Chapter 10

Natural language processing

Krista Lagus, Mikko Kurimo, Timo Honkela, Mathias Creutz, Jaakko J. Väyrynen, Sami Virpioja, Ville Turunen, Matti Varjokallio

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. 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 10.2 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.

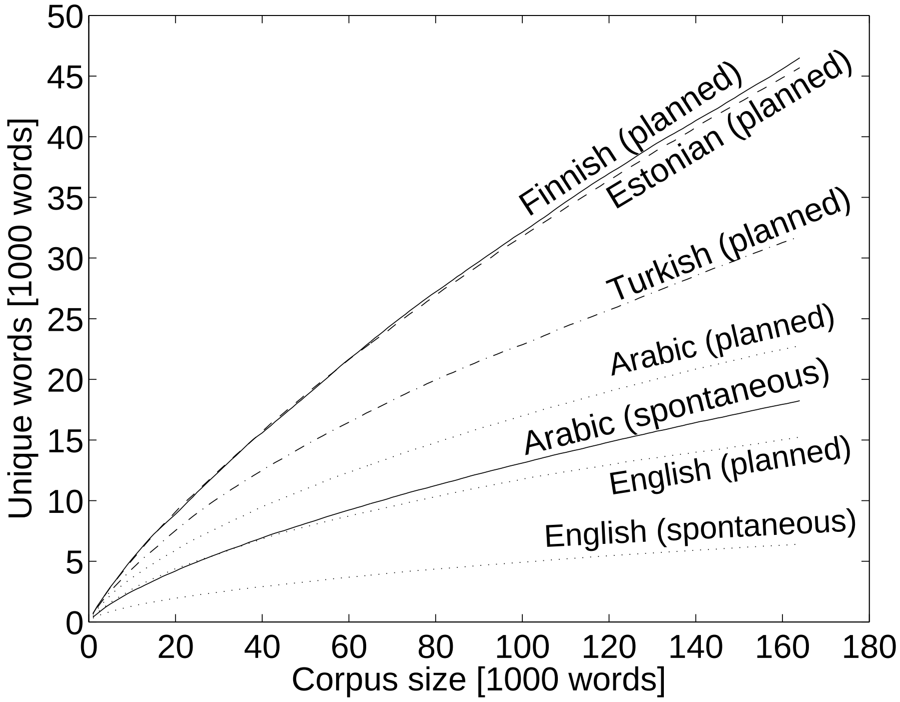

Figure 10.2: The number of different word forms (types) encountered in growing portions of running text (tokens) of various languages.

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].

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. 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" (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, Categories-ML [3] and Categories-MAP [4]. The former is based on Maximum Likelihood (ML) optimization, in combination with some heuristics, whereas the latter applies a more elegant model formulation within the Maximum a Posteriori (MAP) framework. The MAP formulation, along with a thorough comparison to the other Morfessor variants, is provided also in [5] and [10].

Some sample segmentations of Finnish, English, as well as Swedish words, are shown in Figure 10.3. 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 10.3, "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").

Evaluation

In the publications related to the development of Morfessor, the algorithm has been evaluated by comparing the results to linguistic morpheme segmentations of Finnish and English words [1, 2, 3, 4, 5]. In order to carry out the evaluation, linguistic reference segmentations needed to be produced as part of the project, since no available resources were applicable as such. This work resulted in a morphological "gold standard", called *Hutmegs* (Helsinki University of Technology Morphological Evaluation Gold Standard) [11, 12]. 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.

Morfessor algorithms have also been evaluated in the Morpho Challenge competitions described in Section 10.2. Morpho Challenge 2007 included evaluation in four languages (English, Finnish, German and Turkish) and two competitions: comparison against linguistic standards and evaluation in information retrieval tasks. Morfessor managed fairly

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 10.3: 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.*

well in all the evaluations, especially with Finnish and Turkish languages.

Applications

Morfessor has been extensively tested as a component of a large vocabulary speech recognition system. By allowing a compact but flexible vocabulary for the system, Morfessor improves especially recognition of rare words. For several languages such as Finnish, Estonian and Turkish, this approach outperforms the state-of-the-art solutions. The speech recognition experiments are described in Section 8.3.

In addition to speech recognition, Morfessor has been used in speech retrieval and statistical machine translation systems. These experiments are described in Section 8.4 and 13, respectively.

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) and two versions of the Morfessor (Baseline and Categories-ML). Those interested in larger-scale experiments can download the Morfessor program and train models using their own data sets. Two versions are available: Morfessor 1.0 software implements the Morfessor Baseline algorithm described in [13] and Morfessor Categories-MAP 0.9.2 software implements the Morfessor Categories-MAP algorithm described in [4]. During 2007, a monthly average of 10 downloads has been registered for both versions.

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*, Volume 4, Issue 1, Article 3, January 2007.

[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. Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Doctoral thesis, Dissertations in Computer and Information Science, Report D13, Helsinki University of Technology, Espoo, Finland, 2006.

[11] 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.

[12] 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.

[13] 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.

10.2 Morpho Challenge

Morpho Challenge is a series of scientific competition organized by Adaptive Informatics Research Centre for an evaluation of unsupervised morpheme analysis algorithms. The challenge is part of the EU Network of Excellence PASCAL Challenge Program and in 2007 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 two times: the results of the 2005 challenge were published in a workshop in April 2006 in Venice, Italy [1]. The 2007 challenge workshop was held in September 2007 in Budapest, Hungary [2, 3].

In the original challenge, the words were segmented in unsupervised morphemes and the results were evaluated by a comparison to linguistic gold standard morphemes. The organizers also used the results to for training statistical language models and evaluated the models in large vocabulary speech recognition experiments [1]. The 2007 challenge was a more difficult one requiring morpheme analysis of words instead of just segmentations into smaller units. The evaluation of the submissions was performed by two complementary ways: *Competition 1*: The proposed morpheme analyses were compared to a linguistic morpheme analysis gold standard by matching the morpheme sharing word pairs [2]. *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 [3]. 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.

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, Matti Varjokallio. Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard – Morpho Challenge 2007. In *Working Notes of the CLEF 2007 Workshop.* Edited by Alessandro Nardi and Carol Peters. 19-21 September, Budapest, Hungary.

[3] Mikko Kurimo, Mathias Creutz, Ville Turunen. Unsupervised Morpheme Analysis Evaluation by IR experiments – Morpho Challenge 2007. In *Working Notes of the CLEF 2007 Workshop.* Edited by Alessandro Nardi and Carol Peters. 19-21 September, Budapest, Hungary.

10.3 Emergence of linguistic features using independent component analysis

We have been able to show that Independent Component Analysis (ICA) [1] applied on word context data provides distinct features that reflect syntactic and semantic categories [2]. The difference to latent semantic analysis (LSA) is that the analysis finds features or categories that are not only explicit but can also 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.

It is important to compare the capability of single features or feature pairs to separate categories because this measures how well the obtained features correspond with the categories. In fact, when all features are used, the separation capabilities of ICA and LSA are comparable because the total information present is the same. We have also shown that the emergent features match well with categories determined by linguists by comparing the ICA results to linguistic word category information [3].

We have shown how the features found by the ICA method can be further processed by simple nonlinear methods, such as thresholding, that gives rise to a sparse feature representation of words [4, 5]. We performed thresholding for each found word feature vector separately. The values closest to zero were set to zero and only a selected number of features were left to their original values. An analogical approach can be found from the analysis of natural images, where a soft thresholding of sparse coding is a denoising operator.

We compared the original representation and the thresholded representations in multiple choice vocabulary tasks, which measure the semantic information captured by the representation. An illustrative result is shown in Figure 10.4, which compares the feature thresholding with the two methods, latent semantic analysis and independent component analysis. The graph shows that the thresholded ICA representation is able to capture the most important semantics with fewer components, as the quality of the thresholded ICA representation degrades more slowly than both LSA representations. Several tests were run with three languages, including two different corpora, with quite similar results.

We have also shown how independent component analysis gives rise to a multilingual word feature space when trained with a parallel corpus [6]. The feature space created by the found features is also multilingual. Words that are related in different languages appear close to each other in the feature space, which makes it possible to find translations for words between languages. Table 10.1 shows the closest words for the English word 'finland' in the feature space, which include different forms of the Finnish equivalent, but also the name of a neighboring country ('sweden') as well as Austria ('itävalta'). The latter might be caused by shared work during the Finnish EU presidency. The single features also carry multilingual semantic information, as can be seen from Table 10.2, that lists the most prominent words in three features.

The attained 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 see that further processing of the features is possible and thresholding produces a more sparse representation that can have greater interpretability without too much information loss. The method is also applicable to multilingual textual data, and is able to find representations where the multilingual semantic space can be used to mine translations and related words.

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

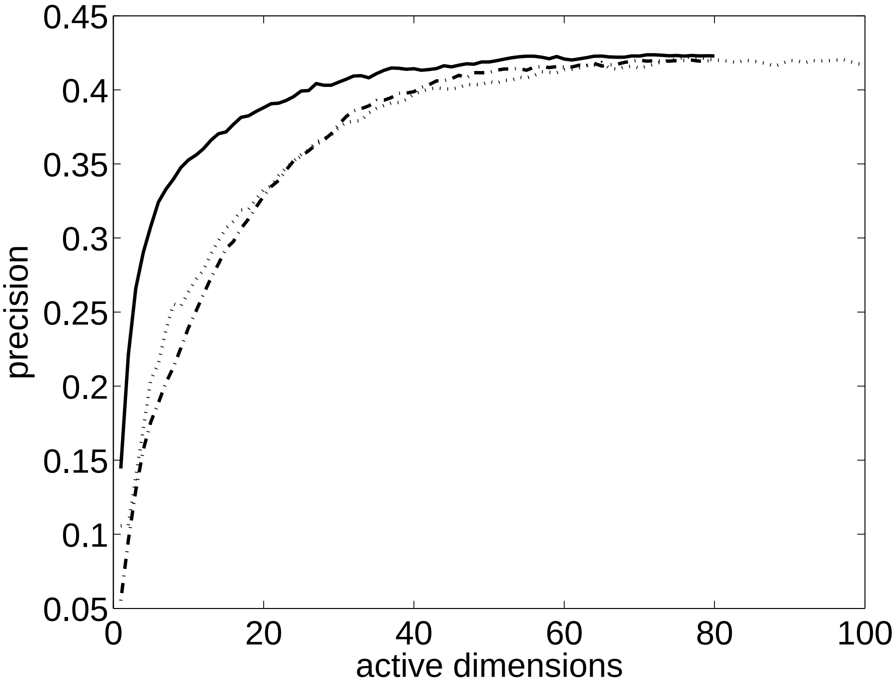

Figure 10.4: Rates of correctly answered questions with unthresholded LSA (dotted), LSA with thresholding with 80 components (dashed) and ICA with thresholding with 80 components (solid) set w.r.t. the number of non-zero features (after thresholding). The features where calculated from free electronic English books extracted from the Gutenberg project. The test questions were based on synonyms and related words extracted from the Moby thesaurus.

Table 10.1: The closest words in the multilingual feature space to the word 'finland'.

word	match
finland	1.00
suomen	0.83
suomi	0.82
sweden	0.79
suomessa	0.77
austria	0.73
...	...

on specific learning mechanisms.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis.* John Wiley & Sons, 2001.

[2] T. Honkela, and A. Hyvärinen. Linguistic feature extraction using independent component analysis. In *Proceedings of IJCNN 2004, International Joint Conference on*

Table 10.2: Most prominent words for three example features (columns) that list clearly related words in both languages.

saksan	values	eroja
ranskan	rauhan	different
germany	demokratian	difference
france	vapauden	välillä
french	democracy	erilaista
german	ihmisoikeuksien	differences
sweden	arvoja	erot
netherlands	solidarity	toisiaan
ranska	peace	disparities
belgian	arvojen	eri
ruotsin	kunnioittaminen	erilaiset
saksa	oikeusvaltion	differ
italian	principles	differing
kingdom	continent	eroavat
...	...	...

Neural Networks, Budapest, Hungary, 25–29 Jul 2004, pp. 279–284.

[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 NORSIG 2004, Sixth Nordic Signal Processing Symposium*, Espoo, Finland, 9–11 Jun 2004, pp. 300–303.

[4] J. J. Väyrynen, L. Lindqvist and T. Honkela. Sparse distributed representations for words with thresholded independent component analysis. In *Proceedings of IJCNN 2007, International Joint Conference on Neural Networks*, Orlando, Florida, 12–17 Aug 2007, pp. 1031–1036.

[5] J. J. Väyrynen, T. Honkela and L. Lindqvist. Towards explicit semantic features using independent component analysis. In: M. Sahlgren and O. Knuttson (eds.), *Proceedings of SCAR 2007 Workshop, Semantic Content Acquisition and Representation*, SICS Technical Report T2007-06, Swedish Institute of Computer Science, Stockholm, Sweden, ISSN 1100-3154, Tartu, Estonia, 24 May 2007, pp. 20–27.

[6] J. J. Väyrynen and T. Lindh-Knuutila. Emergence of multilingual representations by independent component analysis using parallel corpora. In *Proceedings of SCAI 2006, Ninth Scandinavian Conference on Artificial Intelligence*, Espoo, Finland, 25–27 Oct 2006, pp. 101–105.