Chapter 13

Natural language modeling

Krista Lagus and Mikko Kurimo

13.1 Semantic analysis of Finnish words and sentences

Krista Lagus

The study of lexical organization, i.e. the study of how word-related information may be efficiently represented is important for many natural language applications, in particular, for smoothing language models. Furthermore, analysis of words as they are used in large corpora may reveal insight on the semantic properties of words that is relevant to linguistic research as well.

As shown in [1,6], the self-organizing map (Chapter 9) can be applied for clustering English word forms based on the words that have appeared in their immediate contexts. I Finnish, however, the rich inflectional morphology poses a challenge as the vocabularies built on inflected word forms are typically very large. Moreover, also the inflections, some of which correspond to prepositions and function words in Englis, carry relevant semantic information [5]. Furthermore, the much less restricted word order compared to English is likely to cause more variation in the immediately nearby words.

In this project, we have applied the SOM algorithm to visualize and cluster common Finnish verbs based on averaged morphosyntactic features in the verb contexts (see [2,3]). Verbs were selected for study for two reasons: there exists a semantic reference classification of Finnish verbs [5] for comparing the results, and the semantic representation of verbs is considered an interesting problem in linguistics [4,5].

The corpus consisted of 17 million words of Finnish magazines and newspaper articles. In the first experiment, 25 verbs and their inflected forms were examined. In the second experiment, the 600 most frequent Finnish verbs were organized on a map, shown in Figure 13.1.

The results in both experiments show that even the simple features used, namely a set of morphological properties of the nearby words, collected over a large number of instances, were found to be suitable for obtaining automatically a semantic clustering and organization of verbs. When compared to a reference classification of Finnish verbs [5], the obtained clustering shows a somewhat different perspective or world view than Pajunen's. In particular, the organization of verbs on the map reflects the importance of cultural, social, and emotional dimensions in lexical organization.

Future research includes the study of verbs based on various other features (e.g. nouns in the context of the verb), and the organization of words from other syntactic categories.

References

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

[2] K. Lagus and A. Airola. Analysis of functional similarities of Finnish verbs using the self-organizing map. In *ESSLLI'01 Workshop on The Acquisition and Representation of Word Meaning,* August 2001.

[3] K. Lagus. Studying similarities in term usage with self-organizing maps. In *Proceedings of NordTerm'01,* June 2001.

[4] B. Levin. *English Verb Classes and Alternations: a Preliminary Investigation.* The University of Chicago Press, Chicago and London, 1993.

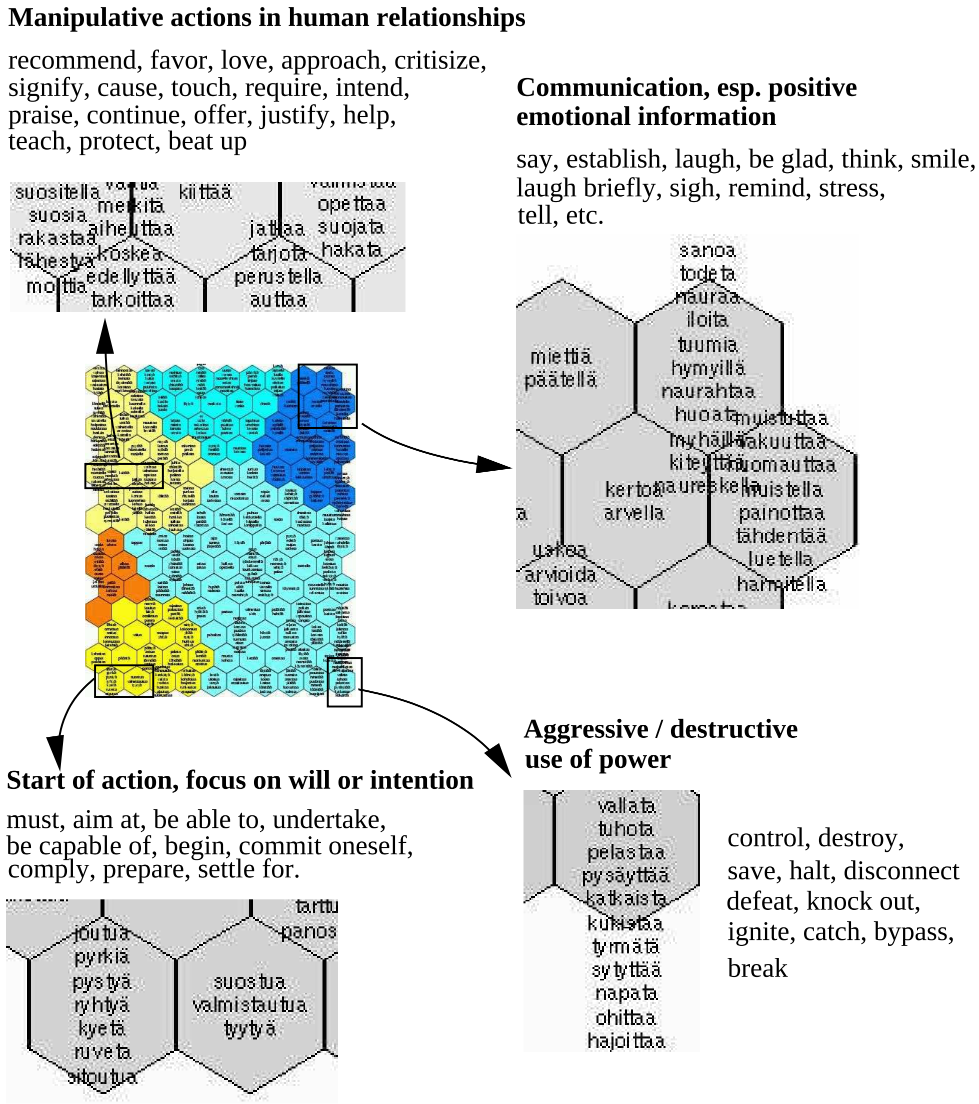

Figure 13.1: A map of the 600 most frequent verbs (base forms) in the Newspaper corpus. The verbs were organized on the basis of the distribution of morphological features in one preceding and two succeeding words, collected over all instances of the verb in any inflected form. The contents of four sample map regions are shown in the insets. In the reference classification (pp. 157–165 in [5], many of the verbs e.g. in the lower right corner indicating 'destructive use of power' are further divided into two specific categories, namely (1) break verbs (*tuhota* 'destroy', *katkaista* 'break', *hajoittaa* 'break down') and (2) fight verbs (*pysäyttää* 'stop', *kukistaa* 'defeat', *tyrmätä* 'knock out'). Similar categories can be found in [4] for English verbs.

[5] A. Pajunen. *Argumenttirakenne. Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä.* Suomalaisen Kirjallisuuden Seura, 2001.

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

13.2 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 14.2) 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 [3] 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 [1] 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.

It has been previously shown that good topically organized clustering of large text collections can be achieved efficiently using the WEBSOM method (see Section 10 or [2]). 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 13.2.

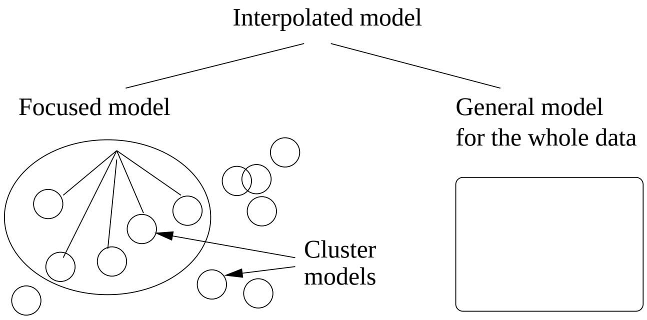

Figure 13.2: *A focusing language model obtained as an interpolation between topical cluster models and a general model.*

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.

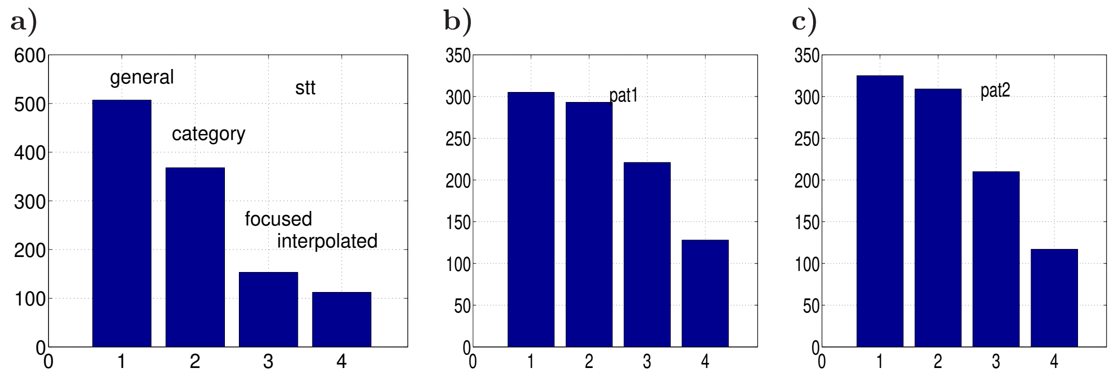

Figure 13.3: The perplexities of the different language models, **a)** for the Finnish STT news corpus, **b)** for smaller patent corpus and **c)** for larger patent corpus. The explanation of the bars in each figure, from left to right: 1. general model for the whole corpus, 2. category-specific model 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 13.3 indicate that the focusing model is superior in terms of perplexity when compared to a general "monolithic" trigram model of the whole data set. 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 next step in this project, for which the experiments are currently being performed, is to examine how well the obtained improvements in modeling translate to advancing speech recognition accuracy.

References

[1] 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.

[2] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks*, vol. 11, number 3, pp. 574–585. May 2000.

[3] 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*, volume 1, pages 737–730, 2001.