Unsupervised Word Categorization Using Self-Organizing Maps and Automatically Extracted Morphs

Mikaela Klami and Krista Lagus

Adaptive Informatics Research Centre, Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, Finland
mikaela.klami@hut.fi

Abstract. Automatic creation of syntactic and semantic word categorizations is a challenging problem for highly inflecting languages due to excessive data sparsity. Moreover, the study of colloquial language resources requires the utilization of fully corpus-based tools. We present a completely automated approach for producing word categorizations for morphologically rich languages. Self-Organizing Map (SOM) is utilized for clustering words based on the morphological properties of the context words. These properties are extracted using an automated morphological segmentation algorithm called Morfessor. Our experiments on a colloquial Finnish corpus of stories told by young children show that utilizing unsupervised morphs as features leads to clearly improved clusterings when compared to the use of whole context words as features.

1 Introduction

Gathering lexical information based on authentic word usage has become a reasonable avenue due to the availability of vast language resources. Detailed syntactic and semantic information on words is valuable for a wide variety of fundamental natural language processing tasks, including information retrieval, question answering, and machine translation.

One way of capturing information regarding the syntactic or semantic relatedness of words is through obtaining a classification or a cluster structure of them. Over the years, many researchers have examined the use of statistical, corpus-based methods for clustering words [1,2,3,4,5].

In the task of word clustering, first a set of informative features representing the words is determined, and for each word, the values for the features are recorded. Then, a clustering method is used for grouping the words based on their feature vector values. Typically, the features used are individual words occurring in the immediate context of the words being clustered [1,2,5] or having some other grammatical relationship with the words [6,7]. This is a feasible approach for languages like English with manageable amounts of word inflection. But even there, data sparsity is a problem: many samples of a word form are needed in order to obtain a reliable feature representation for it.

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 912–919, 2006.

For a language like e.g. Finnish which relies heavily on inflection and other morphological processes, the data sparsity problem is yet intensified due to the much larger number of possible context word forms. Finnish word forms typically contain masses of potentially valuable semantic information inside them. As counting the occurrences of individual words is generally an infeasible strategy for such morphologically rich languages, one solution is to use features that utilize the inflectional or derivational properties within words [5,8]. However, suitable morphological analyzers do not exist for all languages, and the manually designed analyzers also typically fail to cope with e.g. language change or the colloquial language found in everyday conversations. Fortunately, in the recent years, tools that discover a rudimentary morphological segmentation automatically from text corpora have emerged, e.g. Linguistica [9] and Morfessor [10].

The purpose of this study was to find out whether the use of completely automatically discovered morphological segmentations can improve the quality of word categorizations for highly inflecting languages. We also selected an unusual and challenging corpus for our task, a collection of stories told by young Finnish children. The data set was chosen in the purpose of studying the language use and language acquisition process of children of different ages, but only the technical results of our work will be presented here due to space limitations.

In our experiments, we employ an unsupervised morphological segmentation algorithm called Morfessor [10] for extracting morphological features for the words that are to be categorized. We then use the morph features in training word category maps using the Self-Organizing Map (SOM) algorithm [11]. We evaluate and compare the utilization of different feature sets, with both whole context word features and sets of Morfessor-extracted morphs. Our experiments show that the use of automatically extracted morph features instead of whole words leads to improved quality of the resulting word category maps. We expect the presented results also to be of interest for solving many other lexical acquisition tasks than word categorization, and there is reason to believe that similar benefits will be obtained for also other languages with rich morphology.

2 Segmenting Words into Morphs Using Morfessor

Morfessor [10,12] is an automated learning algorithm for extracting the morphology of a language from text corpora. Morfessor is able to segment the words of an unlabeled text corpus into morpheme[1]-like units (morphs), and it is especially applicable to highly inflecting, morphologically rich languages like Finnish, Spanish or Turkish. The Morfessor algorithm not only seeks to find the most accurate segmentation possible, but it also learns a representation of the language from the data it was applied to, namely an inventory of the morphs of the language. In this work, we chose to apply the Categories-ML variant of Morfessor, which uses a Hidden Markov Model (HMM) to model morph sequences and estimates the model parameters by maximizing the likelihood of the data. The

[1] Morphemes can be defined as parts of words that constitute the smallest meaningful units in the grammar of a language.

algorithm is based on the Expectation Maximization (EM) principle, and the Viterbi algorithm is used for optimizing the model.

The output of the Morfessor algorithm is a lexicon of the words from the corpus, segmented at the proposed morpheme boundaries into morphs. The more recent versions of Morfessor also label the resulting morphs as either 'STM' (word root), 'PRE' (prefix) or 'SUF' (suffix). For example, the verb form "aivastivat" ('they sneezed') is correctly segmented as "aivast/STM + i/SUF + vat/SUF".

Morfessor has been tested on Finnish and English text corpora with good results [12], making it an able tool for producing the morphological segmentation used in this work. Previously it has been shown that the use of Morfessor-extracted morphs was able to improve Finnish large vocabulary speech recognition accuracy considerably [13].

3 Self-Organizing Maps

The Self-Organizing Map (SOM) [11,14] is an unsupervised neural network algorithm that is able to organize multidimensional data sets into a two-dimensional, ordered map grid. The data samples are then projected on the map so that their relative distances reflect their similarity according to the chosen feature set, i.e. similar input samples will generally be found close to each other on the map.

The map grid of a SOM consists of nodes, each of which corresponds to a prototype vector. Together, the prototype vectors form an approximation of the data set. During the training process of the SOM, sample vectors (or feature vectors) are compared to the prototype vectors, and the samples are mapped according to their Best Matching Unit (BMU) on the map grid. The algorithm thus simultaneously obtains a clustering of the data based on the prototype vectors, and a nonlinear projection of the input data from the multidimensional input space onto the two-dimensional ordered map.

3.1 SOMs in Word Categorization

Word category SOMs are word maps trained on sample word forms from an input text corpus [1,2]. The general idea is to have implicit word categorizations emerge automatically from the input data itself. The similarity or dissimilarity of word forms is usually based on their textual contexts in the corpus, i.e. word forms that have similar elements in their contexts should appear close to each other on the resulting word category SOM.

A set of training words for training a word SOM is usually chosen based on a word form frequency list of the corpus, and a feature vector for each training word is then calculated from the corpus. Traditionally, the feature sets of word SOMs have consisted of the occurrences of whole corpus words in the contexts of the training word samples. The size of the context window of the sample words may vary; in the experiments of this paper, the context window size was fixed at 1, meaning that only the two words that were immediately before and after the training word were considered as its context.

In this work, a novel approach to word SOM feature sets is adopted. Instead of using the context word forms for features as such, they are segmented into morphs by using the Morfessor algorithm, and the context of a training word is now checked for the presence of feature morphs rather than for whole, unsegmented words. Also Lagus et al. [5] utilized morphological features in a study on categorizing Finnish verbs using SOM, but instead of automatically extracted word segments, they used as features only 21 different morphemes obtained with a linguistic morphological analyzer. However, such an approach would be inapplicable here due to the particular colloquial nature of our corpus.

4 Data Set and Feature Extraction

The data set for our experiments consisted of 2642 stories or fairytales in Finnish, told by children aged from 1 to 14. In total, the stories form a Finnish text corpus of 198 036 word forms (after preprocessing).

The stories were collected using a method called *Storycrafting* [15]: they were told orally and transcribed by an adult exactly as they were heard, without correcting any potential mistakes or colloquialisms by the child. This gives the corpus a particular nature as the conversational and authentic use of language of young Finnish children. Because of this challenging, often non-orthographical nature of the data, statistical, completely automated learning methods were assumed to prove especially useful in analyzing it.

The stories were preprocessed, which included stripping metadata, removal of punctuation, changing numbers to the special symbol NUM, and changing uppercase letters to lower-case. Finally, a morphological segmentation of the corpus was obtained using the Categories-ML variant of the Morfessor algorithm.

5 Experiments on SOM Feature Sets

The objective of these experiments was to find out the best way of producing word categorizations with the Self-Organizing Map. The experiments on comparing word SOMs with different types of features are described in the following.

5.1 Evaluation Measure

For evaluating the quality of word SOMs, an automatical evaluation measure was needed. The evaluation measure developed for this work is based on using the part-of-speech (POS) information of the 200 most frequent word forms in the corpus. Each word form was manually assigned a list of all its possible POS tags, according to a recent descriptive book of Finnish grammar [16].

The idea of the evaluation measure is to find out how tightly word forms are clustered on a particular word SOM according to their POS classifications. For each word form on a SOM, a percentage is calculated which tells the portion of the words in the same or the immediately neighboring map node having at least

one POS in common with the word form under examination. An average percentage for the whole word SOM is then calculated over these results of individual word forms. More precisely, the evaluation measure for a single word SOM can be written as $1/N \sum_{i=1}^{N} A_i/B_i$, where N is the number of word forms projected on the SOM, B_i denotes the number of words having their Best Matching Unit (BMU) in the neighborhood of the BMU of the ith word, and A_i is the number of words in the neighborhood sharing at least one POS with the ith word.

Finally, in order to rule out the possibility of chance, the final results for each type of word SOM were calculated over the individual results of 100 randomly initialized word SOMs of that type, yielding the final quality score for the word SOM type. This quality score can be seen as a kind of a POS cluster density score for the emergent word clusters of the particular SOM variant.

5.2 Feature Sets for Different Experiments

We evaluated and compared word SOM variants with different numbers and types of features. The feature sets used will be described below in more detail.

Morph vs. whole context word features. In order to see whether using morph features can improve the quality of a word SOM, two maps were trained: one with the 200 most frequent corpus word forms as features, and one with the 200 most frequent Morfessor-extracted morphs as features (with all types of morphs).

Different types of morph features. In addition to the SOM experiments with morph features in general, a few variants with different types of morph features were trained. Here, two morph type experiments are presented, namely one word SOM with the 200 most frequent root morphs as features, and one with the 80 suffix morphs that Morfessor extracted from the story corpus.

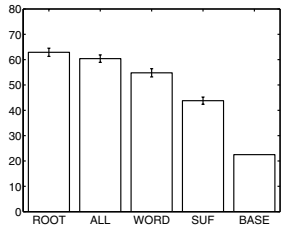
Other experiments. Some experiments on combining together different types of feature morphs as well as on the effect of the number of features in the feature set were also studied, but they will not be further considered in this paper as the results fell between the reported figures.

Baseline. For comparison, also a baseline similarity measure was included. It counts for every word form in the training word set the percentage of other training words sharing at least one POS with it. Again, the result is an average over all the words in the training set. This corresponds roughly to the idea of a SOM organized in a completely random fashion.

6 Results

The evaluation results of the experiments can be found in Fig. 1. As can be seen, all word SOMs clearly outperform the (rather crude) baseline similarity, whether they had morphs or whole context words as features. This indicates that all the word SOM variants that were evaluated succeeded in creating word categorizations which surpass in quality a random organization of the data.

All but one of the word SOMs that utilized morphological information in their feature sets seemed to fare better in the evaluation than the traditional



Feature sets:

ROOT : root morphs (200 most frequent)
ALL : all morphs (200 most frequent)
WORD : whole words (200 most frequent)
SUF : suffix morphs only (80 most frequent)
BASE : baseline similarity measure

Fig. 1. The accuracies of the experiments (error bars mark one standard deviation)

word SOMs with whole context words as features. The best results were yielded by word SOMs with only root morphs as features, but when also suffixes were added to feature sets, the evaluation results seemed to slightly decline. Further, the only morph-featured word SOM variant that actually fared worse than the traditional SOMs with whole context words as features was the one with only suffix morphs in its feature set.

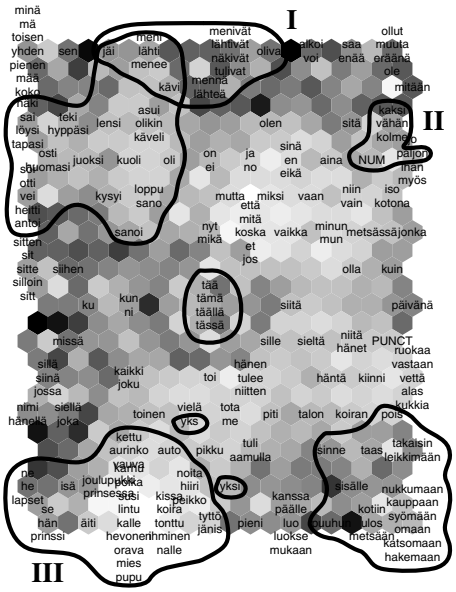
These results imply that, firstly, utilizing automatically discovered morphological units in the feature set of a word SOM does indeed improve the quality of the resulting word category SOM. Due to the highly inflecting nature of Finnish, the better performance of morph features as compared to whole feature words is hardly a surprise. With context words segmented into roots and into a variety of derivational or inflectional affixes, it is clear that some of the data sparsity problems caused by the diversity of Finnish inflected word forms are solved.

Secondly, it seems that not all morphs make equally good features: the best-quality word SOMs were constructed by using only root morphs as features. The fact that the POS-based quality evaluation measure we developed seemed to penalize the inclusion of other types of morphs (prefixes, suffixes) into the feature set appears to imply that most of the semantic and POS information of Finnish word forms is carried in their roots, not so much in the affixes attached to these roots. However, affix morphs should not be too hastily rejected as bad features on the basis of these evaluation results, as it may e.g. be the case that they encode some entirely different characteristic of words than their POS class.

6.1 Example of a Word SOM Analysis

Another objective of our research was to analyze the unique Finnish children's stories corpus (see Sect. 4) using SOMs. With the optimal features discovered in the experiments, we trained word category SOMs on the whole data set and also on story data from different age categories of children. As including also suffix morphs into the feature set appeared to bring some additional benefits from the point of view of analyzing and interpreting the story data, the final analysis maps were trained using a feature set of the 200 most frequent root morphs combined with the 20 most frequent suffixes.

An example word category SOM, constructed using the whole story corpus and the feature set described above, can be found in Fig. 2. Some emergent



I Verbs related to movement
- lähti he/she left
- menivät they went
- tulivat they came

II Numerals
- kaksi two
- paljon a lot of

III Agent nouns and pronouns
- hän he/she
- he they
- lapset children
- prinssi prince
- poika boy
- tyttö girl
- isä father
- äiti mother
- susi wolf
- lintu bird
- hevonen horse

Fig. 2. The U-matrix representation of a word category SOM trained on the whole children's stories corpus. Some interesting clusters of semantically similar words have been manually highlighted, and some words from three of them are presented in more detail on the side of the map (with English translations). Notice for example how the nouns in group III are typical agents in the children's stories.

groups of intuitively similar words have been manually highlighted to the resulting word map. However, we will not go here into a further analysis of the language use of young Finnish children due to space limitations.

7 Conclusions

We trained word category SOMs on Finnish text data using morphologically informed features that were extracted from the corpus itself with an unsupervised morphological segmentation algorithm called Morfessor. Experiments were performed on different kinds of SOM feature sets with morphs and whole context words. The resulting word SOMs were evaluated with an evaluation measure developed for this task, based on a list of part-of-speech -classified word forms.

Our experiments showed that the use of Morfessor-extracted morphs as word SOM features clearly improves the POS density -based quality of the resulting word SOMs, as opposed to using unsegmented context words for features. However, some types of morphs seem to make better features than others. The best resulting word SOMs were trained by using a feature set with only root morphs, chosen from the top of a morph frequency list calculated from the data.

The work described in this paper is the first completely automated categorization of Finnish word forms with morphology-utilizing word SOMs. It is also a

study on the problem of lexical acquisition in a highly inflecting language in general. Particularly, the method we described aims at – and succeeds in – tackling the critical problem of data sparsity, typical to morphologically rich languages (as opposed to e.g. languages like English).

In future, the current work could be extended in many ways. First, the experiments should be re-run on data in other languages, and maybe also on another, larger corpus in Finnish. Further, the morphological features of also the training word itself could be utilized, examining e.g. only some very high-frequency morphs like common suffixes. Finally, the Morfessor-extracted morphological features could be used in many other lexical acquisition tasks, like e.g. finding verb subcategorization patterns or selectional preferences of words.

References

1. Ritter, H., Kohonen, T.: Self-Organizing Maps. Biological Cybernetics **61** (1989) 241–254
2. Honkela, T., Pulkki, V., Kohonen, T.: Contextual relations of words in Grimm tales analyzed by self-organizing map. In: Proceedings of ICANN-95. Volume 2., Paris, EC2 et Cie (1995) 3–7
3. Redington, M., Chater, N., Finch, S.: Distributional information: A powerful cue for acquiring syntactic categories. Cognitive Science **22**(4) (1998) 425–469
4. Schütze, H.: Automatic word sense discrimination. Computational Linguistics **24**(1) (1998) 97–123
5. Lagus, K., Airola, A., Creutz, M.: Data analysis of conceptual similarities of Finnish verbs. In: Proceedings of the CogSci 2002, Fairfax, Virginia (2002) 566–571
6. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of English words. In: ACL 30. (1993) 183–190
7. Schulte im Walde, S.: Clustering verbs semantically according to their alternation behaviour. In: COLING-00. (2000) 747–753
8. Light, M.: Morphological cues for lexical semantics. In: ACL 34. (1996) 25–31
9. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Computational Linguistics **27**(2) (2001) 153–198
10. Creutz, M., Lagus, K.: Unsupervised discovery of morphemes. In: Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02, Philadelphia, Pennsylvania (2002) 21–30
11. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics **43** (1982) 59–69
12. Creutz, M., Lagus, K.: Inducing the morphological lexicon of a natural language from unannotated text. In: Proceedings of AKRR'05, Espoo (2005) 106–113
13. Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J.: Unlimited vocabulary speech recognition with morph language models applied to Finnish. Computer Speech and Language **20**(4) (2006) 515–541
14. Kohonen, T.: Self-Organizing Maps. 3rd edn. Springer, Berlin (2001)
15. Riihelä, M.: The Storycrafting Method. Stakes, Helsinki, Finland (2001)
16. Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T., Alho, I.: Iso suomen kielioppi. Suomalaisen Kirjallisuuden Seura, Helsinki (2004)