Morfessor suggests likely grammatical categories for the segments. Each morph is tagged as a prefix, stem, or suffix. Sometimes the morph categories can resolve the semantic ambiguity of a morph, e.g., Finnish "pää". In Figure 10.2, "pää" has been tagged as a stem in the word "pää+hän" ("in [the] *head*"), whereas it functions as a prefix in "pää+aihe+e+sta" ("about [the] *main* topic").

Demonstration and software

There is an online demonstration of Morfessor on the Internet: `http://www.cis.hut.fi/projects/morpho/`. Currently, the demo supports three languages: Finnish, English, and Swedish. Those interested in larger-scale experiments can download the Morfessor program and train models using their own data sets. The software is described in [12]. Within a period of ten months (April 2005 – January 2006) a monthly average of 18 downloads of the program has been registered.

Further applications

Outside the scope of the current project, Morfessor has been used successfully in the recognition of Turkish [13] as well as Estonian speech. Hagen and Pellom [14] apply Morfessor in English speech recognition intended for oral reading tracking within an interactive reading tutor program for children. Morfessor has also been used in Finnish information retrieval, both in the retrieval of text [15] and spoken documents [16] (Sec. 9.5). Furthermore, in a number of works on language modeling, the segments discovered by Morfessor constitute the basic vocabulary [17, 18, 19] (Section 9.3). Kumlander [20] has analyzed the word splits obtained when running Morfessor on stories told by Finnish children.

References

[1] Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia, Pennsylvania, USA, 2002.

[2] Mathias Creutz. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proc. ACL'03*, pages 280–287, Sapporo, Japan, 2003.

[3] Mathias Creutz and Krista Lagus. Induction of a simple morphology for highly-inflecting languages. In *Proc. 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona, July 2004.

[4] Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, 2005.

[5] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 2006. (Accepted for publication).

[6] C. G. de Marcken. *Unsupervised Language Acquisition*. PhD thesis, MIT, 1996.

[7] M. R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, 1999.

[8] John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

[9] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific Series in Computer Science, Singapore, 1989.

[10] Mathias Creutz and Krister Lindén. Morpheme segmentation gold standards for Finnish and English. Technical Report A77, Publications in Computer and Information Science, Helsinki University of Technology, 2004.

[11] Mathias Creutz, Krista Lagus, Krister Lindén, and Sami Virpioja. Morfessor and Hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, pages 107–112, Tallinn, Estonia, 4 – 5 April 2005.

[12] Mathias Creutz and Krista Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology, 2005.

[13] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, and M. Creutz. On lexicon creation for Turkish LVCSR. In *Proc. Eurospeech'03*, pages 1165–1168, Geneva, Switzerland, 2003.

[14] Andreas Hagen and Bryan Pellom. Data driven subword unit modeling for speech recognition and its application to interactive reading tutors. In *Proceedings of INTERSPEECH 2005*, pages 2757–2760, Lisbon, Portugal, September, 4–8 2005.

[15] Sam Engström. Information retrieval using unsupervisedly segmented morphemes. Special assignment, Laboratory of Computer and Information Science, Helsinki University of Technology, July 2005.

[16] Mikko Kurimo and Ville Turunen. To recover from speech recognition errors in spoken document retrieval. In *Proceedings of Interspeech 2005*, pages 605–608, Lisbon, Portugal, September 2005.

[17] Vesa Siivola and Bryan L. Pellom. Growing an n-gram language model. In *Proceedings of Interspeech 2005*, Lisbon, Portugal, September 2005.

[18] Simo Broman and Mikko Kurimo. Methods for combining language models in speech recognition. In *Proceedings of Interspeech 2005*, pages 1317–1320, Lisbon, Portugal, September 2005.

[19] Sami Virpioja. New methods for statistical language modeling. Master's thesis, Helsinki University of Technology, Department of Computer Science and Engineering, Laboratory of Computer and Information Science, 2005.

[20] Mikaela Kumlander. Forthcoming master's thesis. Master's thesis, University of Helsinki, Department of General Linguistics, 2005.

10.2 Word sense disambiguation using document maps

Krister Lindén, Krista Lagus

A single word may have several senses or meanings, for example "was *heading* south/the newspaper *heading* is", or "Church" as an institution versus "church" as a building. Word sense disambiguation automatically determines the appropriate senses of a particular word in context. It is an important and difficult problem with many practical consequences for language-technology applications in information retrieval, document classification, machine translation, spelling correction, parsing, and speech synthesis as well as speech recognition. For a textbook introduction, see [5]. In particular, Yarowsky [6] noted that words tend to keep the same sense during a discourse.

In [2] we introduce a method called THESSOM for word sense disambiguation that uses an existing topical document map, in this case a map of nearly 7 million patent abstracts, created with the WEBSOM method (see [1]). The method uses the document map as a representation of the semantic space of word contexts. The assumption is that similar meanings of a word have similar contexts, which are located in the same area on the self-organized document map. The results confirm this assumption. In this method, the existing general-purpose document map is calibrated, i.e., marked with correct senses, using a subset of data where the ambiguous words have been sense-tagged. The sense-calibrated map can then be utilized as a word sense classifier, for determining a probable correct sense for an ambiguous sample word in context. The data flow of the training and testing procedure is shown in Figure 10.3.

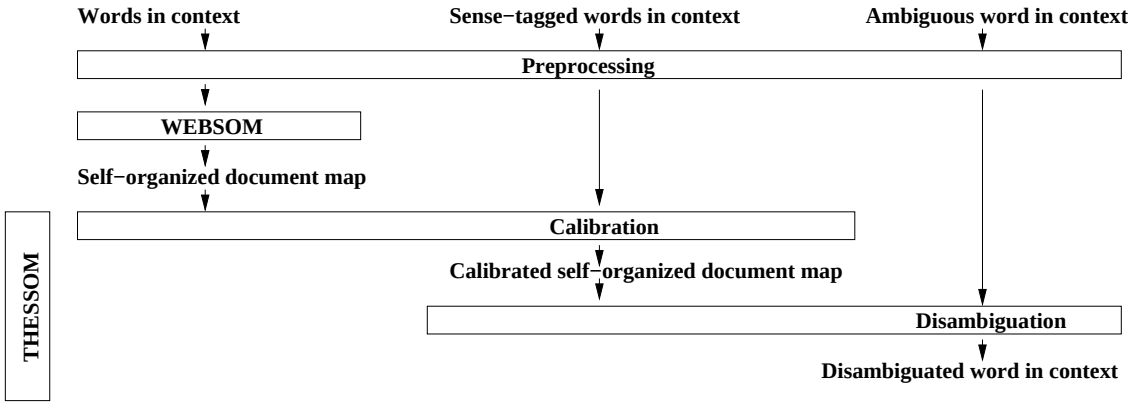

Figure 10.3: Data flow of word sense disambiguation with self-organized document maps

Results on the SENSEVAL-2 corpus (from a word sense disambiguation contest) indicate that the proposed method is statistically significantly better than the baselines, and performs on an average level when compared to the total of supervised methods in the competition. The benefit of the proposed method is that a single general purpose representation of the semantic space can be used for all words and their word senses.

In [3], instead of utilizing one general-purpose document map and merely calibrating (marking) it with particular sense locations, an individual document map is created for each ambiguous word from the training material (short contexts) for that word. Moreover, advanced linguistic analysis was performed using a dependency grammar parser to produce additional features for the document vectors. The training material consisted of a total of 8611 contexts for the 73 ambiguous words, i.e., on the average 118 contexts per word. As

a result, 73 maps were generated, one for each ambiguous word.

In [4], we evaluate the efficacy of various features for word sense disambiguation with THESSOM. We conclude that the syntactic features are the most important for word sense disambiguation and should be chosen if a single feature type needs to be selected. However, their feature space is sparse, so the features based on the base forms of words function as a kind of back-off model with a small but statistically significant improvement to the overall dismabiguation performance of THESSOM.

The algorithm was tested on the SENSEVAL-2 benchmark data and shown to be on a par with the top three contenders of the SENSEVAL-2 competition. It was also shown that adding more advanced linguistic analysis to the feature extraction seems to be essential for improving the classification accuracy. We conclude that self-organized document maps have properties similar to a large-scale semantic structure that is useful for word sense disambiguation.

References

[1] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585, May 2000.

[2] K. Lindén and K. Lagus. Word Sense Disambiguation in Document Space. In *2002 IEEE Int. Conference on Systems, Man and Cybernetics*, Tunisia, October 6–9, 2002.

[3] K. Lindén. Word Sense Disambiguation with THESSOM. In *Workshop on Self-Organizing Maps, WSOM'03 — Intelligent Systems and Innovational Computing*, Kitakyushu, Japan, September 11–14, 2003.

[4] K. Lindén. Evaluation of Linguistic Features for Word Sense Disambiguation with Self-Organized Document Maps. Computers and the Humanities, 2004 (December).

[5] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[6] D. Yarowsky. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 189–196, Cambridge, MA, 1995.

10.3 Topically focusing language model

A statistical language model provides predictions for future words based on the already seen word sequence. This is important, for example, in large vocabulary continuous speech recognition (see Section 9.3) to guide the search into those phoneme sequence candidates that constitute relevant words and sentences. Especially when the vocabulary is large, say 100 000 words, the estimation of the most likely words based on the previous sequence is challenging since all possible words, let alone all word sequences, have never been seen in any data set. For example, there exist 10^{25} sequences of 5 words of a vocabulary of 100 000 words. Thus directly estimating a n:th order Markov model is generally out of the question for values of n larger than 5.

In [1] we proposed a *topically focusing language model* that is built utilizing a topical clustering of texts obtained using the WEBSOM method. The long-term dependencies [2] are taken into account by focusing the predictions of the language model according to the longer-term topical and stylistic properties of the observed speech or text.

In speech recognition suitable text data or the recognizer output can be utilized to focus the model, i.e., to select the text clusters that most closely correspond to the current discourse or topic. Next, the focused model can be applied to speech recognition or to re-rank the result hypothesis obtained by a more general model [6].

It has been previously shown that good topically organized clustering of large text collections can be achieved efficiently using the WEBSOM method (see [3]). In this project, the clustering is utilized as a basis for constructing a focusing language model. The model is constructed as follows:

Cluster a large collection of topically coherent text passages, e.g., paragraphs or short documents using the WEBSOM method. For each cluster (e.g. for each map unit), calculate a separate, small n-gram model. During speech recognition, use transcription history and the current hypothesis to select a small number of topically 'best' clusters. Combine the language models of each cluster to obtain a focused language model. This model is thus focused on the topical and stylistic peculiarities of a history of, say, 50 words. Combine further with a general language model for smoothing. The structure of the resulting combined language model is shown in Figure 10.4.

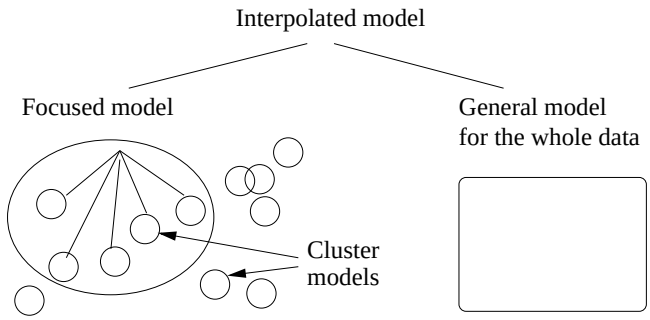

Figure 10.4: *A focusing language model obtained as an interpolation between topical cluster models and a general model.*

As the cluster-specific models and the general model we have used n-gram models of various orders. However, other types of models describing the short-term relationships between words could, in principle, be used as well. The combining operation amounts to a linear interpolation of the predicted word probabilities.

The models were evaluated using perplexity[1] on independent test data averaged over

[1]Perplexity is the inverse predictive probability for all the words in the test document.

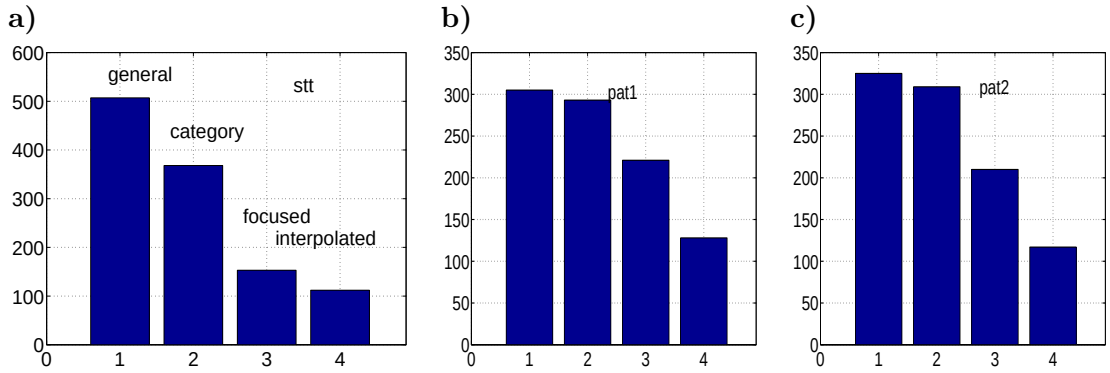

Figure 10.5: The perplexities of the different language models, **a)** for the Finnish STT news corpus, **b)** for smaller patent corpus and **c)** for larger patent corpus. The explanation of the bars in each figure, from left to right: 1. general model for the whole corpus, 2. category-specific model using prior text categories, 3. focusing model using unsupervised text clustering, and 4. the focusing model interpolated with the general model.

documents. The results for the Finnish and English text corpora in Figure 10.5 indicate that the focusing model is superior in terms of perplexity when compared to a general "monolithic" trigram model of the whole data set [4]. The focusing model is, as well, significantly better than the topic category specific models where the correct topic model was chosen based on manual class label on the data. One advantage of unsupervised topic modeling over a topic model based on fixed categories is that the unsupervised model can achieve an arbitrary granularity and a combination of several sub-topics. Finally, the lowest perplexity was obtained by a linear interpolation of word probabilities between the focusing model and the general model.

The first experiments to apply the focusing language models in Finnish large-vocabulary continuous speech recognition are reported in [5]. The results did not show clear improvements over the baseline, but by using a local LM of small but relevant text material, we see, however, that lattice rescoring can decrease the error rate. The preliminary English speech recognition tests indicate as well, that an interpolated model between a huge general LM and a small local LM performs better than the general LM alone. While there are clearly improvements to be made in language modeling, for example, to collect larger amounts of relevant text training data, maybe the most important result of the Finnish speech recognition tests is that the topical focusing works and does not slow down the whole recognition process.

More recently, our studies have been focused on conditions when the interpolation of the local LMs and the general LMs does improve the perplexity and the speech recognition results [6] and on comparisons of different combination algorithms between topic LMs and other LMs [7].

References

[1] V. Siivola, M. Kurimo, and K. Lagus. Large vocabulary statistical language modeling for continuous speech recognition in Finnish. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 737–730, 2001.

[2] R.M. Iyer and M. Ostendorf, "Modeling long distance dependencies in language: Topic mixtures versus dynamic cache model," *IEEE Trans. Speech and Audio Processing*, 7, 1999.

[3] M. Kurimo and K. Lagus. An efficiently focusing large vocabulary language model. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN'02)*, pages 1068–1073, Madrid, Spain, 2002.

[4] K. Lagus and M. Kurimo. Language model adaptation in speech recognition using document maps. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, pages 627–636, Martigny, Switzerland, 2002.

10.4 Emergence of linguistic features using independent component analysis

We have been able to show that Independent Component Analysis (ICA) [2] applied on word context data provides distinct features that reflect syntactic and semantic categories[1]. The analysis gives features or categories that are both explicit and can easily be interpreted by humans. This result can be obtained without any human supervision or tagged corpora that would have some predetermined morphological, syntactic or semantic information. We have also shown that the emergent features match well categories determined by linguists by comparing the ICA results with a tagged corpus [3].

We have also shown that the ICA can be used succesfully for studying the properties of morphemes [5]. We used a large Finnish text corpus in the analysis. As a result we obtained emergent linguistic representations for the morphemes. We have also used the ICA in language modeling. This includes creating an N-gram model for classes derived from ICA features [6].

In the following, we will show several examples of the analysis results from [1]. In considering the feature distributions, it is good to keep in mind that the sign of the features is arbitrary. As was mentioned earlier, this is because of the ambiguity of the sign: one could multiply a component by -1 without affecting the model. Also, the numbering (order) of the components is arbitrary.

Fig. 10.6 shows how the third component is strong in the case of nouns in singular form. A similar pattern was present in all the nouns with three exceptional cases with an additional strong fourth component indicated in Fig. 10.7. The reason appears to be that "psychology" and "neuroscience" share a semantic feature of being a science or a scientific discipline. A similar pattern is also present in words such as "engineering" and "biology". This group of words provide a clear example of distributed representation where, in this case, two components are involved.

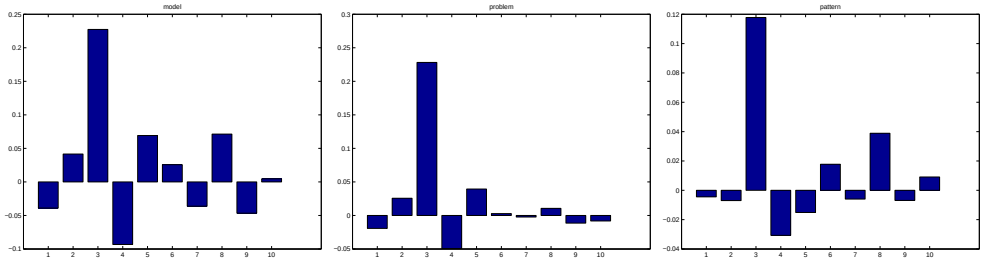

Figure 10.6: ICA features for "model", "problem" and "pattern". For each word, we show the values of the 10 independent components as a bar plot.

An interesting point of comparison for Fig. 10.6 is the collection of plural forms of the same nouns in Fig. 10.8. The third component is strong as with the singular nouns but now there is another strong component, the fifth.

The results include both an emergence of clear distinctive categories or features and a distributed representation. In the emergent representation, a word may thus belong to several categories simultaneously in a graded manner.

We wish that our model provides additional understanding on potential cognitive mechanisms in natural language learning and understanding [4]. Our approach attempts to show that it is possible that much of the linguistic knowledge is emergent in nature and based on specific learning mechanisms.

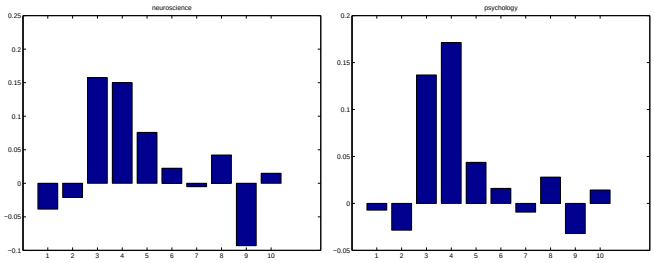

Figure 10.7: ICA features for "neuroscience" and "psychology".

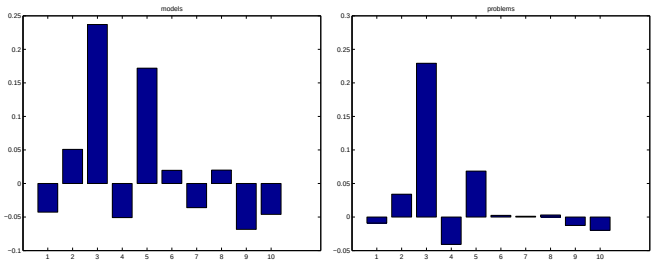

Figure 10.8: ICA features for "models" and "problems".

References

[1] T. Honkela, and A. Hyvärinen. Linguistic Feature Extraction using Independent Component Analysis. In *Proceedings of IJCNN 2004, International Joint Conference on Neural Networks*, Budapest, Hungary, 25-29 July 2004, pp. 279-284.

[2] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[3] J.J. Väyrynen, T. Honkela, and A. Hyvärinen. Independent Component Analysis of Word Contexts and Comparison with Traditional Categories. In: Jarmo M. A. Tanskanen (ed.), Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004, Espoo, Finland, 9-11 June 2004, pp. 300-303.

[4] T. Honkela, A. Hyvärinen, and J. Väyrynen. Emergence of Linguistic Features: Independent Component Analysis of Contexts. In A. Cangelosi, G. Bugmann and R. Borisyuk (eds.), Proceedings of NCPW9, Neural Computation and Psychology Workshop, Plymouth, England, pp. 129-138, 2005.

[5] K. Lagus, M. Creutz, and S. Virpioja. Latent Linguistic Codes for Morphemes using Independent Component Analysis. In A. Cangelosi et al. (eds.), Modeling Language, Cognition and Action, Proceedings of the Ninth Neural Computation and Psychology Workshop, NCPW9, Plymouth, England, pp. 139-144.

[6] Virpioja, S. (2005) New methods for statistical natural language modeling. Master's thesis, Department of Computer Science and Engineering, Helsinki University of Technology.

10.5 SOM-based analysis of words and sentences

Observation of language use provides indirect evidence of the representations that humans utilize. The study of conceptual and cognitive representations that underlie the use of language is important for applications such as speech recognition. By studying large amounts of data it may be possible to induce the conceptual, system-internal representations which provide a grounding for meanings of words.

The self-organizing map [4] can be applied for clustering word forms based on the words that have appeared in their immediate contexts. This has been shown originally for artificially generated sentences in [9] and later for large English text corpora, e.g., in [1]. In Finnish the rich inflectional morphology poses a challenge as the vocabularies built on inflected word forms are typically very large. This problem has successfully been tackled in [6]. In [7], the motivation and methodology of use of the self-organizing maps in conceptual analysis is considered in some detail.

In [5] we were analyzing poems, in particular, Shakespeare's sonnets. Our specific focus was to see what kind information we can find on the "semantic turn" in a sonnet. This is a topic that is related both to the structure of a poem and the meaning of the words used. Our aim was not to present an analysis model that would cover all relevant aspects but to outline one particular approach that can be later extended to cover other points of view.

In [8] we considered the creation of word category maps (cf. e.g. [9, 1]) using ICA-based word features. In earlier studies, a random encoding for each word has been used. Ideally, one could represent each word as a feature vector that would take into account its syntactic and semantic characteristics. This kind of sparse feature representation can be created automatically using independent component analysis (ICA) [3] as we have shown in [2]. In [8], we compared the word category maps both in cases where the random encoding and ICA-based encoding of words were used.

References

[1] T. Honkela. *Self-Organizing Maps in Natural Language Processing.* PhD thesis, Helsinki University of Technology, 1997, Espoo, Finland.

[2] T. Honkela, A. Hyvärinen, and J. Väyrynen. *Emergence of linguistic representation by independent component analysis.* Technical report A72, Helsinki University of Technology, Laboratory of Computer and Information Science, 2003.

[3] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[4] T. Kohonen. *Self-Organizing Maps.* Third, extended edition. Springer, 2001.

[5] O. Kohonen, S. Katajamäki, and T. Honkela. In Search for Volta: Statistical Analysis of Word Patterns in Shakespeare's Sonnets. In: *Proceedings of International Symposium on Adaptive Models of Knowledge, Language and Cognition (AMKLC'05).* Espoo, Finland, June 15-17, 2005. pp. 44-47.

[6] K. Lagus, A. Airola, and M. Creutz. Data analysis of conceptual similarities of Finnish verbs. In *Proceedings of the CogSci 2002, the 24th annual meeting of the Cognitive Science Society*, pages 566–571. Fairfax, Virginia, August 7–10, 2002.

[7] K. Lagus. Miten hermoverkkomallit selittävät kielen oppimista. A.M. Korpijaakko-Huuhka, S. Pekkala and H. Heimo. (eds.) *Kielen ja kognition suhde*. Puheen ja kielen tutkimuksen yhdistyksen julkaisuja 37, 2005.

[8] J.J. Väyrynen and T. Honkela. Word Category Maps based on Emergent Features Created by ICA. In: Heikki Hyötyniemi, Pekka Ala-Siuru and Jouko Seppänen (eds.), Life, Cognition and Systems Sciences, Symposium Proceedings of the 11th Finnish Artificial Intelligence Conference, Finnish Science Center Heureka Vantaa, 1-3 September 2004, pp. 173-185.

[9] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 1989; 61:241-254