AIMS

Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung
Lehrstuhl für Theoretische Computerlinguistik Universität Stuttgart

1998 • VOL. 4 • NO. 3

Inducing Lexicons with the EM Algorithm

Previous Issues of AIMS:

Vol.1 (1) 1994: *Wigner Distribution in Speech Research.* Master thesis by Wolfgang Wokurek (in German), papers by Grzegorz Dogil, Wolfgang Wokurek, and Krysztof Marasek (in English), and a bibliography.

Vol.2 (1) 1995: *Sprachentstörung.* Doctoral Dissertation. University of Vienna, 1994 by Wolgang Wokurek (in German with abstract in English). Full title: Sprachentstörung unter Verwendung eines Lautklassendetektors (Speech enhancement using a sound class detector).

Vol.2 (2) 1995: *Word Stress.* Master thesis by Stefan Rapp (in German) and papers mostly by Grzegorz Dogil, Michael Jessen, and Gabriele Scharf (in English).

Vol.2 (3) 1995: *Language and Speech Pathology.* Master theses by Gabriele Scharf and by Jörg Mayer (in German) and papers by Hermann Ackermann, Ingo Hertrich, Jürgen Konczak, and Jörg Mayer (mostly in German).

Vol.3 (1) 1997: *Tense versus Lax Obstruents in German.* Revised and expanded version of Ph.D. Dissertation, Cornell University, 1996 by Michael Jessen (in English). Full title: Phonetics and phonology of the tense and lax obstruents in German.

Vol.3 (2) 1997: *Electroglottographic Description of Voice Quality.* Habilitationsschrift, University of Stuttgart, 1997 by Krysztof Marasek (in English).

Vol.3 (3) 1997: *Aphasie und Kernbereiche der Grammatiktheorie (Aphasia and core domains in the theory of grammar).* Doctoral Dissertation, University of Stuttgart, 1997 by Annegret Bender (in German with abstract in English).

Vol.3 (4) 1997: *Intonation and Bedeutung (Intonation and meaning).* Doctoral Dissertation, University of Stuttgart, 1997 by Jörg Mayer (in German with abstract in English).

Vol.3 (5) 1997: *Koartikulation und glottale Transparenz (Coarticulation and glottal transparency).* Doctoral Dissertation, University of Bielefeld, 1997 by Kerstin Vollmer (in German with abstract in English).

Vol.3 (6) 1997: *Der TFS-Repräsentationsformalismus und seine Anwendung in der maschinellen Sprachverarbeitung (The TFS Representation Formalism and its Application to Natural Language Processing).* Doctoral Dissertation, University of Stuttgart, 1997 by Martin C. Emele (in German).

Vol.4 (1) 1998: *Automatisierte Erstellung von Korpora für die Prosodieforschung (Automated generation of corpora for prosody research).* Doctoral Dissertation, University of Stuttgart, 1998 by Stefan Rapp (in German with abstract in English).

Vol.4 (2) 1998: *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese (Theory-based modelling of German intonation for speech synthesis).* Doctoral Dissertation, University of Stuttgart, 1998 by Gregor Möhler (in German with abstract in English).

Inducings Lexicons with the EM Algorithm

Mats Rooth, Stefan Riezler, Detlef Prescher, Sabine Schulte im Walde, Glenn Carroll, and Franz Beil

Lehrstuhl für Theoretische Computerlinguistik
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Rooth, Mats:
Inducing Lexicons with the EM Algorithm / Mats Rooth; Stefan Riezler; Detlef Prescher; Sabine Schulte im Walde; Glenn Carroll; Franz Beil. - Stuttgart, 1998.
AIMS - Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, Vol. 4, No. 3, 1998, Stuttgart, Germany.
ISSN 1434-0763

Lehrstuhl für Theoretische Computerlinguistik
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstraße 12
70174 Stuttgart

www: `http://www.ims.uni-stuttgart.de/projekte/gramotron/`
email: `gramotron@ims.uni-stuttgart.de`

Contents

EM-Based Clustering for NLP Applications

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil

Abstract

We present a probabilistic clustering method defining the clustering task as induction of hidden parameters of a probability model on dyadic linguistic data. Induction of unobserved class parameters is accomplished as statistical inference in the framework of maximum-likelihood estimation from incomplete data via the EM algorithm. We present illustrative examples of induced clusters from large sets of English and German data. The clustering models are evalutated on a pseudo-disambiguation task testing the power of the models to generalize over unseen data. We present two applications of the clustering models to natural language processing applications. The first task concerns the induction of semantic labels for subcategorization slots of English and German lexical verbs. Based on induced clusters of subcategorization frames and their nominal fillers, induction of labels of the respective slots is accomplished by a further application of EM. We report experiments with large sets of German and English intransitive and transitive verbs and outline a linguistic interpretation of the learned representations. A second application concerns the use of clustering models on English verb-noun combinations to select among candidate English translations of German nouns. We present an evaluation of clustering models on a pseudo-disambiguation task for noun-ambiguity. Furthermore, we evaluate the models on a gold standard extracted from a bilingual corpus. The performace of cluster-based target-language disambiguation is reported for models of various sizes and compared to simpler statistical disambiguation methods.

5.1 Introduction

Probabilistic clustering methods for natural language applications mainly focus on the following two tasks: (i) induction of smooth probability models on language data, (ii) automatic discovery of class-structure in natural language. The first task uses clustering as a solution to the problem of sparse data. In such applications the class-structure itself is not of interest, rather data clusters are consulted as general back-up sources of information when information about specific events is sparse or missing in the input. In contrast to this, for the second task we are interested in the structure of the induced clusters as a statistical semantics underlying the data in question. In the following we will present two applications of a probabilistic clustering model each of which stresses one of these tasks.

Clustering is conceptualized in our approach as induction of a class-based probability model on dyadic linguistic data—a sample of verb-noun pairs extracted from the maximal probability parses of large unannotated English and German data. Classes are introduced as hidden parameters of the probability model on verb-noun pairs. Induction of these unobserved parameters is accomplished in the framework of maximum likelihood estimation from incomplete data via the EM algorithm.

We present some illustrative examples of induced clusters and report the results of a harder evaluation of the models on a pseudo-disambiguation task testing the power of the models to generalize over unseen data.

The first application we will present is the induction of semantically annotated lexica based on induced clusters. In this application the stress is clearly on automatic discovery of class-structure. Based on clusters of subcategorization frames of verbs and their nominal fillers, induction of labels for the respective slots of subcat frames is accomplished by a further application of EM. We report results on experiments with observations derived from large English and German corpora. Furthermore, we outline an interpretation of the learned representations as theoretical-linguistic decompositional lexical entries.

The second application concentrates on the smoothing aspect of the clustering models. Here class-based probability models on English verb-noun combinations are used to disambiguate nouns senses in the context of ambiguous translations of nouns from German to English. We present an evaluation of models on a pseudo-disambiguation task for noun-ambiguity. In addition to this cheap evaluation, we present an evaluation on a gold standard which is extracted from a bilingual corpus. We report the performance of cluster-based target-language disambiguation for models of various sizes and compared to simpler methods such as plain maximum likelihood estimation on the training corpus.

5.2 EM-Based Clustering

In our clustering approach, classes are derived directly from distributional data—a sample of pairs of verbs and nouns, gathered by parsing an unannotated corpus and extracting the fillers of grammatical relations. Semantic classes corresponding to such pairs are viewed as hidden variables or unobserved data in the context of maximum likelihood estimation from incomplete data via the EM algorithm. This approach allows us to work in a mathematically well-defined framework of statistical inference, i.e., standard monotonicity and convergence results for the EM algorithm extend to our method.

The basic ideas of our EM-based clustering approach were presented in Rooth (Ms) (see also Rooth (1998)). An important property of our clustering approach is the fact that it is a "soft" clustering method, defining class membership as a conditional probability distribution over verbs and nouns. In contrast, in hard (Boolean) clustering methods such as that of Brown et al. (1992), every word belongs to exactly one class, which because of homophony is unrealistic. The foundation of our clustering model upon a probability model furthermore contrasts with the merely heuristic and empirical justification of similarity-based approaches to clustering (Dagan et al. 1998). The probability model we use can be found earlier in Pereira et al. (1993). However, in contrast to this approach, our statistical inference method for clustering is formalized clearly as an EM-algorithm. Approaches to probabilistic clustering similar to ours were presented recently in Saul and Pereira (1997) and Hofmann and Puzicha (1998). There also EM-algorithms for similar probability models have been derived, but applied only to simpler tasks not involving a combination of EM-based clustering models as in our lexicon induction experiment.

We seek to derive a joint distribution of verb-noun pairs from a large sample of pairs of verbs $v \in V$ and nouns $n \in N$. The key idea is to view v and n as conditioned on a hidden class $c \in C$, where the classes are given no prior interpretation. The semantically smoothed probability of a pair (v, n) is defined to be:

$$p(v,n) = \sum_{c \in C} p(c,v,n) = \sum_{c \in C} p(c)p(v|c)p(n|c)$$

The joint distribution $p(c, v, n)$ is defined by $p(c, v, n) = p(c)p(v|c)p(n|c)$. Note that by construction, conditioning of v and n on each other is solely made through the classes c.

In the framework of the EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997), we can formalize clustering as an estimation problem for a latent class (LC) model as follows. We are given:

- a sample space $\mathcal{Y}$ of observed, incomplete data, corresponding to pairs from $V \times N$,
- a sample space $\mathcal{X}$ of unobserved, complete data, corresponding to triples from $C \times V \times N$,

Class 5 PROB 0.0412		0.0148	0.0084	0.0082	0.0078	0.0074	0.0071	0.0054	0.0049	0.0048	0.0047	0.0046	0.0041	0.0040	0.0040	0.0039	0.0039	0.0039	0.0039	0.0038	0.0038	0.0038	0.0037	0.0035	0.0035	0.0035	0.0034	0.0033	0.0033	0.0033	0.0033
		man	ruth	corbett	doctor	woman	athelstan	cranston	benjamin	stephen	adam	girl	laura	maggie	voice	john	harry	emily	one	people	boy	rachel	ashley	jane	caroline	jack	burun	juliet	blanche	helen	edward
0.0542	ask.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●
0.0340	nod.as:s		●	●	●	●	●	●	●	●		●	●	●		●	●	●			●	●	●	●	●	●	●	●	●	●	●
0.0299	think.as:s	●	●	●	●	●	●			●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●		●	●	●	●
0.0287	shake.aso:s	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●		●	●	●	●	●	●	●	●	●		●	●
0.0264	smile.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●		●	●	●	●	●
0.0213	laugh.as:s	●	●		●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
0.0207	reply.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●			●				●	●
0.0167	shrug.as:s	●		●	●	●	●	●	●	●	●	●	●	●			●	●			●	●		●	●	●	●	●	●	●	●
0.0148	wonder.as:s	●	●	●	●	●	●			●	●	●				●	●	●		●	●	●	●	●	●	●	●	●	●	●	●
0.0141	feel.aso:s	●		●	●	●				●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●		●	●	●	●
0.0133	take.aso:s	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●
0.0121	sigh.as:s	●	●	●	●	●	●		●	●	●		●			●	●	●		●		●	●	●	●	●	●		●	●	●
0.0110	watch.aso:s	●	●	●		●	●	●	●	●	●	●		●		●	●	●	●	●	●	●		●		●	●	●		●	
0.0106	ask.aso:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●			●		●			●	●	●
0.0104	tell.aso:s	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●
0.0094	look.as:s	●	●	●	●	●	●	●	●	●	●	●				●		●	●	●	●	●				●	●	●			●
0.0092	give.aso:s	●	●		●	●			●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
0.0089	hear.aso:s	●	●	●	●	●	●	●				●		●	●	●	●		●	●	●			●	●	●	●	●	●		
0.0083	grin.as:s	●	●	●	●	●	●	●	●		●	●	●	●		●	●	●			●	●	●	●	●	●	●			●	●
0.0083	answer.as:s	●	●	●	●	●	●	●	●					●	●	●	●	●	●	●	●		●		●	●		●		●	●
0.0081	explain.as:s		●		●	●			●	●	●		●	●	●	●	●	●	●		●	●	●	●				●	●	●	●
0.0079	frown.as:s	●	●		●	●				●	●	●	●				●	●	●	●	●	●	●	●	●					●	●
0.0076	hesitate.as:s	●	●		●	●				●	●		●		●	●	●	●	●		●	●	●		●	●		●			●
0.0074	stand.as:s	●	●	●	●	●	●	●		●		●	●		●	●	●		●	●	●	●	●		●	●		●			
0.0066	continue.as:s	●	●	●	●		●	●	●	●	●	●	●		●	●	●	●	●	●	●			●				●	●	●	
0.0065	find.aso:s	●	●		●	●			●		●	●	●	●		●	●	●	●	●	●	●		●		●		●	●		●
0.0064	feel.as:s	●	●	●	●	●		●	●		●	●		●			●	●	●	●	●	●	●	●	●	●		●	●		●
0.0062	sit.as:s	●	●	●	●	●	●	●	●			●	●	●		●	●	●	●	●	●	●		●				●	●		●
0.0062	agree.as:s	●	●	●	●		●		●	●	●	●	●	●		●		●	●	●	●	●	●	●		●	●	●	●		●
0.0056	cry.as:s	●	●	●		●		●		●	●	●	●		●				●	●	●	●		●		●					

Figure 5.1: English class 5: communicative action

- a set $X(y) = \{x \in \mathcal{X} \mid x = (c, y),\ c \in C\}$ of complete data related to the observation y,
- a complete-data specification $p_\theta(x)$, corresponding to the joint probability $p(c, v, n)$ over $C \times V \times N$, with parameter-vector $\theta = \langle \theta_c, \theta_{vc}, \theta_{nc} | c \in C, v \in V, n \in N \rangle$,
- an incomplete data specification $p_\theta(y)$ which is related to the complete-data specification as the marginal probability $p_\theta(y) = \sum_{X(y)} p_\theta(x)$.

The EM algorithm is directed at finding a value $\hat{\theta}$ of θ that maximizes the incomplete-data log-likelihood function L as a function of θ for a given sample $\mathcal{Y}$, i.e.,

$$\hat{\theta} = \arg\max_{\theta} L(\theta) \text{ where } L(\theta) = \ln \prod_{y} p_\theta(y).$$

As prescribed by the EM algorithm, the parameters of $L(\theta)$ are estimated indirectly by proceeding iteratively in terms of complete-data estimation for the auxiliary function $Q(\theta; \theta^{(t)})$, which is the conditional expectation of the complete-data log-likelihood $\ln p_\theta(x)$ given the observed data y and the current fit of the parameter values $\theta^{(t)}$ (E-step). This auxiliary function is iteratively maximized as a function of θ (M-step), where each iteration is defined by the map

$$\theta^{(t+1)} = M(\theta^{(t)}) = \arg\max_{\theta} Q(\theta; \theta^{(t)})$$

Note that our application is an instance of the EM-algorithm for context-free models (Baum et al. 1970; Baker 1979), from which the following particularly simple reestimation formulae can be derived. Let $x = (c, y)$ for fixed c and y, and $f(y)$ be the frequency of y in the training sample. Then

$$\begin{aligned} M(\theta_{vc}) &= \frac{\sum_{y \in \{v\} \times N} f(y) p_\theta(x|y)}{\sum_y f(y) p_\theta(x|y)}, \\ M(\theta_{nc}) &= \frac{\sum_{y \in V \times \{n\}} f(y) p_\theta(x|y)}{\sum_y f(y) p_\theta(x|y)}, \\ M(\theta_c) &= \frac{\sum_y f(y) p_\theta(x|y)}{|\mathcal{Y}|}. \end{aligned}$$

Intuitively, the conditional expectation of the number of times a particular v, n, or c choice is made during the derivation is prorated by the conditionally expected total number of times a choice of the same kind is made. As shown by Baum et al. (1970), every such maximization step increases the log-likelihood function L, and a sequence of re-estimates eventually converges to a (local) maximum of L.

5.3 Clustering Examples

In the following, we will present some examples of induced clusters. In one experiment the input to the clustering algorithm was a training corpus of 1,178,698 tokens (608,850 types) of English verb-noun pairs participating in the grammatical relations of intransitive and transitive verbs and their subject and object fillers. The data were gathered from the maximal-probability parses the head-lexicalized probabilistic context-free grammar of Carroll and Rooth (1998) gave for the British National Corpus (117 million words).

Fig. 5.1 shows an induced semantic class out of a model with 35 classes. At the top are listed the 30 most probable nouns in the $p(n|5)$ distribution and their probabilities, and at left are the 30 most probable verbs in the $p(v|5)$ distribution where 5 is the class index. Those verb-noun pairs which were seen in the training data appear with a dot in the class matrix. Verbs with suffix $.as : s$ indicate the subject slot of an active intransitive. Similarily $.aso : s$ denotes the subject slot of an active transitive, and $.aso : o$ denotes the object slot of an active transitive. Thus v in the above discussion actually consists of a combination of a verb with a subcat frame slot $as : s$, $aso : s$, or $aso : o$. Induced classes often have a basis in

Class 17 PROB 0.0265		0.0379	0.0315	0.0313	0.0249	0.0164	0.0143	0.0110	0.0109	0.0105	0.0103	0.0099	0.0091	0.0089	0.0088	0.0082	0.0077	0.0073	0.0071	0.0071	0.0070	0.0068	0.0067	0.0065	0.0065	0.0058	0.0057	0.0057	0.0054	0.0051	0.0050
		number	rate	price	cost	level	amount	sale	value	interest	demand	chance	standard	share	risk	profit	pressure	income	performance	benefit	size	population	proportion	temperature	tax	fee	time	power	quality	supplely	money
0.0437	increase.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
0.0392	increase.aso:o		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●		●	●	●	●	●
0.0344	fall.as:s	●	●	●	●	●	●	●	●		●	●	●	●		●	●	●	●			●	●	●	●		●		●	●	
0.0337	pay.aso:o	●	●	●		●	●		●	●	●			●				●	●	●			●			●	●				●
0.0329	reduce.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
0.0257	rise.as:s	●	●	●	●	●	●	●	●		●		●	●	●	●	●	●		●	●	●	●	●	●		●	●	●	●	
0.0196	exceed.aso:o	●	●	●	●	●	●		●	●	●	●	●			●		●		●	●	●	●		●	●	●	●	●	●	
0.0177	exceed.aso:s	●	●	●	●	●	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●		●			●	●
0.0169	affect.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●		●	●	●	●	●	
0.0156	grow.as:s	●						●	●	●	●			●	●	●	●	●		●	●	●	●					●			●
0.0134	include.aso:s	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●		●	●			●	●	●	●	●	●
0.0129	reach.aso:s	●	●	●	●	●	●	●		●	●			●		●	●		●	●		●	●	●			●	●	●	●	●
0.0120	decline.as:s	●	●	●	●	●		●	●	●	●		●	●	●	●					●	●	●	●					●	●	
0.0102	lose.aso:o	●	●		●		●	●	●	●		●	●	●	●	●	●	●		●		●	●	●		●	●	●	●		●
0.0099	act.aso:s		●	●				●					●							●					●			●		●	
0.0099	improve.aso:o	●	●			●			●			●	●	●		●			●								●	●	●		
0.0088	include.aso:o	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
0.0088	cut.aso:o	●	●	●	●	●		●	●			●	●		●	●		●		●					●	●	●	●			
0.0080	show.aso:o	●	●		●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●				●	●	●		
0.0078	vary.as:s	●	●	●	●	●	●		●	●	●		●						●		●		●	●			●		●		
0.0072	give.aso:o	●	●	●	●	●	●		●	●		●	●	●	●	●	●	●	●	●	●	●	●	●			●	●	●		●
0.0071	carry.aso:o	●	●	●	●	●	●			●			●		●			●		●		●	●		●	●		●	●		●
0.0068	improve.as:s	●	●	●		●		●	●			●	●			●			●								●	●	●		
0.0066	have.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●		●	●	●	●	●	●	●
0.0066	produce.aso:o	●	●		●	●	●	●	●	●	●		●	●		●	●	●	●	●	●		●				●	●	●	●	●
0.0066	get.aso:o	●	●	●	●	●	●	●	●	●	●		●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
0.0064	raise.aso:o	●		●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●				●	●				●	●	●
0.0063	mean.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●		●		●	●	●	●	●
0.0062	receive.aso:o	●		●	●	●	●		●	●	●	●	●	●		●		●	●	●		●			●	●	●	●	●	●	●
0.0058	stand.aso:o		●									●			●												●				●

Figure 5.2: English class 17: scalar change

lexical semantics; class 5 can be interpreted as clustering agents, denoted by proper names, "man", and "woman", together with verbs denoting *communicative action*. Fig. 5.2 shows a cluster involving verbs of *scalar change* and things which can move along scales. Fig. 5.9 can be interpreted as involving different *dispositions* and modes of their execution.

In another experiment, we extracted 418,290 tokens (318,086 types) of pairs of German verbs or adjectives and grammatically related nouns from maximal-probability parses; the corpus parsed was a 4.1 million word corpus of 450,000 German subordinate clauses extracted from a 200 million word corpus of German newspapers. The lexicalized statistical model for German is described in Beil et al. (1998).

The figures 5.8 and 5.3 show two classes out of a model with 35 classes. On the left and at the top are listed the 30 highest probable verb/adjective predicates and nouns appearing as fillers of the verb/adjective slots, ordered according to their probability given the class. Verbal predicates are annotated with subcategorization slots, e.g., *liegen-VPA.np:NP.Nom* denotes the nominative noun-phrase filler of the subject-slot of an active verb *liegen* subcategorizing for a nominative and a prepositional phrase. *tragen-VPA.na:NP.Akk* is the accusative noun-

Class 26 PROB 0.0223		0.0379	0.0226	0.0182	0.0171	0.0137	0.0133	0.0129	0.0086	0.0079	0.0079	0.0073	0.0072	0.0071	0.0067	0.0064	0.0063	0.0061	0.0057	0.0056	0.0053	0.0051	0.0047	0.0046	0.0039	0.0038	0.0037	0.0037	0.0036	0.0036	0.0035
		Ergebnis	Preis	Menge	Anteil	St"uck	Zahl	Gewinn	Kritiker	B"urgerMeister	Angst	Umsatz	Einnahme	Zins	Schicksal	Bus	NachFrage	Ertrag	Figur	Verlust	Frucht	ArbeitsLosigkeit	Kost	Last	WahlErgebnis	Temperatur	Solidarit"at	InflationsRate	Abschlu"s	Import	unterSchrift
0.0601	liegen-VPA.np:NP.Nom	●	●	●	●	●	●	●		●				●	●			●		●			●			●		●			
0.0351	tragen-VPA.na:NP.Akk			●	●									●						●	●		●	●							
0.0230	steigen-VPA.n:NP.Nom		●	●	●		●					●	●	●			●			●		●				●		●			
0.0213	sagen-VPA.n:NP.Nom								●	●																					
0.0182	gering-ADJ		●	●	●		●						●	●				●		●		●	●			●		●			
0.0135	sinken-VPA.n:NP.Nom		●		●			●					●	●			●	●								●					
0.0135	steigen-VPA.np:NP.Nom				●		●										●	●		●		●				●					
0.0119	erkl"aren-VPA.n:NP.Nom									●																					
0.0100	positiv-ADJ	●																●													
0.0091	zur"uck+gehen-VPA.np:NP.Nom				●		●	●				●	●					●										●		●	
0.0087	sinken-VPA.np:NP.Nom				●		●					●		●												●	●			●	
0.0082	sp"at-ADJ												●					●													
0.0077	wirken-VPA.n:NP.Nom																														
0.0071	"andern-VPA.na:NP.Akk														●																
0.0071	regional-ADJ																														
0.0067	Erfolgreich-ADJ																												●		
0.0065	best"atigen-VPA.n:NP.Nom									●	●																				
0.0057	steigen-ADJ		●				●	●				●		●			●					●	●					●			
0.0056	an+steigen-VPA.np:NP.Nom		●		●		●					●										●				●		●			
0.0054	formulieren-VPA.na:NP.Nom									●																					
0.0052	b"ose-ADJ														●																
0.0051	an+steigen-VPA.n:NP.Nom				●			●									●					●	●			●					
0.0051	sitzen-VPA.n:NP.Nom										●																				
0.0049	"ubersteigen-VPA.na:NP.Nom	●	●		●		●					●	●	●			●			●			●					●			
0.0045	ein+setzen-VPP.n:NP.Nom															●															
0.0045	an+erkennen-VPA.na:NP.Akk	●																					●		●				●		
0.0045	entdecken-VPA.nap:NP.Akk										●								●												
0.0043	zu+nehmen-VPA.np:NP.Nom	●			●						●	●	●				●					●								●	
0.0043	betrachten-VPA.na:NP.Akk						●																								
0.0043	senken-VPP.n:NP.Nom		●	●			●							●			●											●			

Figure 5.3: German class 26: scalar change

phrase filler of the object slot of the transitive verb *tragen*, *steigen-VPA.n:NP.Nom* denotes the nominative filler of the subject slot of the intransitive verb *steigen*. Clearly, due to the smaller size of the German input data compared to the English data, German classes are less dense than the English counterparts.

Fig. 5.8 shows a class which can be interpreted as *govermental/public authority*, involving nouns such as *police force* and *public prosecutor's office*. Fig. 5.3 shows a cluster involving scalar motion verbs and things which can move along scales.

5.4 Evaluation of Clustering Models

5.4.1 Pseudo-Disambiguation

We evaluated our clustering models on a pseudo-disambiguation task similar to that performed in Pereira et al. (1993), but differing in detail. The task is to judge which of two verbs v and v' is more likely to take a given noun n as its argument where the pair (v, n) has been cut out of the original corpus and the pair (v', n) is constructed by pairing n with a randomly chosen verb v' such that the combination (v', n) is completely unseen. Thus this test evaluates how well the models generalize over unseen verbs.

The data for this test were built as follows. We constructed an evaluation corpus of (v, n, v') triples from a test corpus of 3,000 types of (v, n) pairs which were randomly cut out of the original corpus of 1,280,712 tokens, leaving a training corpus of 1,178,698 tokens. Each noun n in the test corpus was combined with a verb v' which was randomly chosen according to its frequency such that the pair (v', n) did appear neither in the training nor in the test corpus. However, the elements v, v', and n were required to be part of the training corpus. Furthermore, we restricted the verbs and nouns in the evaluation corpus to the ones which occurred at least 30 times and at most 3,000 times with some verb-functor v in the training corpus. The resulting 1,337 evaluation triples were used to evaluate a sequence of clustering models trained from the training corpus.

The clustering models we evaluated were parameterized in starting values of the training algorithm, in the number of classes of the model, and in the number of iteration steps, resulting in a sequence of 3 × 10 × 6 models. Starting from a lower bound of 50 % for randomly initialized models, accuracy was calculated as the number of times the model decided for $p(n|v) \geq p(n|v')$ out of all choices made. Fig. 5.4 shows the evaluation results for models trained with 50 iterations, averaged over starting values, and plotted against class cardinality. Different starting values had an effect of ± 2 % on the performance of the test. We obtained a value of about 80 % accuracy for models between 25 and 100 classes. Models with more than 100 classes show a small but stable overfitting effect.

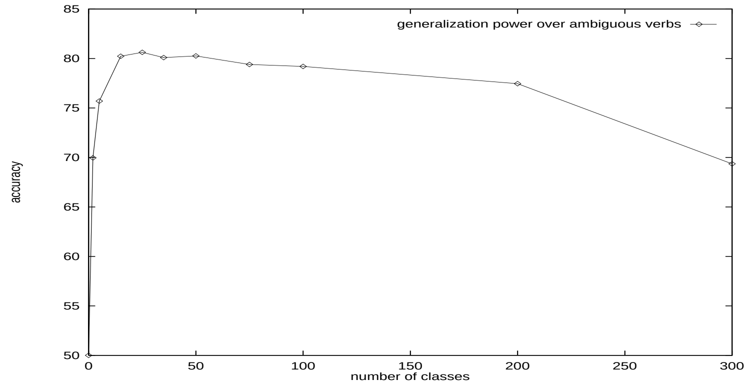

Figure 5.4: Evaluation of English models on pseudo-disambiguation task

The German models were evaluated in a similar way. An evaluation corpus of 886 (v, n, v') triples was extracted from the original corpus of 428,446 verb/adjective-noun tokens, leaving 418,290 tokens for training a sequence of clustering models. Again, the models were parameterized in starting values, number of classes and iteration steps, resulting in a sequence of $3 \times 11 \times 20$ models. Fig. 5.5 shows the evaluation results for models trained with 100 iterations, averaged over starting values, and plotted against class cardinality. We obtained an accuracy of over 75 % for models up to 35 classes. Different starting values had an effect of $\pm$ 2 % on the evaluation results. For models with more than 50 classes again a small overfitting effect can be seen.

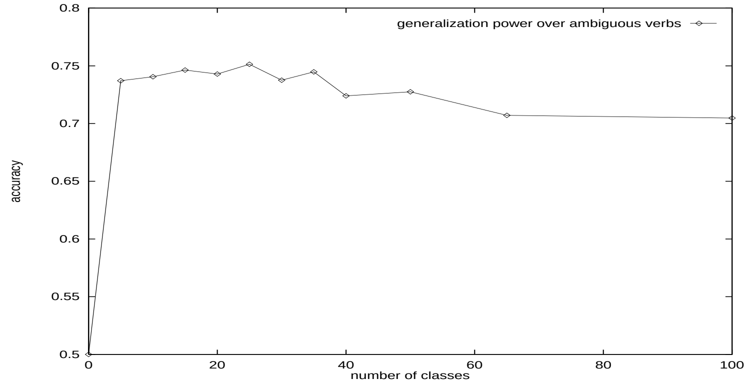

Figure 5.5: Evaluation of German models on pseudo-disambiguation task

5.4.2 Smoothing Power

A second experiment addressed the smoothing power of the model by counting the number of (v, n) pairs in the set $V \times N$ of all possible combinations of verbs and nouns which received

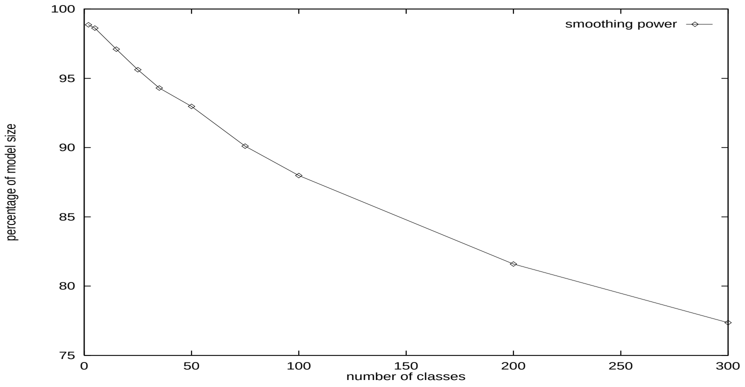

Figure 5.6: Evaluation of English models on smoothing task

a positive joint probability by the model. The $V \times N$-space for the above clustering models included about 425 million (v, n) combinations; we approximated the smoothing size of a model by randomly sampling 1,000 pairs from $V \times N$ and returning the percentage of positively assigned pairs in the random sample. Fig. 5.6 plots the smoothing results for the above models against the number of classes. Starting values had an influence of ± 1 % on performance. Given the proportion of the number of types in the training corpus to the $V \times N$-space, without clustering we have a smoothing power of 0.14 % whereas for example a model with 50 classes and 50 iterations has a smoothing power of about 93 %.

Corresponding to the maximum likelihood paradigm, the number of training iterations had a decreasing effect on the smoothing performance whereas the accuracy of the pseudo-disambiguation was increasing in the number of iterations. We found a number of 50 iterations to be a good compromise in this trade-off.

For German models we observed a baseline smoothing power of 0.012 % which is the relation of the number of types in the German training corpus to the 2.5 billion combinations in the $V \times N$-space for the German experiments. Despite of the fact that this baseline is 10 times smaller than the baseline for the English models, we have a smoothing power of about 32 % for models with 25 classes, which were best in terms of the pseudo-disambiguation task. This is shown in Fig. 5.7. The best compromise in terms of iterations was a number of 100 iterations for the German experiments.

5.5 Application to Lexicon Induction Based on Latent Classes

5.5.1 Motivation

An important challenge in computational linguistics concerns the construction of large-scale computational lexicons for the numerous natural languages where very large samples of lan-

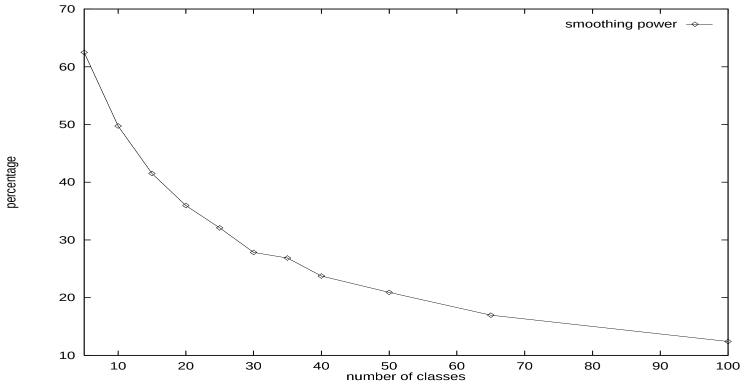

Figure 5.7: Evaluation of German models on smoothing task

guage use are now available. Resnik (1993) initiated research into the automatic acquisition of semantic selectional restrictions. Ribas (1994) presented an approach which takes into account the syntactic position of the elements whose semantic relation is to be acquired. However, those and most of the following approaches require as a prerequisite a fixed taxonomy of semantic relations. This is a problem because (i) entailment hierarchies are presently available for few languages, and (ii) we regard it as an open question whether and to what degree existing designs for lexical hierarchies are appropriate for representing lexical meaning. Both of these considerations suggest the relevance of inductive and experimental approaches to the construction of lexicons with semantic information.

This section presents a method for automatic induction of semantically annotated subcategorization frames from unannotated corpora. We use a statistical subcat-induction system which estimates probability distributions and corpus frequencies for pairs of a head and a subcat frame (Carroll and Rooth 1998). The statistical parser can also collect frequencies for the nominal fillers of slots in a subcat frame. The induction of labels for slots in a frame is based upon estimation of a probability distribution over tuples consisting of a class label, a selecting head, a grammatical relation, and a filler head. The class label is treated as hidden data in the EM-framework for statistical estimation.

5.5.2 Probabilistic Labeling with Latent Classes using EM-estimation

To induce latent classes for the subject slot of a fixed intransitive verb the following statistical inference step was performed. Given a latent class model $p_{LC}(\cdot)$ for verb-noun pairs, and a sample $n_1, \ldots, n_M$ of subjects for a fixed intransitive verb, we calculate the probability of an arbitrary subject $n \in N$ by:

$$p(n) = \sum_{c \in C} p(c, n) = \sum_{c \in C} p(c) p_{LC}(n|c).$$

Class 14 PROB 0.0283		0.0877	0.0303	0.0187	0.0179	0.0161	0.0154	0.0151	0.0150	0.0130	0.0112	0.0094	0.0088	0.0086	0.0085	0.0073	0.0066	0.0066	0.0064	0.0061	0.0061	0.0059	0.0058	0.0057	0.0053	0.0051	0.0051	0.0043	0.0042	0.0041	0.0041
		Polizei	deutschLand	Nation	SPD	USA	Koalition	Sprecher	Verein	StaatsAnwaltschaft	Beh"orde	Bonn	firm	UNO	BundesAmt	Veranstalter	Blatt	Konzern	Magistrat	Sender	oberB"urgerMeister	BundesBank	B"urgerMeister	fern+sehen	Nato	Zeitung	N"ahe	nachrichtenAgentur	PolizeiSprecher	LandesRegierung	rundFunk
0.0581	mit+teilen-VPA.n:NP.Nom	●						●	●		●				●	●		●	●	●			●					●	●		●
0.0220	berichten-VPA.n:NP.Nom	●			●			●			●				●		●			●		●		●		●		●	●		●
0.0141	vereinen-ADJ		●	●																											
0.0120	sagen-VPA.n:NP.Nom	●						●								●							●								
0.0114	Bremer-ADJ	●											●							●			●							●	
0.0113	mit+teilen-VPA.np:NP.Nom			●				●		●	●								●			●	●								
0.0095	erkl"aren-VPA.n:NP.Nom	●						●	●						●						●		●						●		
0.0070	hessisch-ADJ	●			●																									●	●
0.0065	unmittelbar-ADJ																										●				
0.0064	fordern-VPA.na:NP.Nom		●		●	●			●			●							●							●				●	
0.0063	machen-VPA.nad:NP.Nom	●			●	●			●	●									●												
0.0061	verzichten-VPA.np:NP.Nom		●		●	●	●				●			●											●					●	
0.0060	melden-VPA.n:NP.Nom							●							●									●		●		●			●
0.0059	Berliner-ADJ	●							●	●	●									●						●				●	
0.0055	spielen-VPA.nap:NP.Nom		●			●																									
0.0052	Westdeutsch-ADJ												●					●													●
0.0046	statistisch-ADJ														●																
0.0043	meinen-VPA.n:NP.Nom					●		●											●				●								
0.0043	auf+nehmen-VPA.nap:NP.Nom	●	●		●	●				●		●		●																	
0.0042	w"unschen-VPA.nar:NP.Akk																														
0.0041	vor+stellen-VPA.nar:NP.Nom				●		●		●													●			●					●	
0.0041	haben-VPA.nap:NP.Nom	●	●		●	●	●				●				●	●						●		●	●					●	
0.0041	betonen-VPA.n:NP.Nom	●														●					●		●								
0.0040	zust"andig-ADJ									●	●												●								
0.0039	"ubernehmen-VPA.na:NP.Nom	●			●	●	●		●		●	●		●				●	●												
0.0038	nieders"achsisch-ADJ	●			●																									●	
0.0036	"ortlich-ADJ	●							●		●															●					
0.0035	best"atigen-VPA.n:NP.Nom							●		●	●										●		●						●		
0.0035	erz"ahlen-VPA.n:NP.Nom							●																							
0.0033	ein+treffen-VPA.n:NP.Nom	●																								●					

Figure 5.8: German class 14: governmental/public authority

The estimation of the parameter-vector $\theta = \langle \theta_c | c \in C \rangle$ can be formalized in the EM framework by viewing $p(n)$ or $p(c,n)$ as a function of θ for fixed $p_{LC}(.)$. The re-estimation formulae resulting from the incomplete data estimation for these probability functions have the following form ($f(n)$ is the frequency of n in the sample of subjects of the fixed verb):

$$M(\theta_c) = \frac{\sum_{n \in N} f(n) p_\theta(c|n)}{\sum_{n \in N} f(n)}$$

A similar EM induction process can be applied also to pairs of nouns, thus enabling induction of latent semantic annotations for transitive verb frames. Given a LC model $p_{LC}(\cdot)$ for verb-noun pairs, and a sample $(n_1, n_2)_1, \ldots, (n_1, n_2)_M$ of noun arguments (n_1 subjects, and n_2 direct objects) for a fixed transitive verb, we calculate the probability of its noun argument pairs by:

$$\begin{aligned} p(n_1, n_2) &= \sum_{c_1, c_2 \in C} p(c_1, c_2, n_1, n_2) \\ &= \sum_{c_1, c_2 \in C} p(c_1, c_2) p_{LC}(n_1|c_1) p_{LC}(n_2|c_2) \end{aligned}$$

Again, estimation of the parameter-vector $\theta = \langle \theta_{c_1 c_2} | c_1, c_2 \in C \rangle$ can be formalized in an EM framework by viewing $p(n_1, n_2)$ or $p(c_1, c_2, n_1, n_2)$ as a function of θ for fixed $p_{LC}(.)$. The re-estimation formulae resulting from this incomplete data estimation problem have the following simple form ($f(n_1, n_2)$ is the frequency of (n_1, n_2) in the sample of noun argument pairs of the fixed verb):

$$M(\theta_{c1c2}) = \frac{\sum_{n_1, n_2 \in N} f(n_1, n_2) p_\theta(c_1, c_2 | n_1, n_2)}{\sum_{n_1, n_2 \in N} f(n_1, n_2)}$$

Note that the class distributions $p(c)$ and $p(c_1, c_2)$ for intransitive and transitive models can also be computed for verbs unseen in the LC model.

5.5.3 Lexicon Induction Experiments

In a first experiment with English data we used a model with 35 classes. From maximal probability parses for the British National Corpus derived with the statistical parser of Carroll and Rooth (1998), we extracted frequency tables for intransitive verb/subject pairs and transitive verb/subject/object triples. The 500 most frequent verbs were selected for slot labeling. Fig. 5.10 shows two verbs v for which the most probable class label is 5, a class which we earlier described as *communicative action*, together with the estimated frequencies of $f(n)p_\theta(c|n)$ for those ten nouns n for which this estimated frequency is highest.

Fig. 5.11 shows corresponding data for an intransitive scalar motion sense of *increase*.

Fig. 5.12 shows the intransitive verbs which take 17 as the most probable label. Intuitively, the verbs are semantically coherent. When compared to Levin (1993)'s 48 top-level verb classes,

Class 8 PROB 0.0369		0.0385	0.0162	0.0157	0.0101	0.0073	0.0071	0.0063	0.0060	0.0060	0.0057	0.0055	0.0052	0.0052	0.0051	0.0050	0.0050	0.0048	0.0047	0.0047	0.0046	0.0041	0.0041	0.0040	0.0040	0.0039	0.0038	0.0037	0.0037	0.0036	0.0036
		change	use	increase	development	growth	effect	result	degree	response	approach	reduction	forme	condition	understanding	improvement	treatment	skill	action	process	activity	knowledge	factor	level	type	reaction	kind	difference	movement	loss	amount
0.0539	require.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●			●	●		●
0.0469	show.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●		●	●	●	●
0.0439	need.aso:o	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●		●	●		●	●	●	●	●	●		●
0.0383	involve.aso:o	●		●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●
0.0270	produce.aso:o	●	●	●	●	●		●	●	●	●	●	●	●	●	●			●		●	●	●	●	●	●	●	●	●	●	●
0.0255	occur.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●		●			●	●	●		●	●	●	●	●	●	●	●	
0.0192	cause.aso:s	●	●	●	●	●	●	●		●		●	●	●			●		●	●	●	●	●	●	●	●	●	●	●	●	●
0.0189	cause.aso:o	●	●	●	●	●	●	●	●	●				●		●			●		●			●		●	●	●	●	●	●
0.0179	affect.aso:s	●	●	●	●	●	●	●	●	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
0.0162	require.aso:s	●	●	●	●		●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
0.0150	mean.aso:o	●	●	●	●	●	●		●		●	●	●	●	●	●	●		●	●	●	●		●	●		●	●	●	●	●
0.0140	suggest.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●		●	●	●	●	●	●	●		●	●	●		●
0.0138	produce.aso:s	●	●	●	●	●	●	●	●	●	●	●	●	●			●		●	●	●		●	●	●	●	●	●	●	●	●
0.0109	demand.aso:o	●	●	●	●			●	●	●	●		●		●	●	●	●	●		●	●		●			●				●
0.0109	reduce.aso:s	●	●	●	●	●			●	●	●	●	●				●	●	●	●	●	●	●	●					●	●	●
0.0097	reflect.aso:o	●	●	●	●	●	●		●	●	●	●	●	●	●	●		●		●	●	●	●	●	●	●	●	●	●	●	●
0.0092	involve.aso:s	●	●			●	●	●		●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●
0.0091	undergo.aso:o	●	●	●	●		●		●					●		●	●			●						●	●			●	●
0.0086	increase.aso:s	●	●	●	●	●	●	●		●	●	●				●	●	●	●	●		●	●	●	●				●	●	●
0.0081	allow.aso:o	●	●	●	●	●			●		●	●			●	●	●		●	●				●			●		●	●	●
0.0079	include.aso:o			●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●		●	●	●	●	●
0.0075	make.aso:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●
0.0075	encourage.aso:o	●	●	●	●	●			●	●	●	●			●	●		●		●	●			●	●		●		●	●	●
0.0073	saw.aso:o	●	●	●	●	●	●	●	●		●	●	●	●	●	●			●					●	●	●	●	●	●		●
0.0072	create.aso:s	●	●	●	●	●	●	●		●	●	●	●	●		●			●	●	●		●	●	●	●	●	●	●	●	
0.0070	affect.aso:o	●	●		●	●	●	●	●	●	●	●	●	●	●				●	●	●			●	●	●	●	●	●		●
0.0069	imply.aso:o	●	●	●		●			●		●	●	●	●	●	●			●	●		●		●	●		●	●			●
0.0068	achieve.aso:o	●	●	●		●	●	●	●	●		●	●		●	●		●		●				●		●	●		●	●	●
0.0066	find.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●
0.0062	describe.aso:o	●	●	●	●	●	●	●		●	●		●	●				●	●	●	●	●	●	●	●	●	●	●	●	●	●

Figure 5.9: English class 8: dispositions

we found an agreement of our classification with her class of "verbs of changes of state" except for the last three verbs in the list in Fig. 5.12 which is sorted by probability of the class label.

Fig. 5.13 shows the most probable pair of classes for *increase* as a transitive verb, together with estimated frequencies for the head filler pair. Note that the object label 17 is the class found with intransitive scalar motion verbs; this correspondence is exploited in the next section.

Further experiments were done with two German models with 35 and 50 classes respectively. The data for these experiments were extracted from the maximal probability parses of a 4.1 million word corpus of German subordinate clauses, parsed with the lexicalized probabilistic grammar of Beil et al. (1998). Fig. 5.14 shows the subjects of the intransitive verb *bekanntgeben* (make public). The nouns are classified with probability 0.999999 to class 14, which was described in Sect. 5.3 as class of *govermental/public-authority*. The numbers in the column show the estimated frequencies of the subject fillers.

Fig. 5.15 shows the subjects of the intransitive verb *steigen* (rise) which belong with

blush 5	0.982975	*snarl* 5	0.962094
constance	3	mandeville	2
christina	3	jinkwa	2
willie	2.99737	man	1.99859
ronni	2	scott	1.99761
claudia	2	omalley	1.99755
gabriel	2	shamlou	1
maggie	2	angalo	1
bathsheba	2	corbett	1
sarah	2	southgate	1
girl	1.9977	ace	1

Figure 5.10: Lexicon entries: *blush*, *snarl*

increase 17	0.923698		
number	134.147	proportion	23.8699
demand	30.7322	size	22.8108
pressure	30.5844	rate	20.9593
temperature	25.9691	level	20.7651
cost	23.9431	price	17.9996

Figure 5.11: Scalar motion *increase*.

probability 0.67273 to class 26 which was interpreted in Sect. 5.3 as a class of *gradation/scalar change*.

Similar to the English experiments we observe semantic uniformity in the verbs of scalar change. Fig. 5.16 shows 10 intransitive verbs which take class 14 of a 50-classes model (corresponding to class 26 of the 35-class model) as the most probable class to label their respective subject slots. On the basis the most probable class labels these verbs can be summarized as scalar motion verbs. When compared to linguistic classifications of verbs given by Schuhmacher (1986), we found an agreement of our classification with the class of “einfache Änderungsverben” (simple verbs of change) except for the verbs *anwachsen* (increase) and *stagnieren* (stagnate) which were not classified there at all.

An example of the two most probable subject-object class pairs of a transitive verb, *senken* (lower) is shown in Fig. 5.17. Class 14 has been introduced before as *govermental/public authority* and class 26 as *gradation/scalar change*.

Fig. 5.18 shows the transitive verb *dauern* (last/go on) selecting the class-pair $(0, 10)$ with probability 0.957095 as semantic label for its subject and object slots. Class 0 can be interpreted as *project/action-class* and class 10 as class of *time*.

0.977992 decrease	0.560727 drop
0.948099 double	0.476524 grow
0.923698 increase	0.42842 vary
0.908378 decline	0.365586 improve
0.877338 rise	0.365374 climb
0.876083 soar	0.292716 flow
0.803479 fall	0.280183 cut
0.672409 slow	0.238182 mount
0.583314 diminish	

Figure 5.12: Scalar motion verbs

increase $(8, 17)$	0.3097650
development - pressure	2.3055
fat - risk	2.11807
communication - awareness	2.04227
supplementation - concentration	1.98918
increase - number	1.80559

Figure 5.13: Transitive *increase* with estimated frequencies for filler pairs.

5.5.4 Linguistic Interpretation of Latent Class Labels

In some linguistic accounts, multi-place verbs are decomposed into representations involving (at least) one predicate or relation per argument. For instance, the transitive causative/inchoative verb *increase,* is composed of an actor/causative verb combining with a one-place predicate in the structure on the left in Fig. 5.19. Linguistically, such representations are motivated by argument alternations (diathesis), case linking and deep word order, language acquistion, scope ambiguity, by the desire to represent aspects of lexical meaning, and by the fact that in some languages, the postulated decomposed representations are overt, with each primitive predicate corresponding to a morpheme. For references and recent discussion of this kind of theory see Hale and Keyser (1993) and Kural (1996).

We will sketch an understanding of the lexical representations induced by latent-class labeling in terms of the linguistic theories mentioned above, aiming at an interpretation which combines computational learnability, linguistic motivation, and denotational-semantic adequacy. The basic idea is that latent classes are computational models of the atomic relation symbols occurring in lexical-semantic representations. As a first implementation, consider replacing the relation symbols in the first tree in Fig. 5.19 with relation symbols derived from the latent class labeling. In the second tree in Fig 5.19, R_{17} and R_8 are relation symbols with indices derived from the labeling procedure of Sect. 5.5. Such representations can be semantically interpreted in standard ways, for instance by interpreting relation symbols as denoting

bekanntgeben 14	0.999999	(make public)
Sprecher	4	(spokesman)
Polizei	3	(police)
BundesAmt	3	(Federal Agency)
BürgerMeister	2	(mayor)
VorstandsChef	2	(Chairman of the board)
GeschäftsLeitung	2	(manager)
Vorstand	2	(board of management)
unternehmen	1.99996	(company)
WetterAmt	1	(meteorological office)
VolksBank	1	(cooperative bank)

Figure 5.14: Intransitive lexicon entry: *bekanntgeben* (make public)

steigen 26	0.67273	(rise)
Zahl	23.333	(number)
Preis	15.895	(price)
ArbeitsLosigkeit	10.8788	(unemployment)
Lohn	9.72965	(wage)
NachFrage	6.83619	(demand)
Zins	6.80322	(interest)
Auflage	5.22654	(print run)
Beitrag	4.22577	(contribution)
Produktion	4.21641	(output)
GrundstuecksPreis	4	(price of a piece of land)

Figure 5.15: Intransitive lexicon entry: *steigen* (rise)

0.741467 ansteigen	(go up)
0.720221 steigen	(rise)
0.693922 absinken	(sink)
0.656021 sinken	(go down)
0.438486 schrumpfen	(shrink)
0.375039 zurückgehen	(decrease)
0.316081 anwachsen	(increase)
0.215156 stagnieren	(stagnate)
0.160317 wachsen	(grow)
0.154633 hinzukommen	(be added)

Figure 5.16: German intransitive scalar change verbs

senken $(14, 26)$	0.450352	(lower)
BundesBank - LeitZins	5.81457	(Federal bank - base rate)
BundesBank - Zins	2.97838	(Federal bank - interest)
superMarkt - Preis	1	(super market - price)
SommerGeschäft - Verlust	1	(summer business - loss)
BundesBank - DiskontSatz	0.99999	(Federal bank - minimum lending rate)
senken $(14, 14)$	0.147857	
BundesBank - Lombardsatz	0.999973	(Federal bank - rate on loanes on security)
StrafAndrohung - AbtreibungsQuote	0.96842	(threat of punishment - abortion rate)
StrafAndrohung - AbtreibungsZahl	0.96842	(threat of punishment - number of abortions)
FachHandel - LagerKost	0.878333	(stores - storage charges)
Harmonisierung - sozialNiveau	0.764319	(harmonization - social level)

Figure 5.17: Transitive lexicon entries for *senken* (lower)

relations between events and individuals.

Such representations are semantically inadequate for reasons given in philosophical critiques of decomposed linguistic representations; see Fodor (1998) for recent discussion. A lexicon estimated in the above way has as many primitive relations as there are latent classes. We guess there should be a few hundred classes in an approximately complete lexicon (which would have to be estimated from a corpus of hundreds of millions of words or more). Fodor's arguments, which are based on the very limited degree of genuine interdefinability of lexical items and on Putnam's arguments for contextual determination of lexical meaning, indicate that the number of basic concepts has the order of magnitude of the lexicon itself. More concretely, a lexicon constructed along the above principles would identify verbs which are labelled with the same latent classes; for instance it might identify the representations of *grab* and *touch*.

For these reasons, a semantically adequate lexicon must include additional relational con-

dauern $(0, 10)$	0.957095	(last/go on)
Entwirrung - Zeit	2	(disentanglement - time)
BuergerFrageStunde - Stunde	2	(question time - hour)
Prozess - Jahr	2	(trail - year)
schreckensZeit - Jahr	1	(scaring time - year)
ratenZahlung - Jahr	1	(buy in installments - year)

Figure 5.18: Transitive lexicon entry for *dauern* (last/to go on)

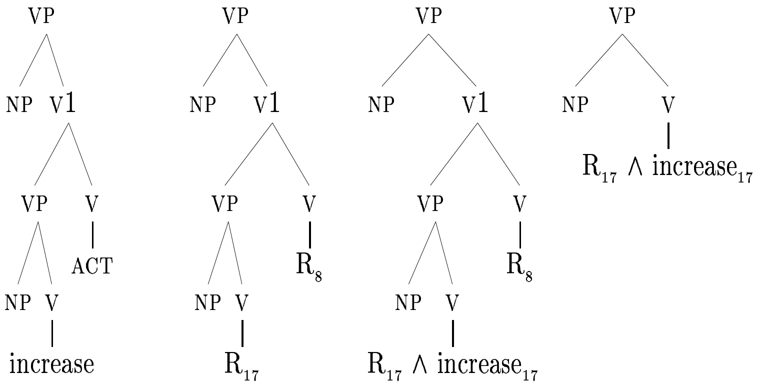

Figure 5.19: First tree: linguistic lexical entry for transitive verb *increase*. Second: corresponding lexical entry with induced classes as relational constants. Third: indexed open class root added as conjunct in transitive scalar motion *increase*. Fourth: induced entry for related intransitive *increase*.

stants. We meet this requirement in a simple way, by including as a conjunct a unique constant derived from the open-class root, as in the third tree in Fig. 5.19. We introduce indexing of the open class root (copied from the class index) in order that homophony of open class roots not result in common conjuncts in semantic representations—for instance, we don't want the two senses of *decline* exemplified in *decline the proposal* and *decline five percent* to have a common entailment represented by a common conjunct. This indexing method works as long as the labeling process produces different latent class labels for the different senses.

The last tree in Fig. 5.19 is the learned representation for the scalar motion sense of the intransitive verb *increase.* In our approach, learning the argument alternation (diathesis) relating the transitive *increase* (in its scalar motion sense) to the intransitive *increase* (in its scalar motion sense) amounts to learning representations with a common component $R_{17} \wedge \mathrm{increase}_{17}$. In this case, this is achieved.

5.5.5 Discussion

We have proposed a procedure which maps observations of subcategorization frames with their complement fillers to structured lexical entries. We believe the method is scientifically interesting, practically useful, and flexible because:

1. The algorithms and implementation are efficient enough to map a corpus of a hundred million words to a lexicon.
2. The model and induction algorithm have foundations in the theory of parameterized families of probability distributions and statistical estimation. As exemplified in the paper, learning, disambiguation, and evaluation can be given simple, motivated formulations.
3. The derived lexical representations are linguistically interpretable. This suggests the possibility of large-scale modeling and observational experiments bearing on questions arising in linguistic theories of the lexicon.
4. Because a simple probabilistic model is used, the induced lexical entries could be incorporated in lexicalized syntax-based probabilistic language models, in particular in head-lexicalized models. This provides for potential application in many areas.
5. The method is applicable to any natural language where text samples of sufficient size, computational morphology, and a robust parser capable of extracting subcategorization frames with their fillers are available.

The pseudo-disambiguation task of Sect. 5.4 provides a hard evaluation of the learned models, allowing comparison to other methods. The discussion of linguistic interpretation suggests entirely independent linguistic evaluations of the adequacy of the learned lexicon.

For instance, by linguistic criteria it is approximately uncontroversial whether a given verb with both intransitive and transitive frames has a subject/object diathesis (as e.g. *increase* does, but *decline* doesn't). This would allow for an evaluation of a procedure (like the one proposed here) which learns a lexicon with hidden linguistic structure.

5.6 Application to Target-Language Disambiguation in Translation

5.6.1 Motivation

In most transfer-based approaches to machine translation the major part of the work is done by symbolic, i.e. non-statistical translation mechanisms. Clearly, a statistical disambiguation (SD) component supporting the symbolic transfer can be very helpful in cases of failure of the symbolic disambiguation component or cases of explosion of ambiguities. Such a supporting SD component should ideally have the following features:

- **robustness**: In order to enable a disambiguation in most cases, one of the central features of SD is its ability to assign a positive though possibly very small probability to nearly every structure under consideration.
- **efficiency**: In cases of explosion of translation ambiguities, SD must facilitate quick resolution.
- **economy**: SD must be able to resolve translation ambiguities with a minimum of information delivered from the symbolic translation component.
- **domain-specificity**: SD must be able to disambiguate translation alternatives according to their use in a specific domain.
- **portability**: A SD component for machine translation systems working with more than one source/target language pair ideally should rely only on information which is automatically obtainable from monolingual corpora.
- **accuracy**: To built a sensible probability model on ambiguous structures, the probabilities assigned to translation alternatives must be gathered by a mathematically well-defined statistical inference mechanism and proven to be useful in a task-oriented evaluation procedure.

To illustrate these features, let us have a look at some recent approaches to statistical disambiguation. One of the most influential works on target word selection in machine translation is described in Dagan and Itai (1994). The SD of Dagan and Itai (1994) uses statistics exclusively on monolingual data. The information gathered is statistics on all grammatical relations in which an ambiguous lexical item participates. In an experiment on translating Hebrew to English, their statistical model was applicable to 70 out of 100 of the ambiguous words, achieving a precision of 91 %. In a second experiment on translating German to English, their statistical model was applicable to 27 out of 54 of the ambigous words, achieving a precision of 78 %. Given the relatively low recall values of 70 % and 50 %, their SD system clearly lacks robustness. The usage of rich information on grammatical relations lets the approach seem uneconomic and presumably not very efficient and portable. However, this approach achieves high precision rates on small test corpora with low averages of 3.3 alternative translations and high averages of 1.4 correct translations.

Another important approach to SD is the work of Melamed (1997). Melamed reports 42 % precision and 35-40 % recall using a gold standard for single-best word-to-word translations (French-English). The test corpus contains several thousand items and is a part of the training corpus. Melamed's best model can be viewed as a robust, economic, domain-specific, and portable SD. Unfortunately, the size of average alternative translations is not reported. Presumably, it is high. More importantly, the models were trained by using a part of the training corpus as test set and the reported recall measures are not high. Thus, the models main drawback is its lack of accuracy.

A further approach to be considered in the context of statistical disambiguation is the approach to word-sense disambiguation presented in Resnik (1997). The connection of monolingual word sense disambiguation to the task of target-language disambiguation can be made by viewing candidate translations of a given lexical item as the different "senses" expressed in terms of the target language. Resnik (1997) cited an average precision of 44.3 % for his experiments on word-sense disambiguation evaluated on verb-object pairs for a random baseline of 28.5 %. His model can be interpreted as a robust, shallow, domain-unspecific, and accurate SD, but it uses a fixed semantic network which is not available in most languages. Thus, in terms of the above classification, it's main drawback is the lack of portability.

To sum up, either the suggested SDs are not robust and economic (Dagan and Ito) or they have a low accuracy even in the domain-specific case (Melamed) or they lack portability (Resnik).

In the following, we will present a new approach to SD which uses the above described latent class models as probabilistic disambiguation device. We will show that this approach yields a robust, economic SD device that is portable and has a high accuracy even in the domain-unspecific case.

5.6.2 Statistical Disambiguation

This section is particularly concerned with the use of a statistical disambiguation component that specifies the translation mapping of ambiguous nouns. This restriction to noun ambiguity can motivated by observations in Buschbeck-Wolf (1997) who states that in contrast to verbal and modifier predicates the disambiguation of nouns is difficult since it is impossible to symbolically fix all contexts in which the one or the other translation is preferred. Thus, for this task, it seems reasonable to rely on statistical information.

Consider for example the German sentence in Fig. 5.20 which is to be translated into English. Suppose that the transfer-module will be presented with an input encoding the noun argument “Karte” to be a direct object of the verbal predicate “spielen”.

GER	Sie spielen ausserdem die Karte der britischen Regierung
ID-76627	You are playing the card UK Government
ENG	You are playing the UK Government card

Figure 5.20: German-English transfer

Furthermore, the symbolic components are supposed to include a word-to-word translation module translating “spielen” to “play” and presenting the possibilities “card”, “chart”, “map”, and “ticket” as translations of “Karte”.

The task of the statistical disambiguator is to specify which of the presented four English nouns should be taken as the proper argument of the predicate “spielen”. The idea employed in our statistical approach is very simple: In order to decide which of the four verb-noun combinations is the best one, we take the pair that is assigned the highest joint probability by a reasonable probability model on such predicate-argument relations. For the above example, the joint probabilities $p_{LC}(\mathrm{play}, \mathrm{card})$, $p_{LC}(\mathrm{play}, \mathrm{chart})$, $p_{LC}(\mathrm{play}, \mathrm{map})$, $p_{LC}(\mathrm{play}, \mathrm{ticket})$ have to be compared. The verb-noun pair with the highest probability will then be taken as the output of the statistical disambiguator. In case there is no unique maximum in this comparison, the system yields the output “don’t-know”.

5.6.3 Disambiguation Experiments

The following experiments were performed with the English clustering models described in Sect. 5.3.

Evaluation on a Pseudo-Disambiguation Task

We evaluated our clustering models on a pseudo-disambiguation task similar to the one described in Sect 5.5, but specified to noun ambiguity. The task is to judge which of two nouns

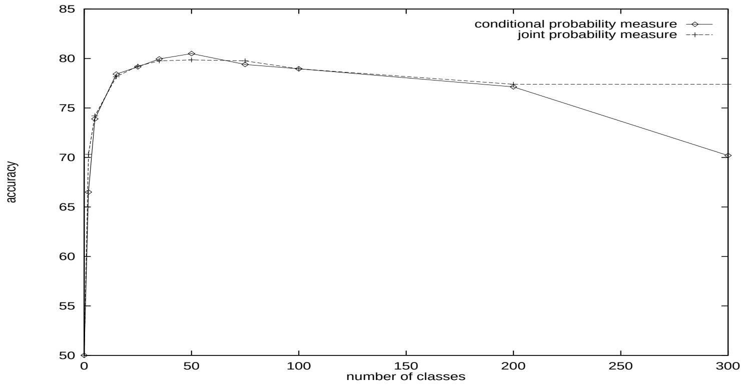

Figure 5.21: Evaluation on pseudo-disambiguation task

n and n' is more likely to appear as argument of a given verb v where the pair (v, n) has been cut out of the original training data and the pair (v, n') is constructed by pairing v with a randomly chosen noun n'. The data were built by constructing an evaluation corpus of (v, n, n') triples from a test corpus of 3,000 (v, n) pairs which was randomly cut out of the original corpus of 1,280,715 tokens. Each verb n in the test corpus was combined with a noun n' which was randomly chosen according to its frequency such that the pair (v, n') did not appear in the training and in the test corpus. Again, the elements v, n', and n had to be part of the reduced training corpus. As above, we restricted the verbs and nouns in the evaluation corpus to the ones which occurred with at least 30 and at most 3000 partners in the original training corpus. The reduced training corpus of 1,178,698 tokens was used to train a sequence of clustering models which were evaluated on the resulting 1,685 evaluation triples.

Clustering models were parametrized in starting values of EM-training, in the number of classes of the model (up to 300), and in the number of iteration steps (up to 50), resulting in a sequence of $3 \times 10 \times 6$ models. Accuracy was calculated as the number of times the model decided for $p(v, n) \geq p(v, n')$ (joint probability measure) resp. $p(v|n) \geq p(v|n')$ (conditional probability measure) out of all choices made. As shown in Fig. 5.21, we obtained an accuracy of ca. 79 % for models between 35 and 100 classes, averaged over different starting values. For models with more than 100 classes we see a small but stable overfitting effect.

Evaluation on a Smoothing Task

Recall the experiments on the smoothing power of LC models as described in Sect. 5.4. There we reported a proportion of the number of types in the English training corpus to the $V \times N$-space of 0.14 % without clustering compared to circa 93 % for clustering models with 50 classes and 50 iterations. As will be shown below, this remarkable increase in smoothing power is one of the key features of our approach to cluster-based SD.

Evaluation on a Golden Standard

To evaluate the output of the statistical disambiguator, we prepared an evaluation corpus from the sentence-aligned debates of the European parliament (mlcc = multilingual corpus for cooperation). Each language is represented by ca. 9 million tokens. The direction of translation was German to English.

angriff	aggression, assault, offence, offense, onset, onslaught, attack , charge, offensive, raid, whammy, inroad
art	fashion, fit, kind, wise, manner, type, species, mode, sort, variety, form, way
aufgabe	abandonment, task, exercise, lesson, giveup, job , office, problem, tax
auswahl	choose, eligibility, selection, choice, choosing, varity, assortment, extract, range, sample
begriff	concept, item, notion, idea
boden	floor, soil, bottom, ground, land
einrichtung	arrangement, constitution, establishment, feature, installation, institution, construction, setup, adjustment, composition, organization
erweiterung	amplification, enhancement, expansion, extension, extention, upgrade, dilatation, dilation, upgrading, add-on, increment
fehler	blemish, blunder, bug, defect, demerit, error, fail, failure, fault, flaw, mistake, shortcom, shortcoming, trouble, slip, blooper, lapse, lapsus
genehmigung	permission, approval, consent, acceptance, approbation, authorization
geschichte	history, story, tale, saga, strip
gesellschaft	companion, companionship, society, company, party, associate
grenze	boundary, frontier, limit, border, periphery, borderline, edge
grund	base, cause, ground, master, matter, reason, bottom root
karte	card, map, ticket, chart
lage	site, situation, position, bearing, layer, tier
mangel	deficiency, fault, lack, privation, scarcity, want, shortage, shortcoming, absence, dearth, demerit, desideration, desideratum, insufficiency, paucity, scarceness
menge	crowd, lot, mass, multitude, plenty, quantity, quantum , quiverful, volume, abundance, amount, aplenty, assemblage , batch, crop, deal, heap, lashings, scores, set, loads, bulk
pruefung	examination, ordeal, scrutiny, test, trial, inspection, exam, testing, tryout, verification, assay, canvass, check, checkup, inquiry, perusal, reconsideration, scruting, exa
schwierigkeit	difficulty, problem, severity, trouble, ardousness, heaviness
seite	page, side, point, aspect, party
sicherheit	certainty, certitude, immunity, safety, security , collateral , secureness, doubtlessness, sureness, guarantee, guaranty, deposit
stimme	voice, vote, tones
termin	date, appointment, meeting, time, term, deadline
verbindung	chain, conjunction, connection, connexion, fusion, incorporation, interconnection, joint , junction, link, compound, alliance , catenation, tie, union, bond, chaining, association, interface, join, liaison, contact, linkage, liaise, touch, relation
verbot	ban, interdiction, prohibition, forbade, forbad, forbiddance
verpflichtung	commitment, committal, duty, obligation, indebtedness , onus, bond, debt, duty, engagement, liability, undertaking
vertrauen	confidence, faith, reliance, trust, confidentialness, trustfulness, assurance, dependence, private, secret
wahl	choice , election, option, ballot, electoral, alternative, poll list
weg	lane, road, way, alley, route, path
widerstand	resistance, resistor, opposition, drag, resistivity
zeichen	char, character, icon, sign, signal, symbol, mark, token, figure, omen
ziel	aim, designation, destination, target, end, goal, object, objective, sightings, intention, prompt, ends
zusammenhang	coherence, context, contiguity, connection
zustimmung	accordance, agreement, approbation, consent, affirmation, allowance, approval, assent, compliance, compliancy, acclamation

Figure 5.22: 35 word dictionary extracted from online-resources

The gold standard was prepared in the following way. We gathered word-to-word translations by online-available dictionaries and eliminated German nouns for which we could not find at least two English translations showing a genuine “semantic” ambiguity. The resulting dictionary is shown in Fig. 5.22. It includes 35 German nouns with an average of 9-10 translations, i.e., in sum 333 English nouns. Based on this dictionary, we extracted all bilingual sentence pairs from the corpus which included a German noun from this dictionary in

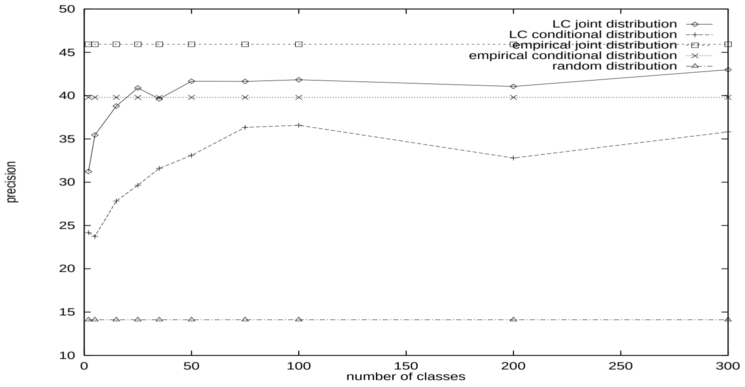

Figure 5.23: Precision of statistical disambiguation versus empirical and random choice

a verb-object position (".aso:o"). Furthermore, the English translation of the object was required to be included in our dictionary and had to appear in a similar verb-object position as the source-object for an acceptable English translation of the German verb. We marked the German noun n_g in the source-sentence, its English translation n_e as appearing in the corpus, and the English lexical verb v_e. This semi-automatic procedure resulted in a test corpus of 1341 sentences.

The goal of the statistical disambiguation was to determine for each marked German noun n_g the dictionary-specified translation $n \in Trans(n_g)$ which resulted in the most probable combination (v_e, n). That is, as a translation of n_g in the context of v_e,

$$n_e = \underset{n \in Trans(n_g)}{\arg\max} \; p_{LC}(v_e, n)$$

is selected if the arg max is defined. Otherwise, the output of the disambiguator is "don't know". The results of our LC-disambiguation tested against the gold standard are shown in Fig. 5.23 (precision) and Fig. 5.24 (recall). The clustering-based statistical disambiguator is compared with the empirical distribution of (v, n)-pairs in the training corpus and a random distribution on (v, n)-pairs. Precision measures the number of times the disambiguator under consideration chooses the same English translation as the human translator defining the gold standard. That is, precision is the number of "correct" translations according to the gold standard divided by the number of "correct" + "incorrect" translation. Recall specifies the number of times the disambiguation component chooses the "correct" translation out of the "correct" + "incorrect" + "don't know" decisions.

Fig. 5.23 shows the precision results for LC-disambiguators using a joint or a conditional probability measure, compared to the joint and conditional empirical distribution and a random distribution. The best result is obtained for the joint empirical distribution of (v, n)-pairs (45.928 %). According to the maximum likelihood paradigm, the LC-disambiguator approaches the empirical distribution in the limit of the number of classes, i.e., we see an im-

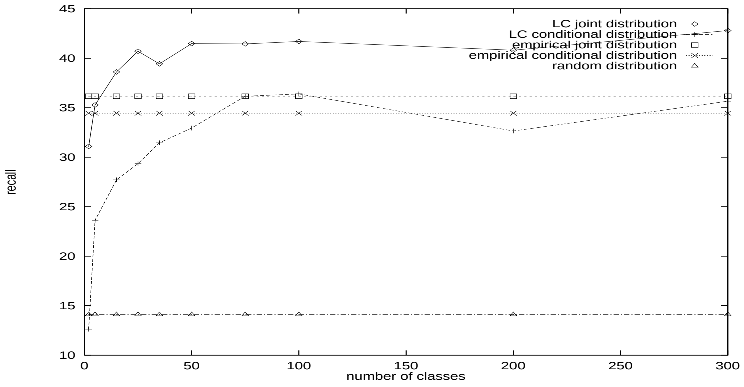

Figure 5.24: Recall of statistical disambiguation versus empirical and random choice

provement from ca. 31 % for LC-models with joint probability measure and 2 classes up to ca. 43 % for LC-models with 300 classes measured with a joint probability measure. The results are worse for both the conditional empirical distribution and the conditional LC-model, but worst for the random distribution (ca. 14 %). Note that these numbers have to be considered in the context of an average of ca. 10 translation choices for each noun.

However, as can be seen in Fig. 5.24, showing the recall results for the disambiguators, the main advantage of our LC-disambiguator is the gain in recall it has over the empirical distribution. That is, the high smoothing power of the LC-model enables a disambiguation decision for nearly every test item, i.e., $\arg\max_{n \in Trans(n_g)} \; p_{LC}(v_e, n)$ is never undefined since there is no case where $p_{LC}(v_e, n) = 0$ for all $n \in Trans(n_g)$. Thus we get a recall percentage of ca. 43 % for the best LC-disambiguators compared to 36.167 % for the joint empirical distribution. Random choice gives a result of ca. 14 % recall.

Note that the curves for the LC-disambiguator have the same shape as the curves resulting from the pseudo-disambiguation task reported above, thus making it a good guess to choose a proper LC-model in terms of class-cardinality from the cheaper pseudo-disambiguation task rather than choosing it from the labor-intensive evaluation on a gold standard.

In a subset of 100 test items, a human judge having access only to v_e and the set of candidates for n_e, i.e. the information used by the model, selected among translations. On this set, human performance was 39 % precision and recall. Performance for class models of size 100, 200, and 300 was 35 %, 39 %, and 45 % respectively (precision = recall). Joint empirical performance was 43 % precision and 34 % recall.

5.6.4 Discussion

We have proposed a statistical disambiguation component based on latent class models on pairs of grammatically related lexical items.

We believe the method meets the main conditions on a powerful statistical disambiguation component because:

1. The LC disambiguator has a high smoothing power, thus it is robust.
2. It resolves translation ambiguities in just one run, so it is quick.
3. It uses only minimal informations from the symbolic transfer component, so it can be called economic.
4. It relies only on monolingual information, so it is portable.
5. Evaluation on a pseudo-disambiguation task (80 % accuracy) and on a golden standard method (43 % accuracy on ca. 10 alternatives in average) shows that the method is accurate.

Bibliography

Abney, S. (1991). Parsing by chunks. In D. Bouchard and K. Leffel (Eds.), *Views on Phrase Structure*. Dortrecht: Kluwer Academic Publishers.

Abney, S. (1995). Chunks and dependencies: Bringing processing evidence to bear on syntax. In *Computational Linguistics and the Foundations of Linguistic Theory*. Stanford, CA: CSLI.

Abney, S. (1996). Chunk stylebook. Technical report, SfS, Universität Tübingen.

Baker, J. (1979). Trainable grammars for speech recognition. In D. Klatt and J. Wolf (Eds.), *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, pp. 547–550.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics 41*(1), 164–171.

Baum, L. E. and G. A. Sell (1968). Growth transformations for functions on manifolds. *Pacific Journal of Mathematics 27*(2), 211–227.

Beckwith, R., C. Fellbaum, D. Gross, and G. A. Miller (1991). Wordnet: A lexical database organized on psycholinguistic principles. In *Lexical Acquisition – Exploiting On-Line Resources to Build a Lexicon*, Chapter 9, pp. 211–232.

Beil, F., G. Carroll, D. Prescher, S. Riezler, and M. Rooth (1998). Inside-outside estimation of a lexicalized PCFG for German. –Gold–. In *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3). IMS, Universität Stuttgart.

Booth, T. L. and R. A. Thompson (1973). Applying probability measures to abstract languages. *IEEE Transactions on Computers C-22*(5), 442–450.

Brown, P., P. deSouza, R. Mercer, V. D. Pietra, and J. Lai (1992). Class-based n-gram models of natural language. *Computational Linguistics 18*(4), 467–479.

Buschbeck-Wolf, B. (1997). Resolution on demand. Technical Report 196, Verbmobil.

Carroll, G. (1997a). *Manual pages for* `charge`*,* `hyparCharge`*, and* `tau`. IMS, Universität Stuttgart.

Carroll, G. (1997b). *Manual pages for* `supar`*,* `ultra`*,* `hypar`*, and* `genDists`. IMS, Universität Stuttgart.

Carroll, G. and M. Rooth (1998). Valence induction with a head-lexicalized PCFG. In *Proceedings of EMNLP-3*, Granada.

Charniak, E. (1995). Parsing with context-free grammars and word statistics. Technical Report CS-95-28, Department of Computer Science, Brown University.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: M.I.T. Press.

Church, K. W., W. A. Gale, P. Hanks, and D. Hindle (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: exploiting on-line resources to build a lexicon.*

Dagan, I. and A. Itai (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics 20*, 563–596.

Dagan, I., L. Lee, and F. Pereira (1998). Similarity-based models of word cooccurrence probabilities. To appear in *Machine Learning.*

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society 39*(B), 1–38.

Duda, R. O. and P. E. Hart (1973). *Pattern classification and scene analysis.* New York: Wiley.

Ersan, M. and E. Charniak (1995). A statistical syntactic disambiguation program and what it learns. Technical Report CS-95-29, Department of Computer Science, Brown University.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong.* Oxford: Oxford Cognitive Science Series.

Gazdar, G., E. Klein, G. K. Pullum, and I. A. Sag (1985). *Generalized Phrase Structure Grammar.* Cambridge, MA: Harvard University Press.

Grimshaw, J. (1990). *Argument Structure.* Cambridge, MA: MIT Press.

Hale, K. and S. Keyser (1993). Argument structure and the lexical expression of syntactic relations. In K. Hale and S. Keyser (Eds.), *The View from Building 20.* Cambridge, MA: MIT Press.

Hindle, D. (1983). User manual for fidditch, a deterministic parser. Technical Memorandum 7590-142, Naval Research Laboratory.

Hindle, D. (1994). A parser for text corpora. In B. Atkins and A. Zampolli (Eds.), *Computational Approaches to the Lexicon.* Oxford: Oxford University Press.

Hindle, D. and M. Rooth (1993). Structural ambiguity and lexical relations. *Computational Linguistics 19*, 103–120.

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

Hughes, J. (1994). *Automatically Acquiring Classification of Words.* Ph. D. thesis, University of Leeds, School of Computer Studies.

Jackendoff, R. (1977). *$\bar{X}$ Syntax: A Study in Phrase Structure.* Cambridge, MA: MIT Press.

Karttunen, L., T. Yampol, and G. Grefenstette (1994). *INFL Morphological Analyzer/Generator 3.2.9 (3.6.4).* Xerox Corporation.

Katz, S. M. (1980). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing 35.*

Klavans, J. L. and M.-Y. Kan (1998, August). The role of verbs in document analysis. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, Montreal, Canada.

Kullback, S. and R. A. Leibler (1951). On Information and Sufficiency. *Annals of Mathematical Statistics 22*, 79–86.

Kural, M. (1996). *Verb Incorporation and Elementary Predicates.* Ph. D. thesis, University of California, Los Angeles.

Lang, B. (1989). A uniform formal framework for parsing. In *Proceedings of the 1st International Workshop on Parsing Technologies (IWPT 1)*, Pittsburgh, PA, pp. 28–42.

Lang, S. (1993). *Algebra* (Third edition ed.). Reading, MA: Addison-Wesley.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation.* Chicago/London: The University of Chicago Press.

Liberman, M. (1992). *ACL-DCI CD ROM 1.* Association for Computational Linguistics.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions.* New York: Wiley.

Melamed, I. D. (1997). Word-to-word models of translational equivalence. Technical Report IRCS, Nr. 98-08, University of Pennsylvania.

Ney, H., U. Essen, and R. Kneser (1994). On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language 8*, 1–38.

Pereira, F., N. Tishby, and L. Lee (1993). Distributional clustering of English words. In *Proceedings of the 31th Annual Meeting of the ACL*, Columbus, Ohio.

Pollard, C. and I. A. Sag (1987). *Information-Based Syntax and Semantics, Vol. 1: Fundamentals.* Stanford, CA: CSLI.

Resnik, P. (1993). *Selection and information: A class-based approach to lexical relationships.* Ph. D. thesis, University of Pennsylvania, CIS Department.

Resnik, P. (1997). Selectional preference and sense disambiguation. Paper presented at: Tagging Text with Lexical Semantics: Why, What, and How?, Workshop in connection with ANLP 97, Washington, D.C.

Ribas, F. (1994). An experiment on learning appropriate selectional restrictions from a parsed corpus. In *Proceedings of COLING-94*, Kyoto, Japan.

Ribas, F. (1995). On learning more appropriate selectional restrictions. In *Proceedings of the 7th Conference of the EACL*, Dublin, Ireland.

Rooth, M. (1998). Two-dimensional clusters in grammatical relations. In *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3). Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Rooth, M. (Ms.). Two-dimensional clusters in grammatical relations. In *Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, AAAI 1995 Spring Symposium Series. Stanford University.

Rooth, M., S. Riezler, D. Prescher, G. Carroll, and F. Beil (1998). EM-based clustering for NLP applications. In *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3). Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Saul, L. K. and F. Pereira (1997). Aggregate and mixed-order Markov models for statistical language processing. In *Proceedings of EMNLP-2.*

Schiller, A. and C. Stöckert (1995). *DMOR*. IMS, Universität Stuttgart.

Schmid, H. (1999). *Manual page for* `lopar`. IMS, Universität Stuttgart.

Schuhmacher, H. (1986). *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben.* Berlin: de Gruyter.

Schulte im Walde, S. (1998). Automatic semantic classification of verbs according to their alternation behaviour. Master's thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Skut, W. and T. Brants (1998). A maximum-entropy partial parser for unrestricted text. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montréal, Québec.

Stowell, T. (1981). *Origins of Phrase Structure.* Ph. D. thesis, MIT.

Wahba, G. and S. Wold (1975). A completely automatic french curve: Fitting spline functions by cross validation. *Commun. Statist. 4*, 1–17.