Emergence of multilingual representations by independent component analysis using parallel corpora

Jaakko J. Väyrynen[⋆]
[⋆]Adaptive Informatics Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 TKK, FINLAND
Jaakko.J.Vayrynen@TKK.Fi

Tiina Lindh-Knuutila[†]
[†]Adaptive Informatics Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 TKK, FINLAND
Tiina.Lindh-Knuutila@TKK.Fi

Abstract

This paper reports the first results on extracting a meaningful representation for words from multilingual parallel corpora. Independent component analysis is used to extract a number of components from statistics calculated for words in contexts. Individual components are meaningful and multilingual and words are represented as a bag of concepts model. The component space created by the extracted components is also multilingual. Words that are related in different languages appear close to each other in the component space, which makes it possible to find translations for words between languages.

1 Introduction

Analysis of symbolic text using statistical methods requires extraction of a numerical representation or statistics. Contextual information has been widely used in statistical analysis of natural language corpora (Deerwester et al., 1990; Honkela et al., 1995; Ritter and Kohonen, 1989). One useful representation for words can be constructed by taking into account the contexts in which they occur. This is based on the distributional hypothesis, which states that words that occur in similar contexts tend to be similar. The length of the context may vary from just neighboring words to the sentence and to the whole document, and thus representing different similarities between words. In this paper, we have contexts that span aligned sentences in different languages and word co-occurrences are calculated inside these contexts. Another approach to produce statistics is to count how many times each word occurs in each document.

We follow here the ideas that knowledge of language emerges from language use and knowledge of language is conceptualization. Knowledge of language includes, for instance, grammar, semantics and similarity of words. As an example of language use, we take a text corpus that represents actual language use and calculate statistics that we try to refine into a compact and meaningful representation.

For the statistics, we study frequencies of words in contexts. Independent component analysis (ICA) is used in the extraction of a representation for words from the collected statistics. ICA is an unsupervised method that finds a feature representation for each word, where the values of the features are continuous and several features can be present in one word at the same time. Unsupervised methods have the advantage that they do not need an explicit training set where a correct representation is known. On the other hand, the representations found by unsupervised methods are emergent, i.e., they follow the data instead of rules used by humans. Sahlgren and Karlgren (2005) applied a method called random indexing to find translations of words from a bilingual parallel corpus. The approach there was to use aligned paragraphs as the contexts and to assign a random vector to each context. The relations of word vectors were examined to find translations of words. Our aim is to find also a meaningful representation that encodes linguistic knowledge.

We hope that independent component analysis can find an emerging structure for the data, providing a meaningful representation for the words. We expect the word representations to have related words closer to each other than unrelated words, and that the emergent features to be meaningful. Especially, we want to see translations of words.

The use of multilingual parallel corpora gives a possibility to analyze how meaningful the features found by ICA are. In order to measure the success of our method and to show its potential, we demonstrate its use using a parallel corpus that has the same text in different languages and in which regions of texts are matched together. The initial experiments were carried out using bilingual material.

2 Data and methods

In this section, we describe our data, a sentence-aligned parallel corpus of Finnish-English text, collection of contextual information, and our method for analyzing words in contexts using independent component analysis. We also discuss our methods for the analysis of the results.

2.1 Data

We selected the English-Finnish pair from the sentence-aligned Europarl parallel corpus of the European Parliament proceedings (Koehn, 2005) for our initial experiments. The texts were preprocessed by replacing all characters with their lowercase versions, numbers with a special symbol and by removing non-word codes and punctuation marks. In the corpus, there were $602,153$ aligned regions, where there were 25 tokens per regions in average for English and 20 for Finnish. There were a total of $15,215,650$ tokens (words in running text) and $50,111$ types (unique word forms) in the English text and $10,437,070$ tokens and $350,972$ different types in the Finnish text.

We began by concatenating the aligned regions for English and Finnish together. This defined the contexts according to the bag of words model, where the word order is not considered. For both languages, we selected $10,000$ most frequent types to be analyzed in the contexts of $1,000$ most frequent types in both languages. After removal of duplicate types, we collected the frequencies of co-occurrences of each word j with the context words in the concatenated regions as a vector $\mathbf{x}_j$, which formed columns of a context-word co-occurrence frequency matrix $\mathbf{X}$ of size $1,983 \times 19,758$. The logarithm of the elements of the matrix $\mathbf{X}$ added by one was taken to dampen the effect of high frequency differences.

The method above differs from the approaches in Ritter and Kohonen (1989) and Honkela et al. (1995) because of the nature of the parallel corpus. For instance, Honkela et al. (1995) calculated contexts for two adjacent words in English text. For sentence aligned corpus, however, there is no match between tokens in across languages inside the aligned regions. Naturally, a more refined alignment would make this possible. It seems reasonable to assume that the use of longer contexts contains information more on the semantic level than the syntactic level, whereas the syntactic level information could be gathered using shorter contexts.

2.2 Independent component analysis

In the classic version of the linear, instantaneous ICA model (Comon, 1994; Hyvärinen et al., 2001), each observed random variable $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ is represented as a weighted sum of independent random variables $\mathbf{s} = (s_1, \ldots, s_k, \ldots, s_n)^T$

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where mixing matrix $\mathbf{A}$ contains the weights which are assumed to be different for each observed variable. Both the mixing matrix $\mathbf{A}$ and the independent components (features) $\mathbf{s}$ are learned in an unsupervised manner from the observed data $\mathbf{x}$. It should be noted that the ICA model leaves the number of components, the order of the components and scale and sign of the components ambiguous.

For our analysis of the observed context-word frequencies, we applied FastICA (Hyvärinen, 1999) algorithm with the standard maximum-likelihood estimation by setting the nonlinearity to the $\tanh$ function, and using symmetric orthogonalization. The dimension of the data was first reduced to 100 by principal component analysis (PCA) in order to decorrelate the data, to reduce overlearning and to get a square mixing matrix $\mathbf{A}$. From the whitened data, i.e., after PCA and normalization of variances, 100 components were extracted with ICA.

The resulting components $\mathbf{s} = (s_1, \ldots, s_n)^T$ create a feature representation in the component space, which makes it possible to analyze how the positions of the words in the two languages are related. Moreover, each component can be analyzed separately by considering which words are most prominent, i.e., have the highest (absolute) value of that component. It seems that the estimated components s_k are skewed in nature. The sign ambiguity of Eq. 1 allows us to have only positively skewed components. Now we need only to consider high positive values, which makes the analysis of the components more simple.

The word vectors $\mathbf{s}_j$ are produced as a resulta of the independent component analysis of the sum of contexts $\mathbf{x}_j$ of the word j. The word vectors $\mathbf{s}_j$ can be analyzed as a bilingual lexicon for the words in the two languages, where each component s_k encodes some interesting feature. This results in a two-fold analysis that is discussed in more detail in the following.

2.3 Comparing word similarities

We can consider the whole component space $\mathbf{s} = (s_1, \ldots, s_k, \ldots, s_n)^T$ and see which words are close to each other. An ideal model would be that translation of a word would be found as the closest point in the space. In addition to that, related words would be closer to each other than unrelated words. This is what latent semantic analysis (LSA) does (Deerwester et al., 1990). We measure similarity between the query word vector $\mathbf{s}_1$ and any other word vector $\mathbf{s}_2$ as the cosine of the angles between the vectors

$$d_{\cos}(\mathbf{s}_1, \mathbf{s}_2) = \frac{\mathbf{s}_1^T \mathbf{s}_2}{\|\mathbf{s}_1\| \|\mathbf{s}_2\|} \qquad (2)$$

where the value $d_{\cos}$ is in the range $[-1, 1]$. The highest value $+1$ is obtained for the query word vector itself.

2.4 Examination of the components

Each component s_k can be thought as a semantically interesting feature, which has related words as the most prominent words and for the rest of the words the component has a value near zero, i.e., we consider the values s_k given for each word and list the corresponding words in descending order.

3 Experiments and results

Preliminary experiments using a bilingual parallel corpus are discussed here. Contextual data was collected into data matrix $\mathbf{X}$ and 100 components were extracted with the FastICA algorithm as reported above. The emergent components define a bilingual lexicon and each component can be analyzed separately.

We analyzed how well the found component space brings translations of words together. Table 1 shows the closest words for the Finnish word ‘suomi’ (left) and its English translation ‘finland’ (right) according to Eq. 2. The closest match is an exact translation of the query word. Considering all the closest matches, there is a clear pattern of country names in both languages: They are clearly related to the word we are comparing it to. It should be mentioned, that the word ‘suomi’ is in fact ambiguous having the meaning of both the name of the country (written with a capital S) and the name of the language. One can see that the English word meaning Finnish language is third on the list.

We repeated the experiment for 14 EU member states names both in English and in Finnish from the time of the recording of the proceedings (years 1996–2003). United Kingdom was excluded because it consists of two separate words. The results were similar to those described above. For member states in English, the corresponding word in some inflected form in Finnish was the closest word form, except for one case, where a possible translation was the second closest word form. The problem seemed to be more difficult to the other direction. Now five of the translations were found from the second closest word form and one translation was the sixth closest word form.

Table 1: Closest words in the ICA space to ‘suomi’ (left) and ‘finland’ (right)

suomi	$d_{\cos}$	finland	$d_{\cos}$
suomi	1.00	finland	1.00
finland	0.82	suomen	0.83
itävalta	0.76	suomi	0.82
finnish	0.74	sweden	0.79
britannia	0.74	suomessa	0.77
saksa	0.73	austria	0.73
...		...	

As an additional short analysis of how well our system performs, we took the 30 most frequent nouns from both the English and the Finnish corpus and checked whether our method could find a corresponding translation for them. As most common words in other word classes often correspond to function words and do not necessarily have direct one-word translation, they were left out of this short analysis. No stemming of the words was performed: The query words included inflective forms as well and a translated form was considered a match regardless of its inflection. The translations were checked manually using NetMot electronic dictionary from Kielikone Ltd.

We checked whether a corresponding term in the other language was found within four closest matching words (that could be in either language). Precision, $precision = C/S$, where C is the number of correct translations and S is the size of the source vocabulary, was used as an evaluation metric.

When Finnish query words were used, the translation precision was 0.67 when only the closest word was examined. The precision rose to 0.77, if two closest words were taken into account and to 0.83 when three or four closest words were considered. Finnish is a more compounding language than English, which is why there were Finnish query words that would be translated into two words in English. An example of this is ‘jäsenvaltiot’ for which the appropriate translation is ‘member states’. Our method could find parts of collocations like this, but in this experiment it was considered a failure. If partial matches like this were taken into account, the precision was 0.96 when four closest words were considered.

For English query words, the precision was 0.53 for the closest word, 0.70 when using the two closest and 0.73 and 0.77 with three and four closest matches, respectively. If partial matches were considered as described above, the precision rose to 0.83 when four closest words were examined.

Compared to the precisions reported in Sahlgren and Karlgren (2005), our initial results with high-frequency words are a little worse than theirs, but on the same level as their reported precision over words that occur more than 100 times. It should be mentioned that Sahlgren and Karlgren (2005) utilized lemmatization and the language pairs, context type and alignment level differed from the ones used here.

We can also study the individual components s_k obtained as they are interesting themselves. As an example we present Table 2 showing the most prominent words for selected three components. It can be seen that the components list words in both languages related to 1) countries, languages and nationalities, 2) values, and 3) differences. The relations between the listed words include, for instance, inflected word forms: ‘arvoja’ (plural partitive of ‘value’), ‘arvojen’ (plural genitive of ‘value’); words used together or close to each other: ‘eroja’ (plural partitive of ‘difference’), ‘välillä’ (‘between’); translations: ‘france’, ‘ranska’; or otherwise closely related words: ‘france’, ‘french’.

Table 2: Most prominent words for three example components (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
...	...	...

The small number of extracted components is naturally not able to produce a unique concept for everything. Instead, the components model the concepts using a sparse bag of concepts model, where only a few of the features are active for each concept.

4 Discussion

Our initial experiments covered only a bilingual case, but the method is directly applicable to a multilingual case if a parallel multilingual corpus is available. Further research will include experiments with the multilingual material using more than two languages as well as a more thorough analysis of the results.

In our experiments, we showed briefly that the analysis of the components produced by ICA can be interesting. As mentioned earlier, using shorter contexts, one is able to obtain ICA-based features that are more syntactic in nature, whereas more semantic features can be found when the context is sufficiently large.

In addition to simply finding the closest matches based on the vector representation, the component analysis can be seen as a tool to dig a little deeper in the level of meaning of words: What do the words (or their underlying meanings) have in common? The use of multilingual corpora gives us a possibility to analyze whether the individual components are meaningful themselves. For instance, we can see what kind of components are active for words in different languages with the same or similar meaning.

Gärdenfors (2000) presents the conceptual spaces model for modeling conceptual representations in a cognitive framework. In that model, the concepts are seen as areas in a multi-dimensional conceptual space built using different quality dimensions (of which the simplest examples could be weight, width, height, or temperature). Our viewpoint is that we are able to obtain representations using the emergent ICA components that are in accord with the conceptual spaces model. Undeterministic statistical analysis of textual data could provide features that align with human perception of the world. Our research focus will be in deeper analysis and understanding of the emergent features.

5 Conclusions

Contextual information for words was analyzed by independent component analysis to produce a component space for words. Our hypothesis was that the obtained components create a meaningful representation instead of a latent representation produced by, for instance, random indexing. We tested our assumption with a bilingual parallel corpus, which enabled us to examine both the found component space and the individual components.

A preliminary experiment utilized a sentence aligned parallel corpora of English-Finnish proceedings of the European parliament. The components obtained form a space in which related words in both languages appear closer to each other than unrelated words. This makes it possible to find translation candidates for words by comparing the similarities of the word vectors in the component space. Moreover, the components clearly encode semantic features using both languages.

We conclude, that the results reported in this paper support our assumption that independent component analysis is able to find components that are meaningful and a good feature representation for words.

References

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of American Society of Information Science*, 41(6):391–407, 1990.

Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought.* MIT Press, 2000.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. Contextual relations of words in Grimm tales analyzed by self-organizing map. In *Proceedings of ICANN-95*, pages 3–7, 1995.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, 2005.

Helge Ritter and Teuvo Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61(4):241–254, 1989.

Magnus Sahlgren and Jussi Karlgren. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering, Special Issue on Parallel Texts*, 11(3): 327–341, 2005.