PROBABILISTIC PARSE SELECTION BASED ON SEMANTIC COOCCURRENCES*

Eirik Hektoen
Computer Laboratory, University of Cambridge
Pembroke Street, Cambridge CB2 3QG, UK
Eirik.Hektoen@cl.cam.ac.uk
http://www.cl.cam.ac.uk/users/eh101

Abstract

This paper presents a new technique for selecting the correct parse of ambiguous sentences based on a probabilistic analysis of lexical cooccurrences in semantic forms. The method is called "Semco" (for semantic cooccurrence analysis) and is specifically targeted at the differential distribution of such cooccurrences in correct and incorrect parses. It uses Bayesian Estimation for the cooccurrence probabilities to achieve higher accuracy for sparse data than the more common Maximum Likelihood Estimation would. It has been tested on the *Wall Street Journal* corpus (in the PENN Treebank) and shown to find the correct parse of 60.9% of parseable sentences of 6–20 words.

1 Introduction

In recent years there have been many proposals for probabilistic natural language parsing techniques, that is, techniques which not only find the possible syntactic derivations for a sentence, but also attempt to determine the most likely parse according to some probabilistic model. A basic example is a probabilistic context free grammar (PCFG), in which each production rule is associated with the conditional probability of it being applied when the left-hand non-terminal occurs in the generation of a sentence (e.g., Baker, 1982; Kupiec, 1991; Pereira and Schabes, 1992). This effectively regards the derivation of a sentence as a top–down recursive stochastic process, starting with the sentence non-terminal and ending with a random sentence in the language being modelled. While a PCFG is a pleasingly simple model, the fact that it assumes the choice of production at each step is only conditional on the left-hand non-terminal is a limitation to its accuracy. Some variations have therefore been proposed, by which the probability of a rule (or more generally, a parse derivation step) is made conditional on an extended view of the preceding derivation. For example, Briscoe and Carroll (1993) associate probabilities with transitions in an LR(1) parse table (partly reflecting the left context and a one-word lookahead), while Black *et al.* (1993) present a model in which virtually any aspect of the partial parse at any point in a derivation may be taken into account.

Both Briscoe and Carroll and Black *et al.* tie the probabilistic analysis to the precise parsing algorithm being used, effectively shifting the emphasis from modelling sentence generation to parse selection as a goal in its own right. Others take this further by separating the parsing and the parse selection completely. Hindle and Rooth (1993), for example, propose a system for resolving PP attachment ambiguities by comparing the degree of statistical association between the possible verb/preposition/noun triples after an arbitrary parser has produced a set of syntactically possible parses. This approach is attractive because it is based on lexical cooccurrences, which may reflect the acceptability of the meaning of a parse, but is less comprehensive than the other methods mentioned, since only a particular kind of ambiguity is covered.

*I am very grateful to Ted Briscoe, John Carroll, Miles Osborne, John Daugman, Gerald Gazdar, Steve Young and the anonymous IWPT-97 reviewers for their valuable advice. The work was made possible by a generous grant from the Norwegian Research Council (*Norsk Forskningsråd*, ST.30.33.221752) and an Overseas Research Students Award from the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom (ORS/9109219).

Yet others have made the opposite move, and presented parsing methods where the probabilistic analysis is used as the main driving force for the parser independently of any linguistically motivated grammar. Both Magerman (1995) and Collins (1996) propose such systems, where complex statistical patterns extracted from a Treebank constitute both the (shallow) syntactic grammar and selectional criteria. The results are systems which are comprehensive in the types of ambiguities they can handle and designed to extract a highly detailed and wide ranging statistical data from the training corpus, but which do not take any advantage of the analytical syntactic rules encoded in a formal grammar, and which do not support the derivation of semantic forms.

The system presented in this paper is essentially a specialised parse selector based on semantic forms derived from the parses found by a separate parser, and can therefore be used with any grammar and parser supporting formal semantics. It is based on lexical cooccurrences in terms of the predicates in the semantic forms, but handles all predicates uniformly and is therefore generally comprehensive in the types of ambiguities covered. It uses a complex statistical analysis to extract a large set of probabilistic parameters from the training corpus, but does not abandon the use of a formal grammar. Significantly, the system uses *Bayesian Estimation*[1] rather than simple Maximum Likelihood Estimation (MLE) for determining cooccurrence probabilities. This appears to be a sufficient response to the high degree of sparseness in the lexical cooccurrence data without the blurring associated with smoothing and clustering techniques (generally required for MLE). It seems reasonable to expect that a parse selection system should benefit from being trained on the same form of data that it is to be applied to—that is, specifically on the selection of the correct parse in sets of possible parses for different sentences rather than the unconditional probability of correct parses in isolation. The focus of the training in this system is therefore the differential distribution of cooccurrences in correct versus incorrect parses. The system is called "Semco" (for semantic cooccurrence analysis), and has been trained and tested on (separate parts of) the *Wall Street Journal* corpus in the PENN Treebank.

2 Definitions

The aim of the Semco analysis is to model cooccurrences of lexical predicates in semantic forms for the purpose of parse selection. A semantic form is here assumed to be a logical expression or description (including unscoped or quasi logical forms) derived in a compositional manner from a syntactic parse tree, and thus in general representing one of several possible interpretations of a sentence. A *cooccurrence* should represent the variable semantic linking of the predicates (generally representing lexical items from the sentence) in such expressions, and is therefore defined as the coincidence of two predicates being applied to the same element (a quantified variable or a constant). More precisely, if the ith argument of the predicate Q and the jth argument of the predicate R are the same, the semantic form is said to include the cooccurrence $(Q.i, R.j)$ understood as an unordered pair.

To see how such cooccurrences can be used for sentence disambiguation, consider the following exchange (from Andy Warhol, 1975):

(1) B: *Is that a female impersonator?*
A: *Of what?*

The expression *female impersonator* is so commonly used as a noun–noun compound ("an impersonator of a female"), that it comes as a surprise here when A's reply requires an alternative reading in which *female* is used as an adjective (giving "an impersonator who is female"). The difference between the two readings is represented by the semantic predicates and cooccurrences shown in Table 1. Note that the cooccurrences are concise representations of the facts that *female* and *impersonator* form a noun compound in Reading 1, while *female* is used as an adjectival modifier of *impersonator* in Reading 2.

The probabilistic analysis treats cooccurrences as elementary, atomic units, so that for a given grammar and lexicon there is a finite set $\mathcal{C}$ of possible cooccurrences. As far as this analysis is concerned, a parse, or

[1] Note that *Bayesian Estimation* (following e.g. Freund, 1992) refers not merely to the use of Bayes's law, but to the particular technique of estimating unknown probabilities by integration over a continuous probability distribution applied in Section 4.

	Reading 1	Reading 2
Categories	$female_{\mathsf{N}}$ $impersonator_{\mathsf{N}}$	$female_{\mathsf{Adj}}$ $impersonator_{\mathsf{N}}$
Predicates	(female x) (impersonator y) (NCOMP y x)	(female x) (impersonator x)
Cooccurrences	(female.0 NCOMP.1) (impersonator.0 NCOMP.0)	(female.0 impersonator.0)

Table 1: Example predicates and cooccurrences

derivation, is regarded as a set of cooccurrences, with the set of all parses being $\mathcal{D} = \{d \mid d \subseteq \mathcal{C}\}$. Similarly, a sentence is regarded as a set of possible parses, such that the set of all possible sentences is $\mathcal{S} = \{s \mid s \subseteq \mathcal{D}\}$.

For each cooccurrence, parse and sentence there is a corresponding *event*—conventionally represented by the corresponding capital letter—referring to the status of a random sentence. More precisely:

- C is the event that the correct parse of the sentence includes the cooccurrence c.
- D is the event that the correct parse is d. It can be expressed as the conjunction of the cooccurrence events for all the cooccurrences in d and the negated cooccurrence events for all other cooccurrences:

$$D = \bigwedge_{c \in d} C \wedge \bigwedge_{c \notin d} \neg C. \tag{2}$$

- S is the event that the correct parse is an element of s, and is simply the disjunction of the corresponding parse events:

$$S = \bigvee_{d \in s} D. \tag{3}$$

3 The Event Space

Given the above definitions of cooccurrence, parse and sentence events, one's first reaction may be to regard the associated probabilities as the relative frequencies of the respective entities in the language—that is, in the correct analyses of the sentences in the training corpus. This is not the only possible definition, however, and would have serious disadvantages for the following analysis. For example, it would mean that the model should reflect the typical number of cooccurrences in an average sentence, making any assumption of independence between cooccurrences inappropriate (since the probabilities of any additional cooccurrence would diminish once the number has passed this average). It would also make it impossible for the training to be based on the differential distribution of cooccurrences in correct and other syntactically possible, but incorrect, parses of the training sentences, since such alternative parses would have no particular status in the model.

Alternative definitions of the basic probabilities are possible because all actual references to probabilities will be conditional on some given sentence. The only requirement of the probabilistic model is that it predicts the relative, conditional probabilities for different parses for any given sentence, while the unconditional probabilities of different sentences in the corpus are never directly relevant.

To derive a suitable definition of the basic event space here, we will take the presumed independence of all cooccurrence events as the starting point, based on the view that the model ought to satisfy the following basic assumption:

(4) **Basic Assumption:** The relative probabilities of any two parses depends only on the cooccurrences that distinguish between them, and not on any cooccurrence present in both of them, any cooccurrence absent in both of them, or any further possible parses of the same sentence.

The independence of the cooccurrence events follows by considering two parses, d_1 and d_2, that are only distinguished by a single cooccurrence c, say $c \in d_1$ and $d_2 = d_1 \setminus c$. By the basic assumption, the ratio $P(D_1|S) : P(D_2|S)$ is invariant for any such d_1 and d_2 and $s \supseteq \{d_1, d_2\}$, and this ratio must be $P(C) : P(\neg C)$. From the independence of the cooccurrence events and (2) it follows that the unconditional probability of a parse event D is given by

$$P(D) = \prod_{c \in d} P(C) \prod_{c \notin d} P(\neg C). \tag{5}$$

4 Cooccurrence Probability Estimation

The cooccurrence probabilities $P(C)$ represent the basic parameters in the analysis and need to be estimated from the training corpus. In many other probabilistic analyses of corpus data the basic parameters are estimated as the observed relative frequencies of sentences displaying the relevant characteristics, but such a simple approach is not possible here. Instead, Bayesian estimation (see e.g. Freund, 1992) is used, by which the estimate is defined in terms of the distribution of the probability to be estimated regarded as a continuous probabilistic variable.

More precisely, the unknown probability $P(C)$ for a given c is regarded as the continuous probabilistic variable Θ with a value θ in the interval $(0, 1)$. Let s_i for i in $0, \ldots, t$ be all the sentences in the training corpus, and let d_i be the correct parse of each s_i. Let also S_i be the event corresponding to s_i, and let C_i be the event associated with the cooccurrence c with respect to the sentence s_i. The overall status of the corpus with respect to c can then be expressed as the conjunction of the events $\hat{S} = \bigwedge S_i$ and

$$\hat{C} = \bigwedge_{c \in d_i} C_i \wedge \bigwedge_{c \notin d_i} \neg C_i. \tag{6}$$

The distribution of Θ given the observed status of the corpus can then be expressed as the probability density function $\phi(\theta|\hat{C}, \hat{S})$, which according to Bayes's law is given by

$$\phi(\theta|\hat{C}, \hat{S}) = P(\hat{C}|\theta, \hat{S}) \frac{h(\theta|\hat{S})}{P(\hat{C}|\hat{S})}. \tag{7}$$

Here $h(\theta|\hat{S})$ is the probability density function for the distribution of Θ given only the sentences in the corpus (i.e., not knowing the correct parses), and $P(\hat{C}|\hat{S})$ is the (discrete) probability of $\hat{C}$ not knowing θ. In other words, $h(\theta|\hat{S})$ is effectively (for our purposes) the *prior* distribution of Θ (regarding $\hat{S}$ as fixed), while $\phi(\theta|\hat{C}, \hat{S})$ is the *posterior* distribution of the same given the correct disambiguation of the sentences implied by $\hat{C}$.

The probability $P(\hat{C}|\theta, \hat{S})$, that is, the probability of the correct (observed) parse selections in the corpus in terms of the cooccurrence c and given $\theta = P(C)$, is given according to (6) and the fact that each sentence represents an independent draw in the event space by the product

$$P(\hat{C}|\theta, \hat{S}) = \prod_{c \in d_i} P(C_i|\theta, \hat{S}) \prod_{c \notin d_i} \left(1 - P(C_i|\theta, \hat{S})\right). \tag{8}$$

To determine $P(C_i|\theta, \hat{S})$, let n_i and m_i be the number of parses of s_i which do and do not, respectively, include c. Since the ratio of the probabilities of each of the former to each of the latter is $\theta : (1 - \theta)$, we get

$$P(C_i|\theta, \hat{S}) = \frac{n_i \theta}{n_i \theta + m_i(1 - \theta)}. \tag{9}$$

Returning to equation (7), the probability $P(\hat{C}|\hat{S})$ can now be determined from $P(\hat{C}|\theta, \hat{S})$ by the integral

$$P(\hat{C}|\hat{S}) = \int_0^1 P(\hat{C}|\theta, \hat{S}) h(\theta|\hat{S})\, d\theta. \tag{10}$$

This leaves only the prior distribution, $h(\theta|\hat{S})$, which cannot be determined analytically, but which may be estimated empirically from a preliminary estimation of all the cooccurrences in the corpus (e.g., using $h(\theta|\hat{S}) = 1$ as a temporary simplification without seriously affecting the overall distribution).

Having found the posterior distribution $\phi(\theta|\hat{C},\hat{S})$, a straightforward application of the Bayesian estimation method would be to estimate the unknown probability $P(C)$ as the *expected* value of Θ, that is

$$E(\Theta|\hat{C},\hat{S}) = \int_0^1 \phi(\theta|\hat{C},\hat{S})\,\theta\,d\theta. \tag{11}$$

A slight variation of this will be used here, however, based on the observation that the expected value operator in (11) essentially represents a continuous, arithmetic mean of the cooccurrence probability θ weighted by the distribution function $\phi(\theta|\hat{C},\hat{S})$. As such, it would represent a reasonable estimation of $P(C)$ if the result were to be used as a term in a sum of such results, but according to (5) the main use of a cooccurrence probability will be represented by a factor of either $P(C)$ or $1 - P(C)$ in a parse probability. As the net effect of the presence or absence of the cooccurrence in a parse thus is to either multiply or divide the parse probability by $\frac{\theta}{1-\theta}$, a more suitable estimate of $P(C)$ is found by applying the expected value operator to $\ln\frac{\theta}{1-\theta}$, that is, defining the estimate as $\tilde{p}_c$ given by the equation[2]

$$\ln\frac{\tilde{p}_c}{1-\tilde{p}_c} = E\Big(\ln\frac{\Theta}{1-\Theta}|\hat{C},\hat{S}\Big) = \int_0^1 \phi(\theta|\hat{C},\hat{S})\,\ln\frac{\theta}{1-\theta}\,d\theta. \tag{12}$$

From (7), (8), (10) and (12), noting that the integral in (10) does not depend on θ and can therefore be moved outside that in (12), we then get the following overall expression for the cooccurrence probability estimate:

$$\ln\frac{\tilde{p}_c}{1-\tilde{p}_c} = \frac{\int_0^1 P(\hat{C}|\theta,\hat{S})h(\theta|\hat{S})\ln\frac{\theta}{1-\theta}\,d\theta}{\int_0^1 P(\hat{C}|\theta,\hat{S})h(\theta|\hat{S})\,d\theta}. \tag{13}$$

To summarise, the effect of all this is that the posterior probabilistic distribution of the cooccurrence probability is determined from the prior distribution of such probabilities and the observations relating to this cooccurrence in the corpus. The posterior distribution represents the full knowledge we have of the likelihood of different possible values of the cooccurrence probability, and the final estimate is defined as that which represents the best approximation (for our purposes) of this as a fixed value.

5 Implementation Notes

The previous section derived (13) in conjunction with (8) and an empirical estimation of $h(\theta|\hat{S})$ as the main expression for the estimate $\tilde{p}_c$ of a cooccurrence probability $P(C)$. In practice, only sentences with at least one parse that includes c and one that doesn't, that is, for which $n_i, m_i > 0$, have any effect on the result. For any such sentence, moreover, it is enough to record the ratio

$$r_i = \frac{n_i}{m_i} \tag{14}$$

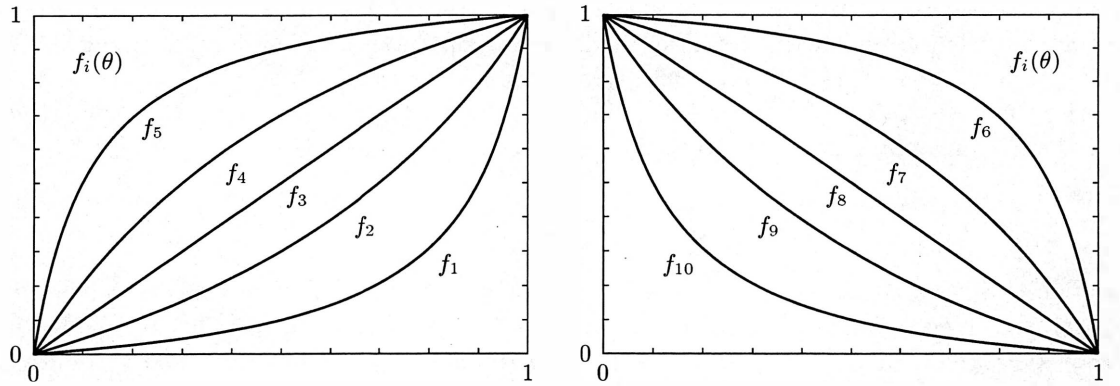
in the training data. Then, by defining

$$f(\theta) = \left[\prod_{n_i,m_i>0} f_i(\theta)\right] h(\theta|\hat{S}) \tag{15}$$

and

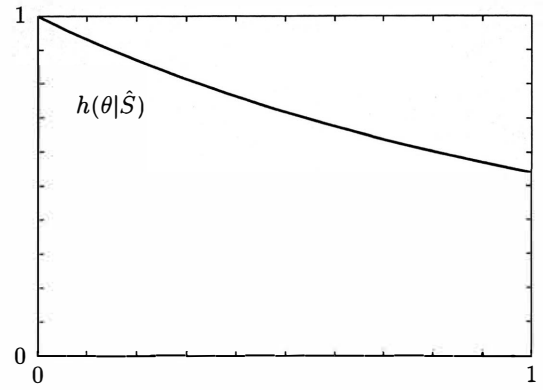
$$f_i(\theta) = \begin{cases} P(C_i|\theta,\hat{S}) &= r_i\theta/(r_i\theta+1-\theta) & \text{if } c \in d_i \\ P(\neg C_i|\theta,\hat{S}) &= (1-\theta)/(r_i\theta+1-\theta) & \text{otherwise,} \end{cases} \tag{16}$$

[2]The convergence of the integral in (12), although $\ln\frac{\theta}{1-\theta} \to \pm\infty$ for $\theta \to 0$ and $\theta \to 1$, follows from the well-known convergence of $\int_0^1 \ln\theta\,d\theta = -1$ and the fact that the relevant probability distribution functions are finite.

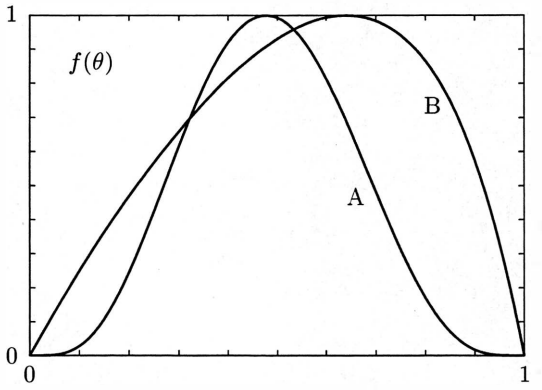


(a) examples of $f_i(\theta)$ for $r_i = \frac{1}{9}, \frac{3}{7}, \frac{5}{5}, \frac{7}{3}, \frac{9}{1}$ (for $i = 1, \ldots, 5$, resp.) and $c \in d_i$

(b) more examples of $f_i(\theta)$ for the same r_i (for $i = 6, \ldots, 10$, resp.) but with $c \notin d_i$



(c) the prior distribution $h(\theta|\hat{S}) = \frac{1.36}{(1+0.36\theta)^2}$ in the *Wall Street Journal* (found by manual curve fitting to the results of a preliminary analysis)



(d) possible posterior distributions based on the preceding functions:
A: $f(\theta) = \left[\prod_1^{10} f_i(\theta)\right] h(\theta|\hat{S})$,
B: $f(\theta) = \left[\prod_{2,7} f_i(\theta)\right] h(\theta|\hat{S})$

Figure 1: Cooccurrence probability density functions. All functions are scaled to give a maximum value of 1 with no significance for the cooccurrence probability estimation.

equation (13) is reduced to

$$\ln \frac{\tilde{p}_c}{1 - \tilde{p}_c} = \frac{\int_0^1 f(\theta) \ln \frac{\theta}{1-\theta} \, d\theta}{\int_0^1 f(\theta) \, d\theta}. \tag{17}$$

The functions $f_i(\theta)$ and $f(\theta)$ represent scaled probability density functions of Θ: $f_i(\theta)$ is (proportional to) the apparent distribution based only on sentence i; $f(\theta)$ is (proportional to) the posterior distribution based on all the relevant sentences as well as the prior distribution $h(\theta|\hat{S})$. Fig. 1 shows some typical examples of the $f_i(\theta)$ functions, the $h(\theta|\hat{S})$ found to represent the distribution of cooccurrence probabilities in the *Wall Street Journal* corpus, and two possible forms of $f(\theta)$ derived from them.

The actual computation of $\tilde{p}_c$ will have to be by numerical integration of (17) based on the r_i values extracted from the training corpus for each cooccurrence.[3] This may seem rather costly in computational terms, but is feasible in practice with a careful implementation. In particular, it is generally not necessary to compute $\tilde{p}_c$ for all the cooccurrences in the training corpus, since in most cases only a small fraction of them will ever be

[3]To avoid arithmetic overflow at or near $\theta = 0$ and $\theta = 1$, however, it is convenient to rewrite the nominator in (17) as $\int_0^1 \big([f(\theta) - f(0)] \ln \theta - [f(\theta) - f(1)] \ln(1 - \theta)\big) d\theta - f(0) + f(1)$, and omit the logarithms when the preceding factors are 0.

required. Instead, a practical implementation may compile indexed files with the r_i values extracted from the corpus for each cooccurrences and compute the small number of cooccurrence probabilities required for any sentence when the need arises.

6 Test Results

An implementation of the Semco analysis has been tested with a feature unification-based, medium-wide coverage, syntactic/semantic grammar and sentences from the bracketted form of the *Wall Street Journal* part of the PENN Treebank. Due to limitations in the grammar and parser, the corpus had to be restricted in different ways: It was first limited to sentences of 6–20 words (after the deletion of parenthetical material) which parsed successfully with at most 100 parses (as more ambiguous ones would add disproportionately to the overall cost of the computation). It was further reduced by rejecting sentences for which no parse achieved a minimum threshold similarity with the corpus bracketting, or for which more than 30% of the parses achieved the same, maximal such similarity.

The parse/bracketting similarity measure used was a weighted sum of different counts based on shared constituents (with or without matching labels), crossing brackets, and the overall length of the sentence (for normalising the result). The threshold was set to correspond, typically, to a parse including about three-quarters of the corpus constituents with no crossing brackets. Multiple parses achieving the same, maximal score above this threshold (within the limit mentioned above) were presumed "equally correct" for training and testing purposes. The number of sentences left in the corpus after each step in the selection process were:

Number of sentences of length 6–20 words	20155
... which parsed successfully	12162
... which had up to 100 parses	10852
... which met the similarity measure requirements	7887

The final set of sentences was shuffled and divided into five partitions, allowing a cycle of five training/testing iterations to be run with different splits of the corpus into $\frac{4}{5}$ for training and $\frac{1}{5}$ testing.

Table 2 shows a detailed overview of the main test statistics. The main part of the table refers to a particular division of the corpus, and shows a breakdown by sentence length and ambiguity. This is followed in the last row by the mean results of the five training/testing rounds based on different divisions of the corpus.

The Parseval measures in the table have become a frequently used, common standard for comparison of different systems, but are only partly relevant for a system such as this. The fact that they are taken at constituent level means that they primarily measure the ability of the system to reproduce the precise bracketting in the corpus rather than the correct selection of parses as such. As a result, fully correct parses receive reduced scores where the grammar used by the parser differs significantly from that assumed in the corpus, while incorrect parses are credited to the extent that some of their constituents coincide with those in the corpus. In this system the grammar generally produces much more detailed parses than the corpus (i.e., with many more constituents, e.g. at bar-1 level), meaning that the *precision* rate is severely reduced and effectively rendered meaningless, while the *recall* rate is not similarly badly affected. This systematic imbalance between the corpus bracketing and the parses also prompted the inclusion of two *crossing brackets* rates: where four words are bracketed as "(w_1 w_2) w_3 w_4" in the corpus but parsed as "w_1 ((w_2 w_3) w_4)", for example, it counts as one crossed bracket in the corpus ("xb/c") but two in the parse ("xb/p"). Again, the generally greater number of constituents in the parses than in the corpus means that the latter is artificially high and a poor basis for comparison with other systems.

The table includes three variant Parseval measures. The "non-crossing precision", which is like the standard precision rate except that any constituent in the parse that does not actually cross brackets with the corpus form is assumed to be correct, is intended as an illustration of a possible way to deal with the problem discussed above. The labelled precision and recall rates are the common variations of the Parseval measures where constituents are required to have matching syntactic labels.

	Parseval measures				Variants			Correct parse ranked in top n				
	prec	rec	xb/c	xb/p	nxpr	lpre	lrec	1	2	3	4	5
Results with the original partitioning of the corpus for training/testing:												
All sents	61.3	87.2	0.40	0.79	93.1	58.0	82.5	60.8	75.1	81.3	85.6	89.2
Length												
6–10	70.0	94.3	0.12	0.20	97.2	67.7	91.2	82.4	91.2	94.2	97.0	98.7
11–15	61.5	87.1	0.41	0.76	93.5	58.2	82.5	58.9	75.3	82.7	87.2	90.6
16–20	57.2	83.7	0.67	1.45	91.0	53.5	78.4	41.2	57.9	65.7	71.3	77.1
Ambiguity												
1–20	62.8	89.9	0.31	0.53	95.1	59.8	85.6	69.3	83.1	88.6	92.9	95.6
21–40	59.0	83.1	0.63	1.20	90.9	55.3	77.9	42.5	60.2	69.2	74.7	80.1
41–60	58.1	81.4	0.69	1.47	89.6	54.2	75.8	37.9	56.0	64.7	67.2	71.6
61–80	56.8	78.7	0.59	2.12	85.0	53.4	73.9	35.1	45.9	51.4	55.4	63.5
81–100	58.2	83.2	0.44	1.40	90.1	54.5	78.0	38.5	46.2	51.9	55.8	61.5
The mean results over a cycle of five different partitionings of the corpus:												
All sents	61.3	87.4	0.41	0.77	93.4	58.0	82.7	60.9	74.7	82.2	85.9	89.3

All entries are in %, except xb/c and xb/p which are per sentence. Definitions:

prec	*precision*, the proportion of the constituents in the parses also found in the corpus
rec	*recall*, the proportion of the constituents in the corpus also found in the parses
xb/c, xb/p	*crossing brackets rate*, counted in the corpus form or selected parse, resp.
nxpr	*non-crossing precision*, constituents in the parse that don't cross any brackets in the corpus
lpre, lrec	*labelled precision/recall*, where the syntactic labels in the parse and the corpus must match
1–5	*n best correct selection rate*, sentences for which the correct parse (i.e., the best match with the corpus bracketting) is ranked in the top n parses by the selection algorithm

Table 2: Parse selection accuracy test results

The measures that most directly reflect the practical accuracy of the Semco system in my opinion are those headed "Correct parse ranked in top n", as these show the relative frequencies of fully correct, sentence-level disambiguations ($n = 1$) and near-misses (n in 2–5). In the compilation of these figures, the "correct" parse of each sentence was identified by comparison with the corpus bracketting using the same similarity formula that was used for the training of the system.

The table shows, as one would expect, that the selection accuracy tends to diminish as sentences increase in length or ambiguity, but the differences in the three higher bands of ambiguities are relatively minor (indeed, for many measures the sentences with 81–100 parses do better than those of 41–60 parses). This indicates that the decision to omit sentences of more than 100 parses from the testing is not likely to have affected the overall performance of the system greatly.

7 Conclusions

The results in Table 2 show that the Semco technique achieves relatively high levels of parse selection accuracy, and that it therefore may represent a good practical method of sentence disambiguation. This is reflected in the Parseval recall rate of 87.4%, the average of only 0.41 crossing brackets per sentence (with respect to the corpus bracketting), and in the correct disambiguation of 60.9% of the sentences (with this figure increasing to 74.7% and 82.2% if near-misses ranked 2nd or 3rd are included).

Comparison with other published work is made difficult by the problems with the Parseval measures discussed

	Sentence lengths	Precision	Recall	Labelled precision	Labelled recall	Crossing brackets
Semco	6–20	61.3	87.4	58.0	82.7	0.41
	11–20	59.8	85.8	56.3	80.9	0.51
Magerman (1995)	10–20	90.8	90.3	89.0	88.5	0.49
(SPATTER)	4–40	86.3	85.8	84.5	84.0	1.33
Collins (1996)	1–40			86.3	85.8	1.14

Entries are in %, except the crossing brackets rate, which is per sentence. Note that the *precision* and *labelled precision* rates are poor measures of the Semco system's accuracy because of the more detailed parses produced by the grammar compared to the corpus annotations. The crossing brackets rate included for Semco is that counted against the corpus bracketting ("xb/c").

Table 3: Parseval measures of parse accuracy of different systems

above and by other incompatibilities between the different systems and the precise corpora used. Table 3 shows, however, the main figures from the Semco system and results published by Magerman (1995) (for his "SPATTER" system) and Collins (1996). Considering the strong relationship between sentence length and accuracy shown in Table 2, the most comparable figures in Table 3 are those for Semco with 11-20 words and Magerman's SPATTER system with 10–20 words. Disregarding the precision rates, which are severely biased against Semco, the table shows that while SPATTER performs somewhat better in terms of recall, the difference in the crossing brackets rates is insignificant. (Collins only includes results for sentence of 1–40 words, making any direct comparison with Semco dubious, but the results are broadly similar to Magerman's for 4–40 words.)

To be fair, the Semco results are based only on a subset of suitable, parseable sentences in the corpus (about 39% of the original sentences within the length range), but this is an inevitable consequence of the major differences between the systems. Magerman's and Collins's systems are both based on extracting local syntactic structures along with their statistical distribution directly from the corpus annotation, and are therefore independent of any linguistic analysis of the corresponding syntax. This robustness can be an important advantage—it is very hard to write a formal grammar with something approaching full coverage of naturally occurring languages—but has several disadvantages too. In particular, the lack of a linguistic analysis means that there is no guarantee that a parse generated by these systems represents a meaningful sentence-wide syntax, and the models cannot support interpretation through compositional semantics.

The Semco system is compatible with any parsing technique capable of supporting formal semantics, making it potentially much more useful in a wider, practical NLP system where any form of interpretation is required. The lack of robustness in the experimental set-up discussed in this paper is not a consequence of the Semco technique, but a reflection of the limited coverage of the grammar used. In future developments of this work, it is intended that the Semco system will be used with a fully wide-coverage natural language grammar and an n-best parser that includes mechanisms for handling undergeneration.

To conclude, the main novel aspect of the Semco analysis is the way the probabilistic model is based on the differential distribution of cooccurrences in correct and incorrect parses. This requires an analysis in which probabilities don't represent direct frequencies in the data, but rather correspond to a hypothetical event universe of which real sentences are only a small fragment. Simply put, the event universe includes sentences with any number of cooccurrences—whether they require a million words or ten—but this is no problem for the practical application of the system which is always conditional on a concrete, given sentence. There is no need for normalisation of the probabilities of parses with different numbers of cooccurrences in this analysis, sine the presence and absence of any cooccurrence are regarded as complementary events.

The use of Bayesian estimation for the basic cooccurrence probabilities is partly a requirement from the probabilistic model, but is also a means of dealing with the highly sparse data in a theoretically motivated manner. Where MLE tends to be highly inaccurate for very sparse data—requiring clustering or smoothing—

and directly inappropriate for unseen data, Bayesian estimation finds the result *theoretically expected* to represent the best approximation of the true probability based on a full analysis of the latter's continuous probability distribution. The result is unable to distinguish between different cooccurrences with the same observations (e.g., unseen), as smoothing or clustering might, but are statistically unbiased such that the random errors in the probability estimates for a set of cooccurrences will tend to cancel out and lead to improved accuracy in the parse probabilities. More detailed experiments by Hektoen (1997) show, moreover, that clustering the cooccurrence data in combination with Bayesian estimation reduces the accuracy of the parse selection.

References

Baker, J. 1982. Trainable Grammars for Speech Recognition. *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, pages 547–550.

Black, Ezra, Fred Jelinek, John Lafferty, David M. Magerman, Robert Mercer and Salim Roukos. 1993. Towards History-based Grammars: Using Richer Models for Probabilistic Parsing. In *Proceedings, 31st Annual Meeting of the Association for Computational Linguistics*, pages 31–37, Columbus, Ohio, USA.

Briscoe, Ted and John Carroll. 1993. Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars. *Computational Linguistics*, 19(1):25–59.

Collins, Michael John. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings, 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, Santa Cruz, California, USA.

Freund, John E. 1992. *Mathematical Statistics*, 5th edition. Prentice–Hall International, Inc., Englewood Cliffs, New Jersey, USA.

Hektoen, Eirik. 1997. *Statistical Parse Selection using Semantic Cooccurrences.* PhD thesis, Churchill College, Cambridge University, Cambridge, UK. Available from `http://www.cl.cam.ac.uk/users/eh101`.

Hindle, Donald and Mats Rooth. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1):103–120.

Kupiec, Julian. 1991. A Trellis-Based Algorithm for Estimating the Parameters of a Hidden Stochastic Context-Free Grammar. *DARPA Speech and Natural Language Workshop*, Asilomar, California, USA.

Magerman, David M. 1995. Statistical Decision-Tree Models for Parsing. In *Proceedings, 33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, Cambridge, Massachusetts, USA.

Pereira, Fernando and Yves Schabes. 1992. Inside–Outside Re-estimation for Partially Bracketed Corpora. In *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Newark, Delaware, USA.

Schabes, Yves. 1992. Stochastic Lexicalised Tree-Adjoining Grammars. In *Proceedings of the fifteenth International Conference on Computational Linguistics: COLING-92*, volume 2, pages 426–432, Nantes, France.

Warhol, Andy. 1975. *The Philosophy of Andy Warhol.* Harcourt Brace Jovanovich, San Diego, California, USA.