Novel Properties and Well-Tried Performance of EM-Based Multivariate Clustering

Detlef Prescher
DFKI Language Technology
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
prescher@dfki.de

Abstract

We present three novel properties for EM-based multivariate clustering: simplified re-estimation formulas, a simple pruning technique, and a novel invariance property preserving the characteristics of the given empirical distribution. Evaluation on two tasks shows: EM-based multivariate clustering models require only twice the storage space of the original sample, and these models yield reliable estimates for unknown data. Moreover we refer to selected experiments showing that EM-based multivariate clustering improves several real-world applications.

1 Introduction

EM-based multivariate clustering (EMMC) as introduced by (Müller *et al.* 00) is an application-independent framework for unsupervised learning from multivariate data via the standard Expectation Maximization (EM) algorithm (Dempster *et al.* 77). EMMC provides both good generalization performance on sparse data and structural information on the data-inherent grouping structure.

The primary goal of this paper is to present some newly discovered mathematical properties of EMMC suggesting why EMMC might be useful for natural language processing (NLP) applications. A secondary goal of the paper is to demonstrate that EMMC does indeed improve several NLP applications.

Our approach is based upon three resources. The first resource are simplified re-estimation formulas which achieve deeper insight into EMMC. The second resource consists of a newly discovered invariance property which turns out to be the theoretical explanation of the good smoothing and disambiguation results observed in several NLP applications using EMMC. The third resource is a new pruning technique which leads in practice to the efficient representation of the structural information provided by EMMC.

The paper is organized as follows: in Section 2 we introduce a simplified presentation of EMMC; in Section 3 we present the fundamental invariance property, and in Section 4 we present our new pruning technique. Section 5 is dedicated to some NLP applications which take advantage of EMMC. In Section 6 we discuss our results and in Section 7 we summarize our conclusions.

2 Theory

Multivariate data refers to a domain $\mathcal{Y}$ with two or more finite sets $\mathcal{Y}_i$ of objects in which observations are made for vectors $y \in \mathcal{Y}$ with one element from either set, i.e. $y_i \in \mathcal{Y}_i$. Multivariate data arises naturally in many NLP applications. In the EMMC approach of (Müller *et al.* 00), classes corresponding to multivariate data (3- and 5-dimensional syllable types) are viewed as hidden data in the context of the maximum likelihood estimation from incomplete data via the EM algorithm. The two main tasks of EMMC are (i) the induction of a smooth probability model on the data, and (ii) the automatic discovery of class structure in the data. The aim is to derive a probability distribution $p(y)$ on multivariate data from a large sample. The key idea is to view y as conditioned on an unobserved class $c \in C$, where the classes are given no prior interpretation. The probability of a d-dimensional data type $y = (y_1, .., y_d)$ is defined as:

$$\begin{aligned} p(y) &= \sum_{c \in C} p(c, y) = \sum_{c \in C} p(c) \cdot p(y|c) \\ &= \sum_{c \in C} p(c) \prod_{i=1}^{d} p(y_i|c) \ . \end{aligned}$$

Note that the independence assumption

$$p(y|c) = \prod_{i=1}^{d} p(y_i|c)$$

for each class c makes clustering feasible since an independence assumption is not adequate for the whole sample, and therefore forces the sample to break into clusters. In general, the number

$\tilde{p}(y_1, y_2)$	smile	laugh	increase	fall	$\tilde{p}(y_1)$
man	0.2				0.2
woman	0.2	0.1			0.3
number			0.1		0.1
price			0.3	0.1	0.4
$\tilde{p}(y_2)$	0.4	0.1	0.4	0.1	

$p(y_1, y_2)$	smile	laugh	increase	fall	$p(y_1)$
man	0.16	0.04			0.2
woman	0.24	0.06			0.3
number			0.08	0.02	0.1
price			0.32	0.08	0.4
$p(y_2)$	0.4	0.1	0.4	0.1	

Figure 1: Invariant marginals of empirical (left) and model distribution (right)

$|C|$ of classes will be experimentally determined such that the assumption is optimally met. The EM algorithm (Dempster *et al.* 77) is directed at maximizing the incomplete data log likelihood $L = \sum_y \tilde{p}(y) \ln p(y)$ as a function of the probability distribution p for a given empirical probability distribution $\tilde{p}$.

Let $f(y)$ be the frequency of a multivariate data type y, and

$$f_c(y) = f(y) \cdot p(c|y)$$

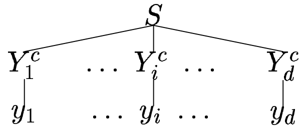
the so-called *class-based frequency* of y annotated with class c. Note that $p(c|y)$ can be interpreted as a class-membership probability "$p(y \in c)$" since $\sum_c p(c|y) = 1$. So-called *marginal class-based frequencies* $f_c(y_i) = \sum_{y \in \mathcal{H}(y_i)} f_c(y)$ of an object $y_i \in \mathcal{Y}_i$ can be computed by summing up the class-based frequencies of all data types lying in the $(d-1)$-dimensional hyper plane $\mathcal{H}(y_i) = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_{i-1} \times \{y_i\} \times \mathcal{Y}_{i+1} \times \ldots \times \mathcal{Y}_d$. Finally, $|f| = \sum_{y \in \mathcal{Y}} f(y)$ is the total frequency of the sample, as $|f_c|$ are the total class-based frequencies. Parameter updates $\hat{p}(c)$, $\hat{p}(y_i|c)$ can thus be computed by $(c \in C,\ y_i \in \mathcal{Y}_i,\ i = 1,..,d)$:

$$\begin{array}{rcl} \hat{p}(c) & = & |f|^{-1} \cdot |f_c| \ , \\ \hat{p}(y_i|c) & = & |f_c|^{-1} \cdot f_c(y_i) \ . \end{array}$$

Proof: Following the lines of (Prescher 01b) the proof consists of two steps: (i) A given EMMC model is equivalent to a simple probabilistic regular grammar $(c \in C,\ y_i \in \mathcal{Y}_i,\ i = 1,..,d)$:

$$\begin{array}{ll} S \to Y_1^c \ldots Y_d^c & (p(c)) \\ Y_i^c \to y_i & (p(y_i|c)) \end{array}$$

Here, S is the starting symbol, and Y_i^c is a nonterminal symbol. A "sentence" $y_1 \ldots y_i \ldots y_d$ has exactly $|C|$ "syntax trees":



Each "syntax tree" has the probability of

$$p(c) \prod_{i=1}^{d} p(y_i|c)$$

yielding a "sentence probability" of

$$\sum_{c \in C} p(c) \prod_{i=1}^{d} p(y_i|c)$$

which is equal to the EMMC probability of the type $y = y_1 \ldots y_d$. (ii) Applying the EM re-estimation formulas for context-free grammars (see e.g. (Prescher 01a) for a discussion of these formulas in the context of the well-known inside-outside algorithm)

$$\hat{p}(r) = \frac{\sum_y f(y) \sum_x p(x|y) \cdot f_r(x)}{\sum_y f(y) \sum_x p(x|y) \cdot f_A(x)}$$

(r denotes a rule, A its left-hand side, y a sentence, x its syntax trees, and $f_r(x)$ and $f_A(x)$ count how often r and A occure in x) to the regular grammar of the first step, (Prescher 01b) shows that this yields the proposition **q.e.d.**

Note that the shown re-estimation formulas are simplified versions of the formulas presented in (Müller *et al.* 00). The simplified formulas reveal that the re-estimation process of EMMC is based on recursive application of two basic techniques:

- marginalization (via hyper planes), and
- normalization (of marginal frequencies).

In the following, we present an invariance property relying on these simple re-estimation techniques.

3 Invariant Marginal Distributions

For a given training corpus, unknown data can be defined as the set of data types not occurring in the training corpus but in the union of all evaluation corpora which are reasonable for an application in mind. Given this definition, real-world data generally splits into two parts, (i) the huge set of unknown data, and (ii) the tiny set of training data. For this reason, the most important property of all probability models is their ability to deal with unknown data, i.e. to assign unknown (but reasonable) data types a non-zero probability. Since model parameters are solely estimated from the training corpus, without any help from other sources of information about the data, it is commonly accepted that estimates for unknown data are unreliable. In contrast to this general setting, we aim to derive an invariance property of EMMC which indicates that these models yield reliable estimates for unknown data, at least for the huge amount of unknown data types[1] lying in the given domain $\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_d$.

The left hand table of Figure 1 shows the empirical distribution $\tilde{p}(y_1, y_2) = |f|^{-1} \cdot f(y_1, y_2)$ and its two marginal distributions, $\tilde{p}(y_1)$ and $\tilde{p}(y_2)$, of a toy sample of 2-dimensional verb-noun data. The empirical probability of 0.1 in the third column and third row of this table indicates that 10 % of the tokens of the sample matches the data type *(woman,laugh)*, whereas the empty entry in the third column and the second row indicates that the data type *(man, laugh)* has the empirical probability of zero, and does not occur in the sample. Given this sample as training data, the pair *(woman, laugh)* is a known data type, whereas the pair *(man, laugh)* is an unknown data type. It follows by definition, that unknown data types have an empirical probability of zero. However, the table also includes unreasonable data, like *(number, smile)*, with an empirical probability of zero.

The table also shows the two marginal distributions $\tilde{p}(y_1)$ and $\tilde{p}(y_2)$. The value of 0.4 in the second column and sixth row indicates that *smile* occurs in 40% of the data tokens in the sample, whereas the value of 0.3 in the sixth column and third row indicates that *woman* occurs in 30%. The right hand table of Figure 1 shows a smoothed probability distribution of this sample. Smoothed probability distributions ideally assign unknown data types, in this case *(man,laugh)* and *(number, fall)*, a non-zero probability (0.04 and 0.02, respectively), whereas unreasonable data types keep their zero-probability. Of course, the shown probability distribution is only one example out of the set of all possible smoothed probability distributions. However, a closer look at Figure 1 shows that it is a very special probability distribution: its two marginal distributions $p(y_1)$ and $p(y_2)$ are identical to the marginal distributions, $\tilde{p}(y_1)$ and $\tilde{p}(y_2)$, of the empirical distributions. We call this property an invariance property. Invariance properties are very useful in settings where a given object must be carefully modified, since they force the modification process to respect certain constraints. In our case, the empirical distribution $\tilde{p}(.)$ is modified to the smoothed distribution $p(.)$, but the corrections are regarded as minimal, because the marginal distributions are invariant. Thus, it may be conjectured, that the smoothed distribution preserves some characteristical properties of the empirical distribution, i.e. of the given corpus.

class 0 0.5	woman 0.6 man 0.4	smile 0.8 laugh 0.2
class 1 0.5	price 0.8 number 0.2	increase 0.8 fall 0.2

Figure 2: EMMC model given the toy sample

Figure 2 shows the two classes of an EMMC model trained on the sample shown in the left hand table of Figure 1. The model was trained with 10 iterations and randomly initialized starting parameters. The first column displays the class index and the class probability, the nouns and their probabilities are listed in descending order in the second column, as are the verbs in the third column. Note, the smoothed probability distribution given in the right hand table of Figure 1 can be computed with these model parameters. This observation indicates that EMMC models possibly have invariant marginal distributions. We show in the following two steps that this is indeed true: the following formula shows that the i^{th} marginal distribution $p(y_i)$ of an EMMC model can be easily computed using class probabilities and the probabilities of y_i given the classes:

$$\begin{aligned}
p(y_i) &= \sum_{y \in \mathcal{H}(y_i)} p(y) \\
&= \sum_{y \in \mathcal{H}(y_i)} \sum_{c \in C} p(c) \prod_{j=1}^{d} p(y_j|c) \\
&= \sum_{c \in C} p(c) \sum_{y \in \mathcal{H}(y_i)} \prod_{j=1}^{d} p(y_j|c) \\
&= \sum_{c \in C} p(c) \cdot p(y_i|c) \prod_{j \neq i} \sum_{y_j \in \mathcal{Y}_j} p(y_j|c) \\
&= \sum_{c \in C} p(c) \cdot p(y_i|c) \ .
\end{aligned}$$

[1] For example, in the smoothing experiment of Section 4 we experimented with a sample of about 600 000 observed verb-noun types lying in a domain of about 400 000 000 (reasonable and unreasonable) verb-noun types.

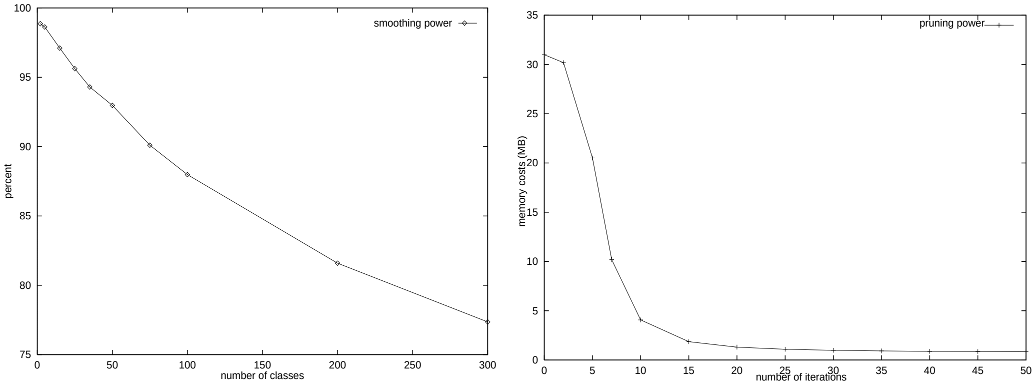


Figure 3: Evaluation via smoothing (left) and pruning (right)

Note that the shown computation only requires the model property of EMMC. In a second step, we additionally use the re-estimation formulas of EMMC. We derive:

$$\begin{aligned}
\hat{p}(y_i) &= \sum_{c \in C} \hat{p}(c) \cdot \hat{p}(y_i|c) \\
&= \sum_{c \in C} |f|^{-1} \cdot |f_c| \cdot |f_c|^{-1} \cdot f_c(y_i) \\
&= \sum_{c \in C} |f|^{-1} \sum_{y \in \mathcal{H}(y_i)} f_c(y) \\
&= \sum_{y \in \mathcal{H}(y_i)} |f|^{-1} \cdot f(y) \sum_{c \in C} p(c|y) \\
&= \sum_{y \in \mathcal{H}(y_i)} \tilde{p}(y) \\
&= \tilde{p}(y_i) \ .
\end{aligned}$$

This result shows that the marginal distributions $p(y_i)$, $(i = 1 \ldots d)$, are invariant during the re-estimation process of the EM algorithm, and we conjecture that $p(.)$ preserves the main characteristical properties of the given corpus[2].

4 Pruning and Smoothing

It is usually the case that more training data will improve a probabilistic model. However, as the size of the training data increases, the size of EMMC models increases, which can lead to models that are too large for practical use. To deal with this problem, we propose a novel pruning technique which deletes infrequent data types class-wise from the model.

Class-based frequencies $f_c(y)$ play a major role in the context of EMMC: (i) they are very successfully applied to resolve lexical ambiguities (Section 5), (ii) their marginals $f_c(y_i)$ are the key variables of the simplified re-estimation formulas (Section 2). As a consequence, our implementation of EMMC uses the class-based frequencies $f_c(y)$ as internal parameters. Unfortunately, it can be shown that the size of the internal parameters does not decrease during the pure mathematical re-estimation procedure since positivity of the model parameters is a second invariance property of EMMC (at least for each finite number of iterations). For this reason, a model with $|C|$ classes would require about $|C|$-times disk space than the original data. Therefore, a simple class-wise cutoff technique was used to prune EMMC models. After each iteration step of the EM algorithm, we deleted all infrequent class-based frequencies

$$f_c(y) < \epsilon \cdot |C|^{-1} \ .$$

Based on experience, we chose $\epsilon = 0.001$ in all of our recent experiments[3]. The cutoff $\epsilon \cdot |C|^{-1}$ decreases as the number of classes increases. This is the desired effect since the class-based frequencies $f_c(y)$ generally decrease as the number of classes increases. The latter follows since the sample frequency distributes among the classes: $f(y) = \sum_{c \in C} f_c(y)$. Thus, it makes sense to use the cutoff ϵ for $f_c(y) \cdot |C|$ rather than for $f_c(y)$.

The EMMC models were empirically evaluated by a pruning task. Input to the clustering algorithm was a corpus of about 27 000 verb-object pairs extracted from the Penn Wall Street Journal. We trained a 2-dimensional clustering model with 35 classes and 50 iterations. The right-hand

[2] Here is an additional, more formal argument: Maximization of the log likelihood $L(p)$ is equivalent to minimization of the Kullback-Leibler divergence $D(\tilde{p}||p)$. Thus, we should choose a model p as close as possible to the given empirical distribution $\tilde{p}$. Obviously, any kind of invariance can guide this process.

[3] as well as in our implementation (Prescher 01c).

NLP application	Improvement of performance using EMMC
Induction of semantically annotated lexicons	∞
Identification of collocations/idioms	0%
Grapheme-to-phoneme conversion	3% (30%)
Machine translation	7 %
Stochastic lexicalized parsing	13-16 %

Figure 4: Improvement of performance of several NLP applications using EMMC

side of Figure 3 shows the required storage space of the models (in MB) during the re-estimation procedure which dramatically decreases between 3 and 10 iterations. Finally, models trained with more than 15 iterations require only twice the storage space of the original sample.

Moreover it is important to note that the best EMMC models used in real-world applications (see Section 5) were trained with about 10 to 100 iterations. Thus, our pruning method is effective at reducing the model size without significantly reducing the model performance.

Unfortunately, the presented pruning technique can have a negative effect with respect to the smoothing behaviour: (i) EMMC models can be expected to be good smoothers, because the model probability $p(y)$ of a multivariate data type y is positive if a class $c \in C$ exists such that the conditional probability $p(y|c)$ is positive. (ii) Unfortunately, the used pruning technique eliminates a data type from the model if these conditional probabilities are small enough.

As a consequence, it is necessary to evaluate the EMMC models by a smoothing task. The so-called *smoothing power* of a multivariate clustering model is defined as the number of data types which receive a positive probability by the model, normalized by the size of the set $\mathcal{Y}_1 \times \ldots \times \mathcal{Y}_d$ of all possible (including unreasonable) multivariate data types. Using the invariance property, it follows for EMMC models with only one class ($|C| = 1$) that $p(y) = \prod_{i=1}^{d} p(y_i) = \prod_{i=1}^{d} \tilde{p}(y_i)$. Obviously, these models have a smoothing power of 100%. However, we expect that the smoothing power of EMMC models decreases as the number of classes increases since $p \to \tilde{p}$ as $|C| \to \infty$. The left-hand side of Figure 3 shows the smoothing results of EMMC models trained (constantly with 50 iterations) on 2-dimensional English verb-noun data (Rooth *et al.* 99). For example, a model with 35 classes had a smoothing power of about 95% which is about 700 times better than the smoothing power of the empirical probability distribution which has a value of 0.14%. Starting values had an effect of only 1% on the performance.

5 Well-Tried Performance

From a practical point of view, the task of a stochastic model is to decide among alternative analyses proposed by the symbolic analysis component of a given application. During the last ten years, evaluation via a so-called pseudo-disambiguation task has become very popular (e.g. (Pereira *et al.* 93), (Rooth *et al.* 99), (Müller *et al.* 00)). The simple task is to judge which of two objects is more likely to appear in the context of a given observation y, where an participating object y_i is compared with a randomly chosen object y'_i (e.g. *man smile* with *number smile* (i=1) or *man increase* (i=2)). Pseudo-disambiguation offers two important technical advantages: huge evaluation suites (with several thousand test items) can be automatically constructed, and free model parameters can be automatically determined by the evaluation results.

Unfortunately, the pseudo-disambiguation task seems very artificial and unrealistic to some people, and thus, it seems necessary to evaluate the EMMC models in real-world applications on a large number of randomly selected examples of a real-world corpus. For this purpose, we refer to some selected NLP applications and show that the performance of these applications were often dramatically improved using EMMC.

First, Figure 4 gives an overview. The selected EMMC-based applications are shown in the first column, and the achieved gains in performance are listed in the second column. In detail:

Unsupervised Induction of Semantically Annotated Lexicons. A technique for automatic induction of slot annotations for subcategorization frames was presented by (Rooth *et al.* 99). Possible annotations consist of the hidden classes of a sample of 2-dimensional verb-argument pairs, where EMMC was used to reveal these classes. Induction of slot labeling for subcategorization frames is accomplished by a further application of EM, and applied experimentally on frame observations derived from parsing large corpora. Thus, it can be argued that unsupervised induction of semantically annotated lexicons from free text completely relies on infered EMMC models, and it seems not unfair to attribute an infinite gain of performance to EMMC.

Identification of Collocations/Idioms. An identification method for verb-noun collocations/idioms was investigated by (Prescher & Heid 00). The defining criterion for EMMC is that verbs and arguments freely combine with each other inside each class and thereby semantically characterize the classes. In contrast to this, collocations/idioms do not show this behavior: their lexemes combine seldom with other partners, and the meaning of an idiom can not be computed using the meanings of the lexemes. Obviously, collocations/idioms do not conform to the requirements of EMMC. (Prescher & Heid 00) exploited this observation and presented a method for identification of verb-noun collocations/idioms based on a simple comparison of the empirical distribution $\tilde{p}(v,n)$ with the distribution $p(v,n)$ of an infered EMMC model, where high values of $\tilde{p}(v,n) - p(v,n)$ are expected to identify collocations/idioms. Recent evaluation on 400 items shows that this method yields as good results as current state-of-the-art methods for identification of collocations. Thus, we attribute a gain of 0% performance to EMMC.

Grapheme-to-Phoneme Conversion. A novel method of g2p conversion was presented by (Müller *et al.* 00). Their approach (i) uses a context-free grammar to produce all possible phonemic correspondences of a given grapheme string, (ii) applies a probabilistic syllable model to rank the pronunciation hypotheses taking the product of the syllable probabilities, and (iii) predicts pronunciation by choosing the most probable analysis. The g2p system was evaluated on a test set of about 2000 unknown words. The ambiguity expressed as the average number of analyses per word was about 300. The g2p system using 5-dimensional syllable models achieves an increase of 3% over the performance of the 5-dimensional baseline system using the empirical syllable distribution. If compared to the standard probability model (probabilistic context-free grammars), it achieves an increase of about 35% (Müller 01).

Machine Translation. A novel approach of lexical ambiguity resolution in machine translation was presented by (Prescher *et al.* 00). The problem to be solved is to find a correct translation of a source word using only minimal contextual information. EMMC was used by choosing the target noun $\hat{n}$ (and a class $\hat{c}$) such that $(\hat{n},\hat{c}) = \underset{n\in\mathcal{A},c\in C}{\operatorname{argmax}} \ (f_c(n,v) + p(c|n,v))$, where $\mathcal{A}$ is the set of alternative target-words and v is the governing verb. The term $f_c(n,v) + p(c|n,v)$ is a class-based frequency but based upon a "tuned frequency" of $f(n,v)+1$. The evaluation on a corpus with about 800 bilingual sentence pairs with about 3 translation alternatives on average yields an increase of 7% precision over the baseline system using the empirical noun-verb distribution as disambiguator.

Stochastic Lexicalized Parsing. An approach to stochastic modeling of unification-based grammars was presented by (Riezler *et al.* 00). The approach is based on lexicalized log-linear models and uses EM for estimation from unannotated data. Very similarly to the disambiguation routine used in machine translation, all parses of a sentence were pre-disambiguated according to maximal class-based frequencies of verb-noun heads in certain grammatical relations. The stochastic model was evaluated on a corpus of 550 sentences of a foreign language learner's grammar. The average ambiguity of this corpus is about 5 parses per sentence. An incorporation of the pre-disambiguation routine into the log-linear models improves precision of the stochastic model by about 13%. It is interesting that incorporation of class-based frequencies improves stochastic models by about 16% if treebank training is used instead of EM training.

6 Discussion

EM-based clustering was derived and applied to syntax (Rooth *et al.* 99). Unfortunately, this approach is not applicable to multivariate data with more than two dimensions. However, (Müller *et al.* 00) presented EMMC models with 3- and 5-dimensional syllables and applied these models successfully to phonology.

Restricting the discussion to two-dimensional data, EMMC models can be found earlier in (Pereira *et al.* 93). In contrast to this approach, EMMC is formalized clearly as EM algorithm, whereas (Hofmann & Puzicha 98) propose an annealed version of standard EM.

However, the re-estimation formulas of EMMC play a major role in the proof of the fundamental invariance property of EMMC models. Thus it can be conjectured that the invariance property can not be proven in the approaches of (Pereira *et al.* 93) and (Hofmann & Puzicha 98).

Furthermore, to the best of our knowledge, this is the first time that effective pruning techniques have been presented for clustering approaches to multivariate data.

EMMC may also be compared to the approaches of (Schütze 92) and (Yarowsky 95) to word sense disambiguation. However, a comparison is not straight-forward, because these approaches use large amounts of information in terms of large context windows, which is easily obtainable in IR applications, but often unavailable in situations such as parsing or translation.

There has been a large amount of previous work on smoothing and most methods have been shown to be highly effective. However, the results of Section 5 indicate that the use of class-based frequencies (lexicon induction, machine translation and parsing) provided by EMMC is more advantageous for disambiguation than the pure use of the smoothed model distribution itself (for example in grapheme-to-phoneme conversion). EMMC utilizes the hidden structural information of given data and clearly, this is a feature lying beyond the capability of a pure smoother.

7 Conclusion

We have presented simplified re-estimation formulas of EMMC which allow deeper insight into EMMC theory. As a consequence, these simplified formulas lead to detection of a new invariance property of EMMC.

We have presented a new pruning technique which makes EMMC feasible. Experimental results show that this new pruning technique leads to effective EMMC models of only about twice the size of the original data.

We have shown that EMMC improves several NLP applications, e.g. machine translation, where other clustering methods are not applicable or e.g. lexicon induction, where competing smoothing methods are useless. We believe that the presented invariance property is the basis for the well-tried performance of EMMC.

References

(Dempster *et al.* 77) A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *J. Royal Statist. Soc.*, 39(B):1–38, 1977.

(Hofmann & Puzicha 98) Thomas Hofmann and Jan Puzicha. Unsupervised learning from dyadic data. Technical Report TR-98-042, International Computer Science Insitute, Berkeley, CA, 1998.

(Müller 01) Karin Müller. Probabilistic context-free grammars for syllabification and grapheme-to-phoneme conversion. In *Proceedings of EMNLP'01*, Pittsburgh, 2001.

(Müller *et al.* 00) K. Müller, B. Möbius, and D. Prescher. Inducing probabilistic syllable classes using multivariate clustering. In *Proc. of ACL-2000*, Hong Kong, 2000.

(Pereira *et al.* 93) F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proc. of ACL'93*, 1993.

(Prescher & Heid 00) D. Prescher and U. Heid. Probabilistisches Clustering zur Identifikation von Verb-Nomen-Kollokationen. Presented at DGfS 2000, Marburg, 2000.

(Prescher 01a) Detlef Prescher. Inside-outside estimation meets dynamic EM. In *Proceedings of IWPT-2001 (to appear)*, Beijing, 2001.

(Prescher 01b) Detlef Prescher. *EM-basierte maschinelle Lernverfahren für natürliche Sprachen.* Unpublished PhD thesis, IMS, University of Stuttgart, to appear 2001.

(Prescher 01c) Detlef Prescher. EMMC Toolkit. Software package for EM-based multivariate clustering. IMS, Universität Stuttgart. http://www.ims.uni-stuttgart/~prescher, to appear 2001.

(Prescher *et al.* 00) Detlef Prescher, Stefan Riezler, and Mats Rooth. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of COLING-2000*, Saarbrücken, 2000.

(Riezler *et al.* 00) S. Riezler, D. Prescher, J. Kuhn, and M. Johnson. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and EM training. In *Proc. of ACL-2000*, 2000.

(Rooth *et al.* 99) M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil. Inducing a semantically annotated lexicon via EM-based clustering. In *Proc. 37th Ann. Meeting of the ACL*, College Park, MD, 1999.

(Schütze 92) Hinrich Schütze. Dimensions of meaning. In *Proceedings of Supercomputing '92*, 1992.

(Yarowsky 95) D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of ACL'95*, Cambridge, 1995.