LATENT LINGUISTIC CODES FOR MORPHEMES USING INDEPENDENT COMPONENT ANALYSIS

KRISTA LAGUS MATHIAS CREUTZ SAMI VIRPIOJA

Neural Networks Research Centre, Helsinki University of Technology,
P.O. Box 5400, 02015 HUT, Finland
krista.lagus@hut.fi

We study properties of morphemes by analyzing their use in a large Finnish text corpus using Independent Component Analysis (ICA). As a result, we obtain emergent linguistic representations for the morphemes. On a coarse level, main syntactic categories are observed. On a more detailed level, the components depict potential thematic roles of the morphemes. An interesting question is whether these discovered lower-dimensional representations could be directly utilized in language processing applications.

1. Introduction

In recent years the use of statistical word-based methods for syntactic and semantic analysis has become increasingly popular. In highly inflecting languages, such as Finnish, a word may have hundreds or even thousands of different inflected forms. Many of the word forms are rare, but they nonetheless make up an important part of the words in a text. Due to the large proportion of rare word forms, statistical analysis methods are of limited use when applied directly to word contexts. Moreover, many cognitively interesting phenomena leave traces *within* words and not only on the sentence level. For example, grammatical relations and to some extent thematic roles (Agent, Patient, Theme, Experiencer, Beneficiary, Location, etc; see e.g., (Saeed, 1997)) are often marked using appropriate inflected forms, not so much using word ordering and function words as in English. We take morphemes as basic units of meaning, and segment every word form into a sequence of morphemes, or more specifically into allomorphs (realization variants). We then apply an unsupervised method called Independent Component Analysis (ICA) (see e.g., Hyvärinen, Karhunen & Oja, 2001) to analyze the contexts of the morphemes to discover latent small-dimensional representations for each morpheme. Ideally each component might code for a single cognitive, syntactic, semantic or phonological property. In a related work (Honkela, Hyvärinen & Väyrynen, 2005), ICA was applied to the analysis of English words based on their averaged contexts.

1.1. *On Finnish Morphology*

Word formation in Finnish mainly takes place using agglutination, that is, the concatenation of morphemes (for a detailed description, see Hakulinen, 1979). A Finnish word typically consists of several morphemes, with an alternation of stems and *suffixes* (and sometimes prefixes):

Finnish word; its segmentation	English translation; literal translation of the segments
kahvinjuojallekin; kahvi + *n* + juo + *ja* + *lle* + *kin*	also for [the] coffee drinker; coffee + of + drink + -er + for + also
tietäisimmeköhän; tietä + *isi* + *mme* + *kö* + *hän*	would we really know?; know + would + we + INTERR + indeed

1.2. *Independent Component Analysis (ICA)*

ICA is a statistical analysis method that attempts to find the latent components (also called sources or underlying factors) that generated a data set. The ICA model (Hyvärinen et al., 2001) assumes that the observations **x** are a mixture of the activities of some unknown, independent sources **s**. In the linear case, $\mathbf{x} = \mathbf{A}\,\mathbf{s}$, where **A** is called the mixing matrix. The problem is then to determine **A** and **s** when only the observations **x** are known. A more familiar method may be Principal Component Analysis (PCA) (or Singular Value Decomposition, SVD) which finds *uncorrelated* sources. In contrast, ICA finds *statistically independent* sources, which generally appear to be more interesting since they often correspond more closely to the original underlying processes.

2. Morpheme Context Data and Experiments

We study properties of morphemes by analyzing their use in a large text corpus. For each morpheme *i* we collect a context vector $\mathbf{x}_i$. The context vectors are then analyzed using ICA.

2.1. *Corpus*

The corpus consists of Finnish newspaper text with 30 million words (tokens) and 1.3 million unique word forms (types). The words were first segmented into morphemes according to the gold-standard segmentation of Hutmegs (Creutz & Lindén, 2004). Hutmegs contains linguistic, semi-automatically produced morpheme segmentations for Finnish (and English) words. On average, a word (token) in the corpus consisted of 2.0 morphemes and the average length of a

morpheme was 3.6 letters. As the data morphemes to be studied we chose the 3759 most frequent morphemes.

2.2. *Formation of morpheme context vectors*

As context features we selected the 506 most common morphemes (including the word boundary) occurring in the context of the 3759 chosen data morphemes. For each data morpheme we then formed a 506-dimensional context vector as follows: For each instance of the data morpheme, if one of the context morphemes appeared immediately after the data morpheme within the same word, the corresponding value of the context vector was increased by one. The logarithm of the values was then taken, and the vectors were normalized to unity. Longer contexts could be studied, but our hypothesis was that already a minimal context may lead to interesting results.

2.3. *Application of ICA*

We applied the FastICA algorithm (Hyvärinen, 1999) to obtain the first 50 independent components. As ICA leaves the sign (positive or negative) of the components undetermined, we determined the signs by examining the data distribution for each component. As a result, a significant property is generally shown as a positive value of the component.

3. Results

When looking at the obtained 50 components, 16 seem particular to verbs (8 to inflected and 8 to derived forms), 12 to inflected forms of nouns, 2 to derived forms of nouns (adjectival and caritive, e.g., kodi+ton; home+less), 5 to locations and persons (with some distinction between internal and external locatives). Four components signify adjective stems, of which one is for comparison. Due to vowel harmony many properties appear in the component list twice: once for back vowels and a second time for front vowels. In addition, there are components that are hard to interpret, or are not very specific (e.g. 14 and 30). Figure 1 depicts activations of three morphemes that are country names, along with their interpretations. In Figure 2, three frequent, concrete nouns are displayed.

4. Discussion

Most components appear to signify some interesting property. These include syntactic categories (verb, noun), thematic roles (Location, Beneficiary) cognitive or semantic properties (agent vs. action role), or morpho-phonological properties

(e.g., the use of back or front vowels) of the word segments. Typically a morpheme has only a few active components, thus creating *a sparse coding* for

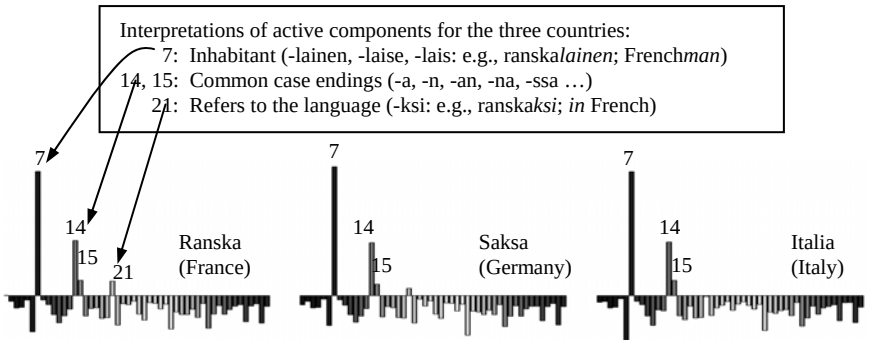

Figure 1: Activations of latent components of three morphemes that are country names: Ranska (France), Saksa (Germany) and Italia (Italy).

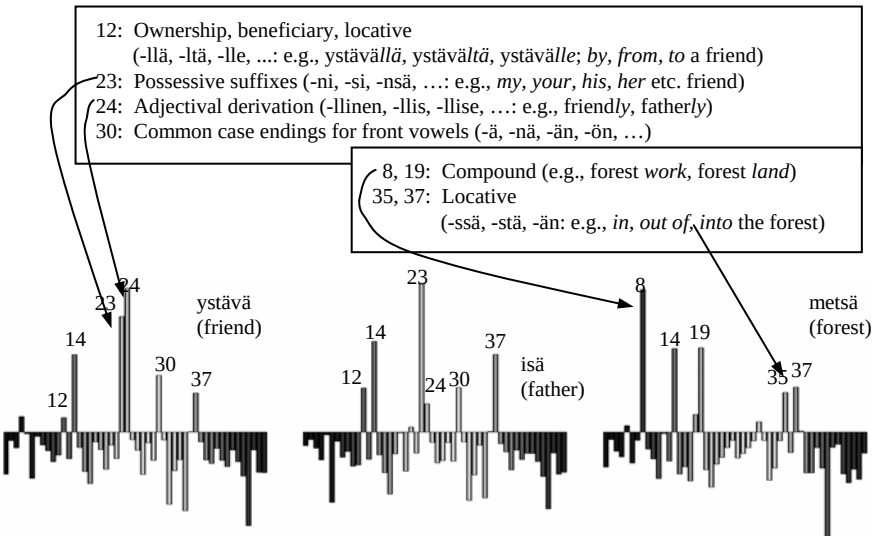

Figure 2: The ICA components for three sample morphemes, ystävä (friend), isä (father) and metsä (forest), along with some interpretations. Component 12 appears to code change and maintenance of the state of ownership and 23 the owner. 24 codes for adjectival role; note that for "friend" this component is strong (friendly), for "father" weaker (fatherly) and for "forest" nonexistent (*forestly).

the morphemes. An ambiguous morpheme is likely to have a higher number of active components, each reflecting one of the morpheme's central properties or uses. On a coarse level, main syntactic categories are observed. On a more detailed level, the components depict potential thematic roles of the morpheme. This result is interesting since in most cases there is no single context feature that

would distinguish such roles, and since many of the context features are themselves ambiguous.

In this experiment we used a linguistic morpheme segmentation for ease of interpretation. It would also be interesting to study what representations ICA discovers for word segments that have been learned in an entirely unsupervised manner, e.g. using the method (Creutz & Lagus, 2004). Questions for future research include discovering suitable context length and optimal number of ICA components. Based on visual inspection, in this case 20 components seemed too few, while 50 was not too many. However, an evaluation method is necessary (such as performance on the TOEFL test of English as a Foreign Language).

The creation of numerical codes for morphemes opens the possibility of using these codes to learn typical patterns of words and to produce new words based on the learned patterns, while operating in a relatively low-dimensional vector space. It is an interesting question whether such codes might be useful for solving a signiticant problem in statistical language modeling, namely how to estimate characteristics (e.g., probability) of future utterances based on observed utterances. Moreover, one may ask whether the component values could be directly utilized by language understanding applications for decoding the meaning of an utterance.

References

Creutz M., Lagus, K. (2004) Induction of a Simple Morphology for Highly-Inflecting Languages. In *Proceedings of the 7th Meeting of the ACL Interest Group in Computational Phonology (SIGPHON), Barcelona* (pp. 43-51).

Creutz M., Lindén K. (2004) Morpheme Segmentation Gold Standards for Finnish and English. Helsinki University of Technology, Publications in Computer and Information Science, Report A77.

Hakulinen L. (1979) *Suomen kielen rakenne ja kehitys* (The structure and development of the Finnish language). Kustannus-Oy Otava, 4th ed.

Honkela T., Hyvärinen A., Väyrynen J. (2004) Emergence of Linguistic Features: Independent Component Analysis of Context (in this volume).

Hyvärinen A. (1999) Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Transactions on Neural Network*, 10(3):626-634.

Hyvärinen A., Karhunen J., Oja E. (2001) *Independent Component Analysis*. John Wiley & Sons.

Saeed J. I. (1997) *Semantics*. Oxford: Blackwell Publishers.