Word Sense Disambiguation Using a Second Language Monolingual Corpus

Ido Dagan*
AT&T Bell Laboratories

Alon Itai†
Technion - Israel Institute of Technology

This paper presents a new approach for resolving lexical ambiguities in one language using statistical data from a monolingual corpus of another language. This approach exploits the differences between mappings of words to senses in different languages. The paper concentrates on the problem of target word selection in machine translation, for which the approach is directly applicable. The presented algorithm identifies syntactic relationships between words, using a source language parser, and maps the alternative interpretations of these relationships to the target language, using a bilingual lexicon. The preferred senses are then selected according to statistics on lexical relations in the target language. The selection is based on a statistical model and on a constraint propagation algorithm, which handles simultaneously all ambiguities in the sentence. The method was evaluated using three sets of Hebrew and German examples and was found to be very useful for disambiguation. The paper includes a detailed comparative analysis of statistical sense disambiguation methods.

1. Introduction

The resolution of lexical ambiguities in non-restricted text is one of the most difficult tasks of natural language processing. A related task in machine translation, on which we focus in this paper, is target word selection. This is the task of deciding which target language word is the most appropriate equivalent of a source language word in context. In addition to the alternatives introduced by the different word senses of the source language word, the target language may specify additional alternatives that differ mainly in their usage.

Traditionally several linguistic levels were used to deal with this problem: syntactic, semantic and pragmatic. Computationally the syntactic methods are the most affordable, but are of no avail in the frequent situation when the different senses of the word show the same syntactic behavior, having the same part of speech and even the same subcategorization frame. Substantial application of semantic or pragmatic knowledge about the

* AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974 (dagan@research.att.com). The work reported here was done while the author was at the Technion - Israel Institute of Technology.
† Department of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel (itai@cs.technion.ac.il).

word and its context requires compiling huge amounts of knowledge, the usefulness of which for practical applications in broad domains has not yet been proven (e.g. (Lenat et al., 1990; Nirenburg et al., 1988; Chodorow, Byrd, and Heidron, 1985)). Moreover, such methods usually do not reflect word usages.

Statistical approaches, which were popular several decades ago, have recently reawakened and were found useful for computational linguistics. Within this framework, a possible (though partial) alternative to using manually constructed knowledge can be found in the use of statistical data on the occurrence of lexical relations in large corpora (e.g. (Grishman, Hirschman, and Nhan, 1986)). The use of such relations (mainly relations between verbs or nouns and their arguments and modifiers) for various purposes has received growing attention in recent research (Church and Hanks, 1990; Zernik and Jacobs, 1990; Hindle, 1990; Smadja, 1993). More specifically, two recent works have suggested using statistical data on lexical relations for resolving ambiguity of PP-attachment (Hindle and Rooth, 1991) and pronoun references (Dagan and Itai, 1990; Dagan and Itai, 1991).

Clearly, statistics on lexical relations can be useful also for target word selection. Consider, for example, the Hebrew sentence extracted from the foreign news section of the daily “Ha-Aretz”, September 1990 (transcripted to Latin letters):

(1)	*Nose*	*ze*	*mana'*	*mi-shtei*	*ha-mdinot*	*mi-lah̲tom*	*'al*	*h̲oze*	*shalom.*
	issue	this	prevented	from-two	the-countries	from-signing	on	treaty	peace

This sentence would translate into English as:

(2) This issue prevented the two countries from signing a peace treaty.

The verb *lah̲tom* has four senses: ‘sign’, ‘seal’, ‘finish’ and ‘close’. The noun *h̲oze* means both ‘contract’ and ‘treaty’, where the difference is mainly in usage rather than in the meaning (in Hebrew the word *h̲oze* is used for both sub-senses).

One possible solution is to consult a Hebrew corpus tagged with word senses, from which we would probably learn that the sense ‘sign’ of *lah̲tom* appears more frequently with *h̲oze* as its object than all the other senses. Thus we should prefer that sense. However, the size of corpora required to identify lexical relations in a broad domain is very large and therefore it is usually not feasible to have such corpora manually tagged with word senses.[1] The problem of choosing between ‘treaty’ and ‘contract’ cannot be solved using only information on Hebrew, because Hebrew does not distinguish between them.

The solution suggested in this paper is to identify the lexical relationships in corpora of the *target* language, instead of the source language. We consider word combinations and count how often they appear in the same syntactic relation as in the ambiguous sentence. For the above example, the noun compound ‘peace treaty’ appeared 49 times

1 Hearst (1991) suggests a sense disambiguation scheme along this line. See Section 7 for a comparison of several sense disambiguation methods.

in our corpus (see Section 4.3 for details on our corpus), whereas the compound 'peace contract' did not appear at all; the verb-object combination 'to sign a treaty' appeared 79 times while none of the other three alternatives appeared more than twice. Thus, we first prefer 'treaty' to 'contract' because of the noun compound 'peace treaty' and then proceed to prefer 'sign' since it appears most frequently having the object 'treaty'. The order of selection is determined by a constraint propagation algorithm. In both cases, the correctly selected word is not the most frequent one: 'close' is more frequent in our corpus than 'sign' and 'contract' is more frequent than 'treaty'. Also, using a model of statistical confidence the algorithm avoids a decision in cases where no alternative is significantly better than the others.

Our approach can be analyzed from two different points of view. From that of monolingual sense disambiguation, we exploit the fact that the mapping between words and word senses varies significantly among different languages. This enables us to map an ambiguous construct from one language to another, obtaining representations in which each sense corresponds to a distinct word. Now it is possible to collect cooccurrence statistics automatically from a corpus of the other language, without requiring manual tagging of senses.[2]

From the point of view of machine translation, we suggest that some ambiguity problems are easier to solve at the level of the target language instead of the source language. The source language sentences are considered as a noisy source for target language sentences, and our task is to devise a target language model that prefers the most reasonable translation. Machine translation (MT) is thus viewed in part as a recognition problem, and the statistical model we use specifically for target word selection may be compared with other language models in recognition tasks (e.g. (Katz, 1987; Jelinek, 1990) for speech recognition). To a limited extent, this view is shared with the statistical machine translation system of Brown et al. (1990), which employs a target language n-gram model (see Section 8 for a comparison with this system). In contrast to this view, previous approaches in MT typically resolve examples like (1) by stating various constraints in terms of the source language (Nirenburg, 1987). As explained above, such constraints cannot be acquired automatically and therefore are usually limited in their coverage.

The experiments we conducted clearly show that statistics on lexical relations are very useful for disambiguation. Most notable is the result for the set of examples of Hebrew to English translation, which was picked randomly from foreign news sections in the Israeli press. For this set, the statistical model was applicable for 70% of the ambiguous words, and its selection was then correct for 92% of the cases. We cite also the results of a later experiment (Dagan, Marcus, and Markovitch, 1993) which tested a weaker variant of our method on texts in the computer domain, achieving a precision of 85%. Both results significantly improve upon a naive method which uses only a priori word probabilities. These results are comparable to recent reports in the literature (see Section 7). It should be emphasized, though, that our results were achieved for a realistic simulation of a broad coverage MT system, on randomly selected examples. We, therefore, believe that our figures reflect the expected performance of the algorithm in a practical

2 A similar observation underlies the use of parallel bilingual corpora for sense disambiguation (Brown et al., 1991; Gale, Church, and Yarowsky, 1992a). As we explain in Section 7, these corpora are a form of a manually tagged corpus, and are more difficult to obtain than monolingual corpora. Erroneously, the preliminary publication of our method (Dagan, Itai, and Schwall, 1991) was cited several times as requiring a parallel bilingual corpus.

implementation. On the other hand, most other results relate to a small number of words and senses which were determined by the experimenters.

Section 2 of the paper describes the linguistic model we use, employing a syntactic parser and a bilingual lexicon. Section 3 presents the statistical model, assuming a multinomial model for a single lexical relationship, and then using a constraint propagation algorithm to account simultaneously for all relations in the sentence. Section 4 describes the experimental setting. Section 5 presents and analyzes the results of the experiment, and cites additional results (Dagan, Marcus, and Markovitch, 1993). In Section 6 we analyze the limitations of the algorithm in different cases and suggest enhancements to improve it. We also discuss the possibility of adopting the algorithm for monolingual applications. Finally, in Section 7 we present a comparative analysis of statistical sense disambiguation methods and then conclude in Section 8.

2. The Linguistic Model

Our approach is first to use a bilingual lexicon to find all possible translations of each lexically ambiguous word in the source sentence, and then use statistical information gathered from target language corpora to choose the most appropriate alternative. To carry out this task we need the following linguistic tools, which are discussed in detail in the following sections:

1. Parsers for both the source language and the target language. These parsers should be capable of locating relevant syntactic relations, such as subject-verb, verb-object etc.
2. A bilingual lexicon which lists alternative translations for each source language word. If a word belongs to several syntactic categories, there should be a separate list for each one.
3. A procedure for mapping the source language syntactic relations to those of the target language.

Such tools have been implemented within the framework of many computational linguistic theories. We have used McCord's implementation of Slot Grammars (McCord, 1990; McCord, 1991). However, our method could have proceeded just as well using other linguistic models.

The linguistic model will be illustrated by the following Hebrew example, taken from the *Ha-Aretz* daily newspaper from September, 1990 (transcripted to Latin letters):

(3)	*Diplomatim*	*svurim*	*ki*	*hitztarrfuto*	*shell*	*Hon Sun*	*magdila*
	diplomats	believe	that	the joining	of	Hon Sun	increases

et	*ha-sikkuyim*	*l-hassagat*	*hitqaddmut*	*ba-sih̲ot.*
	the-chances	for-achieving	progress	in the-talks

Here, the ambiguous words in translation to English are *magdila*, *hitqaddmut* and *si<u>h</u>ot*. To facilitate the reading, we give the translation of the sentence into English, and in each case of an ambiguous selection all the alternatives are listed within curly brackets, the first alternative being the correct one.

(4) Diplomats believe that the joining of Hon Sun { increases | enlarges | magnifies } the chances for achieving { progress | advance | advancement } in the { talks | conversations | calls }.

The following subsections describe in detail the processing steps of the linguistic model. These include locating the ambiguous words and the relevant syntactic relationships among them in the source language sentence, mapping these relationships to alternative relationships in the target language and, finally, counting occurrences of these alternatives in a target language corpus.

2.1 Locating the ambiguous words in the source language

Our model defines the different "senses" of a source word to be all its possible translations to the target language, as listed in a bilingual lexicon. Some translations can be eliminated by the syntactic environment of the word in the source language. For example, in the following two sentences the word 'consider' should be translated differently into Hebrew, due to the different subcategorization frame in each case:

(5) I consider him smart.

(6) I consider going to Japan.

In these examples, the different syntactic subcategorization frames determine two different translations to Hebrew (*ma<u>h</u>shiv* vs. *shoqel*), thus eliminating some of the ambiguity. Such syntactic rules that allow us to resolve some of the ambiguities may be encoded in the lexicon (e.g. (Golan, Lappin, and Rimon, 1988)). However, many ambiguities cannot be resolved on syntactic grounds. The purpose of this work is to resolve the *remaining* ambiguities using lexical cooccurrence preferences, obtained by statistical methods.

2.2 Locating syntactic tuples in source language sentences

Our basic concept is the *syntactic tuple*, which denotes a syntactic relationship between two or more words. It is denoted by the name of the syntactic relation followed by a sequence of words that satisfies the relation, appearing in their base form (without morphological inflections). For example (subject-verb: *man walk*) is a syntactic tuple, which occurs in the sentence "The man walked home".

We assume that our parser (or an auxilliary program) can locate the syntactic relation corresponding to a given syntactic tuple in a sentence. The use of the base form of words is justified by the additional assumption that morphological inflections do not affect the probability of syntactic tuples. This assumption is not entirely accurate, but it has proven practically useful, and reduces the number of distinct tuples.

In our experience, the following syntactic relations proved useful for resolving ambiguities:

- Relations between a verb and its subject, complements and adjuncts, including direct and indirect objects, adverbs and modifying prepositional phrases.
- Relations between a noun and its complements and adjuncts, including adjectives, modifying nouns in noun compounds and modifying prepositional phrases.
- Relations between adjectives or adverbs and their modifiers.

As mentioned earlier, the full list of syntactic relations depends on the syntactic theory of the parser. Our model is general and does not depend on any particular list. However, we have found some desired properties in defining the relevant syntactic relations. One such property is the use of deep, or canonical, relations, as was already identified in (Grishman, Hirschman, and Nhan, 1986). This property was directly available from the ESG parser (McCord, 1990; McCord, 1991), which identifies the underlying syntactic function in constructs such as passives and relative clauses. We have also implemented an additional routine, which modified or filtered some of the relations received from the parser. This postprocessing routine mainly dealt with function words and prepositional phrases to get a set of more informative relations. For example, it combined the subject and complement of the verb "be" (as in "the *man* is *happy*") into a single relation. Likewise, a verb with its preposition and the head noun of a modifying prepositional phrase (as in *sit on* the *chair*) were also combined. The routine was designed to choose relations that impose considerable restrictions on the possible (or probable) lexical tuples. On the other hand, these relations should not be too specific, to allow statistically meaningful samples.

The first step in resolving an ambiguity is to find all the syntactic tuples containing the ambiguous words. For Example (3) we get the following syntactic tuples:

(7)
1. (Subject-verb: *hitztarrfut higdil*)
2. (Verb-object: *higdil sikkuy*)
3. (Verb-object: *hissig hitqaddmut*)
4. (Noun-prepositional phrase: *hitqaddmut b- si$\underline{h}$a*)

(these tuples translate as *joining-increase, increase-chance, achieve-progress* and *progress-in-talks*). In using these tuples, we expect to capture lexical constraints which are imposed by syntactic relationships.

a.	(subj-verb: *hitztarrfut higdil*)	*(subj-verb: joining increase)*	0
		(subj-verb: joining enlarge)	0
		(subj-verb: joining magnify)	0
b.	(verb-obj: *higdil sikkuy*)	*(verb-obj: increase chance)*	20
		(verb-obj: enlarge chance)	0
		(verb-obj: magnify chance)	0
c.	(verb-obj: *hissig hitqaddmut*)	*(verb-obj: achieve progress)*	29
		(verb-obj: achieve advance)	5
		(verb-obj: achieve advancement)	1
d.	(noun-pp: *hitqaddmut b- siha*)	*(noun-pp: progress in talk)*	7
		(noun-pp: progress in conversation)	0
		(noun-pp: progress in call)	0
		(noun-pp: advance in talk)	2
		(noun-pp: advance in conversation)	0
		(noun-pp: advance in call)	2
		(noun-pp: advancement in talk)	0
		(noun-pp: advancement in conversation)	0
		(noun-pp: advancement in call)	0

Table 1
The alternative target syntactic tuples with their counts in the target language corpus.

2.3 Mapping syntactic tuples to the target language

The set of syntactic tuples in the source language sentence is reflected in its translation to the target language. As a syntactic tuple is defined by both its syntactic relation and the words that appear in it, we need to map both components to the target language.

By definition, every ambiguous source language word maps to several target language words. We thus get several alternative target language tuples for each source language tuple that involves an ambiguous word. For example, for tuple 3 in (7) we obtain three alternatives, corresponding to the three different translations of the word *hitqaddmut*. For tuple 4 we obtain nine alternative target tuples, since each of the words *hitqaddmut* and *siha* maps to three different English words (see (4)). The full mapping of the Hebrew tuples in (7) to English tuples appears in Table 1 (the rightmost column should be ignored for the moment). Each of the tuple sets (a-d) in this table denotes the alternatives for translating the corresponding Hebrew tuple.

From a theoretical point of view, the mapping of syntactic relations is more problematic. There need not be a one-to-one mapping from source language relations to target language ones. In many cases the mapping depends on the words of the syntactic tuple, as seen in the following example of translating from German to English.

(8) *Der Tisch gefaellt mir.* — I like the table.

In this example the source language subject (*Tisch*), becomes the direct object (table) in the target, while the direct object (*mir*) in the source language becomes the subject (I) in the target. Therefore, the German lexical tuples

(9) *(subj-verb: Tisch gefaellt)*
(verb-obj: gefaellt mir)

are mapped to the following English lexical tuples

(10) *(verb-obj: like table)*
(subj-verb: I like)

(The Hebrew equivalent is similar to the German structure).

In practice this is less of a problem. In most cases, the source language relation has a direct equivalent in the target language. In many other cases, transformation rules can be encoded, either in the lexicon (if they are word dependent) or as syntactic transformations. These rules are usually available in machine translation systems which use the transfer method, as this knowledge is required in order to generate target language structures.

To further facilitate the mapping of syntactic relations and to avoid errors due to fine distinctions between them, we grouped related syntactic relations into a single “general class”, and mapped this class to the target language. The important classes used were relations between a verb and its arguments and modifiers (counting as one class all objects, indirect objects, complements and nouns in modifying prepositional phrases) and between a noun and its arguments and modifiers (counting as one class all modifying nouns in compounds and nouns in modifying prepositional phrases). The classification enables us to get more statistical data for each class, as it reduces the number of relations. The success of using this general level of syntactic relations indicates that even a rough mapping of source to target language relations is useful for the statistical model.

2.4 Counting lexical tuples in the target language corpus

We now wish to determine the plausibility of each alternative target word being the translation of an ambiguous source word. In our model, the plausibility of selecting a target word is determined by the plausibility of the tuples which are obtained from it. The plausibility of alternative target tuples is in turn determined by their relative frequency in the corpus.

Target syntactic tuples are identified in the corpus similarly to source language tuples, i.e., by a target language parser and a companion routine as described in Section 2.1. The right column of Table 1 shows the counts obtained for the lexical tuples of our example in the corpora we used. The table reveals that the tuples containing the correct target word (‘talk’, ‘progress’, ‘increase’) are indeed more frequent. However, we still need a decision algorithm to analyze the statistical significance of the data, and choose the appropriate word accordingly.

3. The Statistical Model

As seen in the previous section, the linguistic model maps each source language syntactic tuple to several alternative target tuples, where each alternative corresponds to a different selection of target words. We wish to select the most plausible target language word for each ambiguous source language word, basing our decision on the counts obtained from the target corpus, as illustrated in Table 1. To that end we should define a selection algorithm, whose outcome depends on all the lexical tuples in the sentence. If the data obtained from the corpus do not substantially support any one of the alternatives, the algorithm should notify the translation system that it cannot reach a statistically meaningful decision.

Our algorithm is based on a statistical model. However, we wish to point out that we do not see the statistical considerations, as expressed in the model, as fully reflecting the linguistic considerations (syntactic, semantic or pragmatic) that determine the right translation. The model reflects only part of the relevant data, and in addition makes statistical assumptions that are only partially satisfied. Therefore, a statistically based model need not make the correct linguistic choices. The performance of the model can only be empirically evaluated, the statistical considerations serve only as heuristics. The role of the statistical considerations is, therefore, to guide us in constructing heuristics that make use of the linguistic data of the sample (the corpus). Our experience shows that the statistical methods are indeed very helpful in establishing and comparing useful decision criteria which reflect various linguistic considerations.

3.1 The Probabilistic Model

First we discuss decisions based on a single syntactic tuple (as when only one syntactic tuple in the sentence contains an ambiguous word). Denote the source language syntactic tuple T, and let there be k alternative target tuples for T, denoted by $T_1, ..., T_k$. Let the counts obtained for the target tuples be $n_1, ..., n_k$. For notational convenience we number the tuples by decreasing frequency, i.e., $n_1 \geq n_2 \geq ... \geq n_k$.

Since our goal is to choose for T one of the target tuples T_i, we can consider T a discrete random variable with multinomial distribution,[3] whose possible values are $T_1, ..., T_k$. Let p_i be the probability of obtaining T_i, i.e., the probability that T_i is the correct translation for T. We estimate the probabilities p_i by the counts n_i in the obvious way, using the *maximum likelihood estimator* (Agresti, 1990, pp. 40-41). The estimator $\hat{p}_i$ for p_i is

$$\hat{p}_i = \frac{n_i}{\sum_{j=1}^{k} n_j}. \tag{1}$$

The precision of the estimator depends, of course, on the size of the counts in the computation. We will incorporate this consideration into the decision algorithm by using confidence intervals.[4]

3 A variable that can have one of a finite set of values, each of them having a fixed probability.
4 The maximum likelihood estimator is known to give poor estimates when small counts are involved, and there are several methods to improve it (see (Church and Gale, 1991) for a presentation and discussion of several methods). For our needs this is not necessary in most cases, since we are not going to use the estimate itself, but rather a confidence interval for the ratio between two

We now have to establish the criterion for choosing the preferred target language lexical tuple. The most reasonable assumption is to choose the tuple with the highest estimated probability, that is T_1 – the tuple with the largest observed frequency. According to the model, the probability that T_1 is the right choice is estimated as $\hat{p_1}$. This criterion should be subject to the condition that the difference between the alternative probabilities is significant. For example, if $\hat{p_1} = 0.51$ and $\hat{p_2} = 0.49$ the expected success rate in choosing T_1 is approximately 0.5. To prevent the system from making a decision in such cases, we need to impose some conditions on the probabilities p_i.

One possible such condition is that $\hat{p_1}$ exceed a prespecified threshold (or, as we shall describe below, that the threshold requirement be applied to a confidence interval). According to the model, this requirement ensures that the success probability of every decision exceeds the threshold. Even though this method satisfies the probabilistic model, it is vulnerable to noise in the data, which often causes some relatively small counts to be larger than their true value in the sample. The noise is introduced in part by inaccuracies in the model and in part because of errors during the automatic collection of the statistical data. Consequently, the estimated value of p_1 may be smaller than its true value, because other counts in Equation (1) are too large, thus, preventing p_1 from passing the threshold.

To deal with this problem, we have chosen another criterion for significance – the *odds ratio*. We choose the alternative T_1 only if all the ratios

$$\frac{\hat{p_1}}{\hat{p_2}}, \frac{\hat{p_1}}{\hat{p_3}}, \ldots, \frac{\hat{p_1}}{\hat{p_k}}$$

exceed a prespecified threshold. Note that $\frac{\hat{p_i}}{\hat{p_j}} = \frac{n_i}{n_j}$ and since $n_1 \geq n_2 \geq \ldots \geq n_k$, the ratio $\frac{\hat{p_1}}{\hat{p_2}}$ is less than or equal to all the other ratios. Therefore, it suffices to check the odds ratio only for $\frac{\hat{p_1}}{\hat{p_2}}$. This criterion is less sensitive to noise of the abovementioned type than $\hat{p}_1$, since it depends only on the two largest counts.

3.1.1 Underlying Assumptions. The use of a probabilistic model necessarily introduces several assumptions on the structure of the corresponding linguistic data. It is important to point out these assumptions, in order to be aware of possible inconsistencies between the model and the linguistic phenomena for which it is used.

The first assumption is introduced by the use of a multinomial model, which presupposes the following:

Assumption

The events T_i are mutually disjoint.

This assumption is not entirely valid, since sometimes it is possible to translate a source language word to several target language words, such that all the translations are valid. For example consider the Hebrew sentence (from the *Ha-Aretz* daily newspaper, November 27th, 1990) whose English translation is:

estimations (see below).

(11) The resignation of Thatcher is not { related | connected } to the negotiations with Damascus.

In this sentence (but not in others), the ambiguous word *qshura* can equally well be translated to either 'related' or 'connected'. In terms of the probabilistic model, the two corresponding events, i.e., the two alternative English tuples which contain these words, T_1 =(verb-complement: *relate to negotiation*) and T_2 =(verb-complement: *connect to negotiation*) are both correct, thus the events T_1 and T_2 both occur (they are not disjoint). However, we have to make this assumption, since the counts we have, n_i, from which we estimate the probabilities of the T_i's, count actual occurrences of single lexical tuples. In other words, we count the number of times that each of T_1 and T_2 *actually* occur, not the number of times in which each of them *could* occur.

Two additional assumptions are introduced by using counts of the occurrences of syntactic tuples of the *target* language in order to estimate the translation probabilities of *source* language tuples:

Assumption

An occurrence of the source language syntactic tuple T can indeed be translated to one of $T_1, ..., T_k$.

Assumption

Every occurrence of the target tuple T_i can be the translation of only the source tuple T.

Assumption is an assumption on the completeness of the linguistic model. It is rather reasonable, and depends on the completeness of our bilingual lexicon: If the lexicon gives all possible translations of each ambiguous word, then this assumption will hold, since for each lexical tuple T we will produce all possible translations.[5]

Assumption , which may be viewed as a soundness assumptions, does not always hold, since a target language word may be the translation of several source language words. Consider for example the Hebrew tuple T =(verb-obj: *he<u>h</u>ziq lul*). *Lul* is ambiguous, meaning either a playpen or a chicken pen. Accordingly, T can be translated to either T_1 =(verb-obj: *hold playpen*) or T_2 =(verb-obj: *hold pen*). In the context of 'hold' the first translation is more likely and we can, therefore, expect our model to prefer T_1. However, this might not be the case because Assumption is contradicted. 'Pen' can also be the translation of the Hebrew word *'et* (the writing instrument), and thus T_2 can be the translation of another Hebrew tuple, T'=(verb-obj: *he<u>h</u>ziq 'et*). This means that

5 The problem of constructing a bilingual lexicon that is as complete as possible is beyond the scope of this paper. A promising approach may be to use aligned bilingual corpora, especially for augmenting existing lexicons with domain specific terminology (Brown et al., 1993; Dagan, Church, and Gale, 1993). In any case, it seems that any translation system is limited by the completeness of its bilingual lexicon, which makes our assumption a reasonable one.

when translating T we are counting occurrences of T_2 that correspond to both T and T', "misleading" the selection criterion. Section 6.3 illustrates another example in which the assumption is not valid, causing the algorithm to fail to select the correct translation.

We must make this assumption since we use only a target language corpus, which is not related to any source language information.[6] Therefore, when seeing an occurrence of the target language word w, we do not know which source language word is appropriate in the current context. Consequently, we count its occurrence as a translation of all the source language words for which w is a possible translation. This implies that sometimes we use inaccurate data, which introduces noise into the statistical model (See section 6.3 for discussion of an alternative, but expensive, solution, using a bilingual corpus). As we shall see, even though the assumption does not always hold, in most cases this noise does not interfere with the decision algorithm.

3.2 Statistical Significance of the Decision

Another problem we should address is the statistical significance of the data – what confidence do we have that the data indeed reflects the phenomenon. If the decision is based on small counts then the difference in the counts might be due to chance. For example, we should have more confidence on the odds ratio $\frac{\hat{p_1}}{\hat{p_2}} = 3$ when $n_1 = 30$ and $n_2 = 10$, than when $n_1 = 3$ and $n_2 = 1$. Consequently, we shall use a dynamic threshold for $\frac{\hat{p_1}}{\hat{p_2}}$, which is large when the counts are small, and decreases as the counts increase.

A common method for determining the statistical significance of estimates is the use of confidence intervals. Rather than finding a confidence interval for $\frac{\hat{p_1}}{\hat{p_2}}$, we will bound the log odds ratio, $\ln \frac{\hat{p_1}}{\hat{p_2}}$. Since the variance of the log odds ratio is independent of the mean, it converges to the normal distribution faster than the odds ratio itself (Agresti, 1990). We use a one-tailed interval, as we only want to decide whether $\ln \frac{\hat{p_1}}{\hat{p_2}}$ is greater than a specific threshold (i.e., we need only a lower bound for $\ln \frac{\hat{p_1}}{\hat{p_2}}$). Using this method, for each desired error probability $0 < \alpha < 1$, we may determine a value B_α and state that with probability at least $1 - \alpha$ the true value, $\ln \frac{p_1}{p_2}$, is greater than B_α.

The confidence interval of a random variable X with normal distribution is $Z_{1-\alpha}\sqrt{varX}$, where $Z_{1-\alpha}$ is the confidence coefficient, which may be found in statistical tables, and var is the variance. In our case, the size of the confidence interval is $Z_{1-\alpha}\sqrt{var\left[\ln \frac{\hat{p_1}}{\hat{p_2}}\right]}$. In the appendix we approximate the variance by the following

$$var\left[\ln \frac{\hat{p_1}}{\hat{p_2}}\right] \approx \frac{1}{n_1} + \frac{1}{n_2}.$$

The bound we get is thus:

$$\ln\left(\frac{p_1}{p_2}\right) \geq \ln\left(\frac{\hat{p_1}}{\hat{p_2}}\right) - Z_{1-\alpha}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad .$$

6 As explained in the introduction, this is a very important advantage of our method over other methods that use bilingual corpora.

Since $\frac{\hat{p_1}}{\hat{p_2}} = \frac{n_1}{n_2}$ we get:

$$\ln\left(\frac{p_1}{p_2}\right) \geq \underbrace{\ln\left(\frac{n_1}{n_2}\right) - Z_{1-\alpha}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}_{B_\alpha} \tag{2}$$

$B_\alpha(n_1, n_2)$ (or B_α when n_1 and n_2 are understood from the context) is defined to be the right hand side of inequality (2). The meaning of the inequality is that for every given pair n_1, n_2 we know with confidence $1-\alpha$ that

$$\ln\frac{p_1}{p_2} \geq B_\alpha \tag{3}$$

or in other words, B_α is a lower bound for $\ln\frac{p_1}{p_2}$ with this confidence level.

In order to obtain a decision criterion we choose a threshold θ, for B_α, and decide to choose T_1 only if

$$B_\alpha \geq \theta. \tag{4}$$

If inequality (4) does not hold the algorithm makes no decision. The meaning of this criterion is that only if we know with confidence of at least $1-\alpha$ that $\ln\frac{p_1}{p_2} \geq \theta$, we will select the most frequent tuple T_1 as the appropriate one. In terms of statistical decision theory, we say that our null hypothesis is that $\ln\frac{p_1}{p_2} < \theta$, and we will make a decision only if we can reject this hypothesis with confidence at least $1-\alpha$. Note that we cannot compute B_α when one of the counts is zero. In this case we have used the common correction method of adding 0.5 to each of the counts (Agresti, 1990, p. 249).[7]

We shall now demonstrate the use of the decision criterion. In the experiment we conducted we chose the parameters $\alpha = 0.1$, for which $Z_\alpha = 1.282$, and $\theta = 0.2$. Thus, in order to choose T_1 we require that with confidence level of at least 90% the hypothesis should satisfy $\ln\frac{p_1}{p_2} \geq 0.2$ (i.e., $\frac{p_1}{p_2} \geq e^{0.2} = 1.22$). For the alternative translations of tuple c in Table 1 we got $n_1 = 29$ and $n_2 = 5$. For these values $B_\alpha = 1.137$. In this case inequality (4) is satisfied for $\theta = 0.2$, and the algorithm selects the word 'progress' as the translation of the Hebrew word *hitqaddmut*.

In another case we had to translate the Hebrew word *ro'sh*, which can be translated to either 'top' or 'head', in the sentence whose translation is:

(12) Sihanuk stood at the {top | head} of a coalition of underground groups.

The two alternative syntactic tuples were

a. *(verb-pp: stand at head)* 10
b. *(verb-pp: stand at top)* 5

For $n_1 = 10$ and $n_2 = 5$ we get $B_\alpha = -0.009$ (a negative value means that it is impossible to assure with confidence level 90% that $p_1 > p_2$). Since $B_\alpha \leq 0.2$, the

7 In this case, smoothing methods (Church and Gale, 1991) may improve the correction method.

algorithm will refrain from making a decision in this case. This abstention reflects the fact that the difference between the counts is not statistically significant and choosing the first alternative can be wrong in many of the cases (as seen in the five cases that were observed in the corpus).

As mentioned above, our motivation was to find a criterion that depends on a dynamic threshold for $\frac{\hat{p_1}}{\hat{p_2}}$ (or alternatively $\frac{n_1}{n_2}$), so that the threshold will be higher when n_1 and n_2 are smaller. Our criterion indeed satisfies this requirement. If we substitute B_α in Equation (4) we get the following equivalent criterion

$$\ln \frac{n_1}{n_2} \geq \theta + Z_{1-\alpha}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The above inequality clarifies the roles of the two parameters, α and θ: θ specifies a lower bound on $\ln \frac{n_1}{n_2}$, which is independent of the sample size; α reflects the statistical significance: if α is decreased (i.e., we require more confidence), $Z_{1-\alpha}$ will increase, and, therefore, the component dependent on the sample size will increase. Since this component is in inverse relation to n_1 and n_2, the penalty for decreasing α increases when the sample size decreases. From this analysis we can derive the criterion for choosing the parameters: If we wish to use small counts, then α should be small, and θ depends on the required ratio between n_1 and n_2. The optimal values of the parameters should be determined empirically, and might depend on the corpora and parsers we use.

3.3 Sentences with Several Syntactic Relations

In the previous section, we assumed that the source sentence contains only one ambiguous syntactic tuple. In general there may be several ambiguous words which appear in several tuples. We should take advantage of the occurrence patterns of all of the tuples to reach a decision. Since different relations may favor different translations for an ambiguous word, we should devise a strategy for selecting a consistent translation for all words in the sentence. We have used the following constraint propagation algorithm, which receives as input the list of all source tuples, along with their alternative translations to target tuples:

1. Compute B_α of each source tuple. If the largest B_α is less than the threshold, θ, then stop.
2. Let T be the source tuple for which B_α is maximal. Select the translation for the ambiguous words (or word) in T according to T_1 (the most frequent target alternative for T). Remove T from the list of source tuples.
3. Propagate the constraint: Eliminate target tuples that are inconsistent with this decision. If now some source tuples become unambiguous, remove them from the list of source tuples.
4. Repeat this procedure for the remaining list of source tuples, until all ambiguities have been resolved, or the maximal B_α is less than θ.

To illustrate the algorithm, we consider Table (1) using the parameters $\alpha = 0.1$ and $\theta = 0.2$. The largest value of B_α occurs for the tuple (verb-object: *higdil sikkuy*), for which *higdil* can be translated as 'increase', 'magnify', and 'enlarge'. The first alternative appeared $n_1 = 20$ times and the other alternatives did not appear at all, $(n_2 = n_3 = 0)$. Adding the correction factor and computing B_α yields $B_\alpha(n_1 + 0.5, n_2 + 0.5) = B_\alpha(20.5, 0.5) = 1.879 > 0.2 = \theta$. Therefore, the word 'increase' is chosen as the translation of *higdil.* Since this word appears also in the tuple (Subject-verb: *hitztarrfut higdil*) the target tuples that include alternative translations of *higdil* are deleted. Thus

(13) (subj-verb: *joining enlarge)*
(subj-verb: *joining magnify)*

were deleted. This leaves us with only one alternative, *(subj-verb: joining increase)*, as a possible translation of this Hebrew tuple, which is, therefore, removed from the input list.

We now recompute the values of B_α for the remaining tuples. The maximal value is obtained for the tuple

(14) (verb-object: *hissig hitqaddmut*)

where $B_\alpha(29, 5) = 1.137 > \theta$. We, therefore, choose the word 'progress' as a translation for *hitqaddmut.* Since this word, *hitqaddmut*, also appears in the tuple (noun-prepositional phrase: *hitqaddmut b- siha*) we delete the six target tuples that are inconsistent with the selection of 'progress' (those containing the words 'advance' and 'advancement'). There now remain only three alternative target tuples for *hitqaddmut b- siha.*

We now recompute the values of B_α. The maximum value is $B_\alpha(7.5, 0.5) = 0.836 > \theta$ (note that because tuples inconsistent with the previous decisions were eliminated, n_2 dropped from 2 to 0, thus increasing B_α). Thus, 'talk' is selected as the translation of *siha.* Now all the ambiguities have been resolved and the procedure stops.

In the above example all the ambiguities were resolved since in each stage the value of B_α exceeded the threshold $\theta = 0.2$. In some cases not all ambiguities are resolved, though the number of ambiguities may decrease.

It should be noted that other methods may be proposed for combining the statistics of several syntactic relations. For example, it may make sense to multiply estimates of conditional probabilities of tuples in different relations, in a way which is analogous to n-gram language modeling (Jelinek, Mercer, and Roukos, 1992). However, such an approach will make it harder to take into account the statistical significance of the estimate (a criterion which is missing in standard n-gram models). In our set of examples, the constraint propagation method proved to be successful, and did not seem to introduce any errors. Further experimentation, on much larger data sets, is needed to determine which of the two methods (if any) is substantially superior to the other.

4. The Experiment

In order to evaluate the proposed disambiguation method, we implemented and tested the method on a random set of examples. The examples consisted of a set of Hebrew paragraphs, and a set of German paragraphs. In both cases the target language was English. The Hebrew examples consisted of ten paragraphs picked at random from Foreign News sections of the Israeli press. The paragraphs were selected from several news items and articles that appeared in several daily news papers. The target language corpus consisted of American newspaper articles, and the Hansard corpus of the proceedings of the Canadian parliament. The domain of foreign news articles was chosen to correspond to some of the topics which appear in the English corpus.[8] The German examples were chosen at random from the German press, without restricting the topic.[9]

Since we did not have a translation system from Hebrew or German to English, we simulated the steps such a system would perform. Hence, the results we report measure the performance of just the target word selection module, and not the performance of a complete translation system. The latter can be expected to be somewhat lower for a real system, depending on the performance of its other components. Note, however, that since the disambiguation module is highly immune to noise it might be more useful in a real system: In such a system some of the alternatives would be totally erroneous. Since the corresponding lexical tuples would typically not be found in the corpora, they would be eliminated by our module.

The experiment is described in detail in the following subsections. It provides an example of a thorough evaluation which is carried out without having a complete system available. We specifically describe the processing of the Hebrew data, which was performed by a professional translator, supervised by the authors. The German examples were processed very similarly.

4.1 Locating ambiguous words

To locate ambiguous words we simulated a bi-lingual lexicon and syntactic filters of a translation system. For every source language word, the translator searched all possible translations using a Hebrew-English dictionary (Alcalay, 1990). The list of translations proposed by the dictionary was modified according to the following guidelines, to better reflect the lexicon of a practical translation system:

1. Eliminate translations that would be ruled out for syntactic reasons, as explained in Section 2.1.
2. Consider only content words, ignoring function words and proper nouns.
3. Assume that multi-word terms, such as "prime minister", appear in the lexicon as complete terms. Thus we did not consider each of their

8 The corpus includes many irrelevant topics as well, which introduce noisy data with respect to the given domain.
9 The German examples were prepared by Ulrike Schwall from the IBM Scientific center, Heidelberg, Germany.

constituents separately. Also, we did not consider source language words that should be translated to a multi-word target phrase.

4. Eliminate rare and archaic translations that are not expected in the context of foreign affairs in current press.

5. The professional translator added translations that were missing in the dictionary.

In addition, each of the remaining target alternatives for each source word was evaluated as to whether it was a suitable translation in the current context. This evaluation was later used to judge the selections of the algorithm. If all the alternatives were considered suitable, then the source word was eliminated from the test set, since any decision for it would have been considered successful.

We ended up with 103 Hebrew and 54 German ambiguous words. For each Hebrew word we had an average of 3.27 alternative translations, and an average of 1.44 correct translations. The average number of translations of a German word was 3.26, and 1.33 correct translations.

4.2 Determining the syntactic tuples and mapping them to English

Since we did not have a Hebrew parser, we have simulated the two steps of determining the source syntactic tuples and mapping them to English by reversing the order of these steps, in the following way. First, the sample sentences were translated manually, as literally as possible, into English. Then, the resulting English sentences were analyzed, using the ESG parser and the post processing routine (see Section 2.2), to identify the relevant syntactic tuples. The tuples were further classified into "general classes", as described in Section 2.3. The use of these general classes, which was intended to facilitate the mapping of syntactic relations from one language to another, also facilitated our simulation method and caused it to produce realistic output.

At the end of the procedure, we had, for each sample sentence, a data structure similar to Table 1 (without the counts).

4.3 Acquiring the statistical data

The statistical data were acquired from the following corpora:

- Texts from The Washington Post – 40 million words.
- The Hansard corpus of protocols of the Canadian Parliament – 85 million words.
- Associated Press news items – 24 million words.

However, the effective size of the corpora was only about 25 million words, due to two filtering criteria. First, we considered only sentences whose length did not exceed 25 words, since longer sentences required excessive parse time and contained many parsing errors. Second, even 35% of the shorter sentences failed to parse and had to be eliminated. The syntactic tuples were located by the ESG parser and the post processing routine mentioned earlier.

For the purpose of evaluation, we gathered only the data required for the given test examples. Within a practical machine translation system, the disambiguation module would require a data base containing all the lexical tuples of the corpus with their frequency counts. In the current research project we did not have the computing resources necessary for constructing the complete data base (the major cost being parsing time). However, such resources are not needed in order to *evaluate* the proposed method. Since we evaluated the method only on a relatively small number of random sentences, we first constructed the set of all "relevant" target tuples, i.e. tuples that should be considered for the test sentences. Then we scanned the entire corpus and extracted only sentences that contain both words from at least one of the "relevant" tuples. Only the extracted sentences were parsed, and their counts recorded in our data base. Even though this data base is much smaller than the full data base, for the ambiguous words of the test sentences both data bases provide the *same* information. Thus, the success rate for the test sentences is the same for both methods, while requiring a considerably smaller amount of resources at the research phase.

The problem with this method is that for every set of sample sentences the entire corpus has to be scanned. Thus, a practical system would have to preprocess the corpus to construct a data base of the entire corpus. Then, in order to resolve ambiguities, only this data base need be consulted.

After acquiring all the relevant data, the algorithm of Section 3.3 was executed for each of the test sentences.

5. Evaluation

Two measurements, *applicability* and *precision*, are used to evaluate the performance of the algorithm. The applicability (coverage) denotes the proportion of cases for which the model performed a selection, i.e. those cases for which the bound B_α passed the threshold. The precision denotes the proportion of cases for which the model performed a correct selection out of all the applicable cases.

We compare the precision of our method, which we term "TWS" (for Target Word Selection"), to that of the "Word Frequencies" procedure, which always selects the most frequent target word. In other words, the Word Frequencies method prefers the alternative which has the highest a priori probability to appear in the target language corpus. This naive "straw-man" is less sophisticated than other methods suggested in the literature but it is useful as a common benchmark since it can be easily implemented. The success rate of the Word Frequencies procedure can serve as a measure for the degree of lexical ambiguity in a given set of examples, and thus different methods can be partly compared by their degree of success relative to this procedure.

Out of the 103 ambiguous Hebrew words, for 33 the bound B_α did not pass the threshold, achieving an applicability of 68%. The remaining 70 examples were distributed according to Table 2. Thus the precision of the statistical model was 91% (64/70)[10] while

		Word Frequencies		
		correct	*incorrect*	**Total**
TWS	*correct*	42	22	64
	incorrect	2	4	6
	Total	44	26	70

Table 2
Hebrew-English translation: Comparison of *TWS* and *Word Frequencies* methods for the 70 applicable examples.

relying just on Word Frequencies yields 63% (44/70), providing an improvement of 28%. The table demonstrates that our algorithm corrects 22 erroneous decisions of the Word Frequency method, but makes only two errors that the word frequency method translates correctly. This implies that with high confidence our method greatly improves the Word Frequencies method.

The number of Hebrew examples is large enough to permit a meaningful analysis of the statistical significance of the results. By computing confidence intervals for the distribution of proportions we claim that with 95% confidence our method succeeds in at least 86% of the applicable examples. This means that though the figure of 91% might be due to a lucky selection of the random examples, there is only 5% chance that the real figure is less than 86% (for the given domain and corpus). The confidence interval was computed as follows:

$$p \geq \hat{p} - Z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{64}{70} - 1.65\sqrt{\frac{\frac{64}{70}\cdot\frac{6}{70}}{70}} = 0.86,$$

where $\alpha = 0.05$ and the variance is estimated by $\frac{\hat{p}(1-\hat{p})}{n}$.

With the same confidence, our method improves the Word Frequencies method by at least 18% (relative to the actual improvement of 28% in the given test set). Let p_1 be the proportion of cases where our method succeeds and the Word Frequencies method fails ($p_1 = \frac{22}{70}$) and p_2 be the proportion of cases where the Word Frequencies method succeeds and ours fails ($p_2 = \frac{2}{70}$). The confidence interval is for the difference of proportions in multinomial distribution and is computed as follows:

$$\begin{aligned} p_1 - p_2 &\leq \hat{p_1} - \hat{p_2} - Z_{1-\alpha}\sqrt{var(\hat{p_1} - \hat{p_2})} \\ &= \hat{p_1} - \hat{p_2} - Z_{1-\alpha}\frac{1}{\sqrt{n}}\sqrt{\hat{p_1}(1-\hat{p_1}) + \hat{p_2}(1-\hat{p_2}) + 2\hat{p_1}\hat{p_2}} \\ &= \frac{22}{70} - \frac{2}{70} - 1.65\frac{1}{\sqrt{70}}\sqrt{\frac{22\cdot(70-22) + 2\cdot(70-2) + 2\cdot 22\cdot 2}{70^2}} = 0.18. \end{aligned}$$

Out of the 54 ambiguous German words, for 27 the bound B_α did not pass the threshold (applicability of 50%). The remaining 27 examples were distributed according

10 An a posteriori observation showed that in three of the six errors the selection of the model was actually acceptable, and the a priori judgment of the human translator was too restrictive. For example, in one of these cases the statistics selected the expression 'to begin the talks' while the human translator regarded this expression as incorrect and selected 'to start the talks'. If we consider these cases as correct then there are only three selection errors, getting 96% precision.

		Word Frequencies *correct*	Word Frequencies *incorrect*	Total
TWS	*correct*	15	6	21
	incorrect	0	6	6
	Total	15	12	27

Table 3
German-English translation: Comparison of *TWS* and *Word Frequencies* methods for the 27 applicable examples.

to Table 3. Thus, the precision of the statistical model was 78% (21/27), while relying just on Word Frequencies yields 56% (15/27). Here our method corrected 6 errors of the Word Frequencies method, without causing any new errors. We attribute the lower success rate for the German examples to the fact that they were not restricted to topics that are well represented in the corpus. This poor correspondence between the training and testing texts is reflected also by the low precision of the Word Frequencies method. This means that the a priori probability of the target words, as estimated from the training corpora, provides a very poor prediction of the correct selection in the test examples. Relative to the a priori probability, the precision of our method is still 22% higher.

5.1 Additional Results

Recently, Dagan, Marcus and Markovitch have implemented a variant of the disambiguation method of the current paper. This variant was developed for evaluating a method which estimates the probability of word combinations which do not occur in the training corpus (Dagan, Marcus, and Markovitch, 1993). In this section we quote their results, providing additional evidence for the effectiveness of the TWS method.

The major difference between the TWS method, as presented in this paper, and the variant described in (Dagan, Marcus, and Markovitch, 1993), which we term TWS', is that the latter does not use any parsing for collecting the statistics from the corpus. Instead, the counts of syntactic tuples are approximated by counting cooccurrences of the given words of the tuple within a short distance in a sentence. The approximation takes into account the relative order between the words of the tuple, such that occurrences of a certain syntactic relation are approximated only by word cooccurrences which preserve the most frequent word order for that relation (e.g. an adjective precedes the noun it modifies).

The TWS' method still assumes that the source sentence to be translated is being parsed, in order to identify the words which are syntactically related to an ambiguous word. This model is therefore relevant for translation systems that use a parser for the source language, but may not have available a robust target language parser.

The corpus used for evaluating the TWS' method consists of articles posted to the USENET news system. The articles were collected from news groups that discuss computer related topics. The length of the corpus is 8,871,125 words (tokens), and the lexicon size (distinct types, at the string level) is 95,559. The type of text in this corpus is quite noisy, including short and incomplete sentences as well as much irrelevant information, such as person and device names. The test set used for the experiment consists of 78 Hebrew sentences that were taken out of a book about computers. These sentences were

		Word Frequencies		
		correct	*incorrect*	**Total**
TWS' method	*correct*	120	28	148
	incorrect	3	22	25
	Total	123	50	173

Table 4
Comparison of *TWS'* and *Word Frequencies* methods for the 173 applicable examples.

processed as described in Section 4, obtaining a set of 269 ambiguous Hebrew words. The average number of alternative translations per ambiguous word in this set is 5.8, and 1.35 correct translations.

Out of the 269 ambiguous Hebrew words, for 96 the bound B_α did not pass the threshold, achieving an applicability of 64.3%. The remaining 173 examples were distributed according to Table 4. For the words that are covered by the TWS' method, the Word Frequencies method has a precision of 71.1%$(\frac{123}{173})$, while the TWS' method has a precision of 85.5%$(\frac{148}{173})$. As can be seen in the table, the TWS' method is correct in almost all the cases it disagrees with the Word Frequencies method (28 out of 31). The applicability and precision figures in this experiment are somewhat lower than those achieved for the Hebrew set in our original evaluation of the TWS method (Table 2). We attribute this to the fact that the original results were achieved using a parsed corpus, which was about 2.5 times larger and of much higher quality than the one used in the second experiment. Yet, the new results give additional support for the usefulness of the TWS method, even for noisy data provided by a low quality corpus, without any parsing or tagging.[11]

6. Analysis and Possible Enhancements

In this section we give a detailed analysis of the selections performed by the algorithm, and in particular analyze when it failed. The analysis of these modes suggests possible improvements of the model and also indicates its limitations. As described earlier, the algorithm's failure includes either the cases for which the method was not applicable (no selection) or the cases in which it made an incorrect selection. The following paragraphs list various reasons for both types. At the end of the section, we discuss the possibility of adapting our approach to monolingual applications.

6.1 Correct Selection

In the examples that were treated correctly by our method, such as the examples given in the previous sections, the statistics succeeded in capturing two major types of disambiguating data. In preferring 'sign-treaty' over 'seal-treaty' (in Example 1), the statistics reflect the relevant semantic constraint. In preferring 'peace-treaty' over 'peace-contract',

11 It should be mentioned that the work of Dagan, Marcus and Markovitch (1993) includes further results, evaluating an enhancement of the TWS method using their similarity-based estimation method. This enhancement is beyond the scope of the current paper, and is referred to in the next section.

the statistics reflect the lexical usage of 'treaty' in English which differs from the usage of 'contract'.

6.2 Inapplicability

Insufficient data. This was the reason for nearly all the cases of inapplicability. In one of our examples, for instance, none of the alternative relations 'an investigator of corruption' (the correct one) or 'researcher of corruption' (the incorrect one) was observed in the parsed corpus. In this case it is possible to perform the correct selection if we use only statistics about the cooccurrence of 'corruption' with either 'investigator' or 'researcher' in the same local context, without requiring any syntactic relationship. Statistics on cooccurrence of words in a local context were used recently for monolingual word sense disambiguation (Gale, Church, and Yarowsky, 1993; Gale, Church, and Yarowsky, 1992b; Schütze, 1992; Schütze, 1993) (see Section 7 for more details, and (Church and Hanks, 1990; Smadja, 1993) for other applications of these statistics). It is possible to apply these methods using statistics for the target language, and thus incorporate them within the framework proposed here for target word selection. Finding an optimal way of combining the different methods is a subject for further research. Our intuition, though, as well as some of our initial data, suggest that statistics on word cooccurrence in the local context can substantially increase the applicability of the selection method.

Another way to deal with the lack of statistical data for the specific words in question is to use statistics about similar words. This is the basis for Sadler's *Analogical Semantics* (Sadler, 1989) which has not proved effective. His results may be improved if more sophisticated methods and larger corpora are used to establish similarity between words (such as in (Hindle, 1990)). In particular, an enhancement of our disambiguation method, using similarity-based estimation (Dagan, Marcus, and Markovitch, 1993), was evaluated recently. In this evaluation the applicability of the disambiguation method was increased by 15%, with only a slight decrease in the precision. The increased applicability was achieved by disambiguating additional cases in which statistical data was not available for any of the alternative tuples, while it was available for other tuples containing similar words.

Conflicting data. In very few cases two alternatives were supported equally by the statistical data, thus, preventing a selection. In such cases, both alternatives are valid at the independent level of the syntactic relation, but may be inappropriate for the specific context. For instance, the two alternatives of 'to take a job' or 'to take a position' appeared in one of the examples, but since the general context was about the position of a prime minister only the latter was appropriate. In order to resolve such ambiguities it may be useful to consider also cooccurrences of the ambiguous word with other words in the broader context (e.g. (Gale, Church, and Yarowsky, 1993; Yarowsky, 1992)). For instance, the word 'minister' seems to cooccur in the same context more frequently with 'position' than with 'job'.

In another example both alternatives were appropriate also for the specific context. This happened with the German verb *werfen*, which may be translated (among other options) as 'throw', 'cast' or 'score'. In our example, *werfen*, appeared in the context of 'to throw/cast light' and these two correct alternatives had equal frequencies in the corpus ('score' was successfully eliminated). In such situations any selection between the alternatives will be appropriate and, therefore, any algorithm that handles conflicting data would work properly. However, it is difficult to automatically decide when both alternatives are acceptable, and when only one of them is.

6.3 Incorrect Selection

Using the inappropriate relation. One of the examples contained the Hebrew word *matzav*. This word has several translations, two of which are 'state' and 'position'. The phrase which contained this word was: 'to put an end to the {state | position} of war.' The ambiguous word is involved in two syntactic relations, being a complement of 'put' and also modified by 'war.' The corresponding frequencies were:

(15)	verb-comp:	put-position	320
	verb-comp:	put-state	18
	noun-nobj:	state-war	13
	noun-nobj:	position-war	2

The bound of the odds ratio (B_α) for the first relation was higher than for the second, and, therefore, this relation determined the translation as 'position.' However, the correct translation should be 'state,' as determined by the second relation.

These data suggest that while ordering the relations (or using any other weighting mechanism) it may be necessary to give different weights to the different types of syntactic relations. For instance, it seems reasonable that the object of a noun should receive greater weight in selecting the noun's sense than the verb for which this noun serves as a complement.

Further examination of the example suggests another refinement of our method: it turns out that most of the 320 instances of the tuple *(verb-comp: put position)* include the preposition "in", as part of the common phrase "put in a position". Therefore, these instances should not be considered for the current example, which includes the preposition "to". However, the distinction between different prepositions was lost by our program, as a result of using equivalence classes of syntactic tuples (see Section 2.3). This suggests that we should not use an equivalence class when there is enough statistical data for specific tuples.[12]

Confusing senses. In another example, the Hebrew adjective *qatann* modified the noun *sikkuy*, which means 'prospect' or 'chance'. The word *qatann* has several translations, two of which are 'small' and 'young'. In this Hebrew word combination, the correct sense of *qatann* is necessarily 'small'. However, the relation that was observed in the corpus was 'young prospect', relating to the human sense of 'prospect' which appeared in sport articles (a promising young person). This borrowed sense of 'prospect' is necessarily inappropriate, since in Hebrew it is represented by the equivalent of 'hope' (*tiqwa*), and not by *sikkuy*.

The source of this problem is Assumption : a target tuple T might be a translation of several source tuples and while gathering statistics for T we cannot distinguish between the different sources, since we use only a target language corpus.

12 We thank the anonymous reviewer for suggesting this point.

A possible solution is to use an aligned bilingual corpus, as suggested by Sadler (1989), Brown et al. (1991) and Gale et al. (1992a). In such a corpus the occurrences of the relation 'young prospect' will be aligned with the corresponding occurrences of the Hebrew word 'tiqwa', and will not be used when the Hebrew word *sikkuy* is involved. Yet, it should be brought to mind that an aligned corpus is the result of manual translation, which can be viewed as including a manual tagging of the ambiguous words with their equivalent senses in the target language. This resource is much more expensive and less available than an untagged monolingual corpus, while it seems to be necessary only for relatively rare situations. Therefore, considering the tradeoff between applicability and precision, it seems better to rely on a significantly larger monolingual corpus than on a smaller bilingual corpus. An optimal method might exploit both types of corpora, where the somewhat more accurate, but more expensive, data of a bilingual corpus is augmented by the data of a much larger monolingual corpus.[13]

Lack of deep understanding. By their nature, statistical methods rely on large quantities of shallow information. Thus, they are doomed to fail when disambiguation can rely only on deep understanding of the text and no other surface cues are available. This happened in one of the Hebrew examples, where the two alternatives were either 'emigration law' or 'immigration law' (the Hebrew word *hagira* is used for both subsenses). While the context indicated that the first alternative is correct (emigration from the Soviet Union), the statistics (which were extracted from texts related to North America) preferred the second alternative. In order to translate the above phrase the program would need deep knowledge, to an extent which seems to far exceed the capabilities of current systems. Fortunately, our results suggest that such cases are quite rare.

6.4 Monolingual applications

The results of our experiments in the context of machine translation suggest the utility of a similar mechanism even for word sense disambiguation within a single language. In order to select the right sense of a word, in a broad coverage application, it is useful to identify lexical relations between word senses. However, within corpora of a single language it is possible to identify automatically only relations at the word level, which are, of course, not useful for selecting word senses in that language. This is where other languages can supply the solution, exploiting the fact that the mapping between words and word senses varies significantly between different languages. For instance, the English words 'sign' and 'seal' (from example (1) in the introduction) correspond to two distinct senses of the Hebrew word *la<u>h</u>tom* . These senses should be distinguished by most applications of Hebrew understanding programs. To make this distinction, it is possible to perform the same process that is performed for target word selection, by producing all the English alternatives for the lexical relations involving *la<u>h</u>tom*. Then the Hebrew sense which corresponds to the most plausible English lexical relations is preferred. This process requires a bilingual lexicon which maps each Hebrew sense separately into its possible translations, similar to a Hebrew-Hebrew-English lexicon (analogous to the Oxford English-English-Hebrew dictionary of Hornby et al. (1986), which lists the senses of an English word, along with the possible Hebrew translations for each of them).

13 Even though there are large quantities of translated texts, experience has shown that it is much harder to obtain large bilingual corpora than large monolingual corpora. As mentioned earlier, a bilingual corpus of moderate size can be valuable when constructing a bilingual lexicon, thus justifying the effort of maintaining such a corpus.

In some cases, different senses of a Hebrew word map to the same word in English. In these cases, the lexical relations of each sense cannot be identified in an English corpus, and a third language is required to distinguish among these senses. Alternatively, it is possible to combine our method with other disambiguation methods, that have been developed in a monolingual context (see the next section). As a long term vision, one can imagine a multilingual corpora based environment, which exploits the differences between languages to facilitate the acquisition of knowledge about word senses.

7. Comparative Analysis of Statistical Sense Disambiguation Methods

Until recently, word sense disambiguation seemed to be a problem for which there was no satisfactory solution for broad coverage applications. Recently, several statistical methods have been developed for solving this problem, suggesting the possibility of robust, yet feasible, disambiguation. In this section we identify and analyze basic aspects of a statistical sense disambiguation method, and compare several proposed methods (including ours) along these aspects.[14] This analysis may be useful for future research on sense disambiguation, as well as for the development of sense disambiguation modules in practical systems. The basic aspects which will be reviewed are:

1. Information sources used by the disambiguation method.
2. Acquisition of the required information from training texts.
3. The computational decision model.
4. Performance evaluation.

The first three aspects deal with the components of a disambiguation method, as would be implemented for a practical application. The fourth is a methodological issue, which is relevant for developing, testing and comparing disambiguation methods.

7.1 Information sources

We identify three major types of information which were used in statistical methods for sense disambiguation:

1. Words appearing in the local, syntactically related, context of the ambiguous word.
2. Words appearing in the global context of the ambiguous word.
3. Probabilistic syntactic and morphological characteristics of the ambiguous word.

14 The reader is referred to some of these recent papers for thorough surveys of work on sense disambiguation (Hearst, 1991; Gale, Church, and Yarowsky, 1992a; Yarowsky, 1992).

The first type of information is the one used in the current paper, where words which are syntactically related to an ambiguous word are used to indicate its most probable sense. Statistical data on the cooccurrence of syntactically related words with each of the alternative senses reflect semantic and lexical preferences and constraints of these senses. In addition, these statistics may provide information about the topics of discourse which are typical for each sense.

Ideally, the syntactic relationships between words should be identified using a syntactic parser, in both the training and the disambiguation phases. Since robust syntactic parsers are not widely available, and those that exist are not always accurate, it is possible to use various approximations in order to identify relevant syntactic relationships between words. Hearst (1991) uses a stochastic part of speech tagger and a simple scheme for partial parsing of short phrases. The structures achieved by this analysis are used to identify approximated syntactic relationships between words. Brown et al. (1991) make even weaker approximations, using only a stochastic part of speech tagger, and defining relationships such as "the first verb to the right" or "the first noun to the left". Finally, Dagan et al. (1993) (see Section 5.1) assume full parsing at the disambiguation phase, but no preprocessing at the training phase, in which a higher level of noise can be accommodated.

A second type of information is provided by words which occur in the global context of the ambiguous word (Gale, Church, and Yarowsky, 1993; Gale, Church, and Yarowsky, 1992b; Yarowsky, 1992; Schütze, 1992). Gale et al. and Yarowsky use words that appear within 50 words in each direction from the ambiguous word.[15] Statistical data is stored about the occurrence of words in the context of each sense, and is matched against the context in the disambiguated sentence. Cooccurrence in the global context provides information about typical topics which are associated with each sense, where a topic is represented by words which commonly occur in it. Schutze (1992; 1993) uses a variant of this type of information, where context vectors are maintained for character four-grams, instead of words. In addition, the context of an occurrence of an ambiguous word is represented by cooccurrence information of a second order, as a set of context vectors (instead of a set of context words).

Compared with cooccurrence within syntactic relations, information about the global context is less sensitive to fine semantic and lexical distinctions, and is less useful when different senses of a word appear in similar contexts. On the other hand, the global context contains more words, and is therefore more likely to provide enough disambiguating information, in cases where this distinction can be based on the topic of discourse. From a general perspective, these two types of information represent a common tradeoff in statistical language processing: the first type is related to a limited amount of deeper, and more precise, linguistic information, while the second type provides a large amount of shallow information, which can be applied in a more robust manner. The two sources of information seem to complement each other, and may both be combined in future disambiguation methods.[16]

---

15 The size of the context was determined experimentally, based on evaluations of different sizes of context. This optimization was performed for the Hansard corpus of the proceedings of the Canadian Parliament. In general, the size of the global context depends on the corpus, and typically consists of a homogeneous unit of discourse.

16 See also (Gale, Church, and Yarowsky, 1992b, pp. 58-59) and (Schütze, 1992; Schütze, 1993) for methods of reducing the number of parameters when using global contexts, and (Dagan, Marcus, and Markovitch, 1993) for increasing the applicability of the use of local context, in cases where

Hearst (1991) incorporates a third type of statistical information to distinguish between different senses of nouns (in addition to the first type discussed above). For each occurrence of a sense, several syntactic and morphological characteristics are recorded, such as whether the noun modifies or is modified by another word, whether it is capitalized and whether it is related to certain prepositional phrases. Then, in the disambiguation phase, a best match is sought between the information recorded for each sense and the syntactic context of the current occurrence of the noun. This type of information resembles the information which is defined for lexical items in lexicalist approaches for grammars, such as possible subcategorization frames of a word. The major difference is that Hearst captures probabilistic preferences of senses for such syntactic constructs. Grammatical formalisms, on the other hand, usually specify only which constructs are possible, and at most distinguish between optional and obligatory ones. Therefore, the information recorded in such grammars cannot distinguish between different senses of a word which potentially have the same subcategorization frames, though in practice each sense might have different probabilistic preferences for different syntactic constructs.

It is clear that each of the different types of information provides some information which is not captured by the others. However, as the acquisition and manipulation of each type of information requires different tools and resources, it is important to assess the relative contribution, and the "cost effectiveness", of each of them. Such comparative evaluations are not available yet, not even for systems that incorporate several types of data (e.g. (McRoy, 1992)). Further research is therefore needed to compare the relative importance of different information types, and to find optimal ways of combining them.

7.2 Acquisition of training information

When training a statistical model for sense disambiguation, it is necessary to associate the acquired statistics with word *senses*. This seems to require manual tagging of the training corpus with the appropriate sense for each occurrence of an ambiguous word. A similar approach is being used for stochastic part of speech taggers and probabilistic parsers, relying on the availability of large, manually tagged (or parsed), corpora for training. However, this approach is less feasible for sense disambiguation, for two reasons. First, the size of corpora required to acquire sufficient statistics on lexical cooccurrence is usually much larger than that used for acquiring statistics on syntactic constructs or sequences of parts of speech. Second, lexical cooccurrence patterns, as well as the definition of senses, may vary a great deal across different domains of discourse. Consequently, it is usually not sufficient to acquire the statistics from one widely available "balanced" corpus, as common for syntactic applications. A sense disambiguation model should be trained on the same type of texts for which it will be applied, thus increasing the cost of manual tagging.

The need to disambiguate a training corpus before acquiring a statistical model for disambiguation is often termed as the *circularity* problem. In the following paragraphs we discuss different methods that have been proposed to overcome the circularity problem, without exhaustive manual tagging of the training corpus. In our opinion, this is the most critical issue in developing feasible sense disambiguation methods.

---

there is no direct statistical evidence.

Bootstrapping. Bootstrapping, which is a general scheme for reducing the amount of manual tagging, was proposed also for sense disambiguation (Hearst, 1991). The idea is to manually tag an initial set of occurrences for each sense in the lexicon, acquiring initial training statistics from these instances. Then, using these statistics, the system tries to disambiguate additional occurrences of ambiguous words. If such an occurrence can be disambiguated automatically with high confidence, the system acquires additional statistics from this occurrence, as if it were tagged by hand. Hopefully, the system will incrementally acquire all the relevant statistics, demanding just a small amount of manual tagging. The results in (Hearst, 1991) show that at least 10 occurrences of each sense have to be tagged by hand, and in most cases 20-30 occurrences are required to get high precision. These results, which were achieved for a small set of pre-selected ambiguous words, suggest that the cost of the bootstrapping approach is still very high.

Clustering occurrences of an ambiguous word. Schutze (1992; 1993) proposes a method which can be viewed as an efficient way of manual tagging. Instead of presenting all occurrences of an ambiguous word to a human, these occurrences are first clustered using automatic clustering algorithms.[17] Then a human is asked to assign one of the senses of the word to each cluster, by observing several members of the cluster. Each sense is thus represented by one or more clusters. At the disambiguation phase, a new occurrence of an ambiguous word is matched against the contexts that were recorded for these clusters, selecting the sense of that cluster which provides the best match.

It is interesting to note that the number of occurrences that had to be observed by a human in the experiments of Schutze is of the same order as in the bootstrapping approach: 10-20 members of a cluster were observed, with an average of 2.8 clusters per sense. As both approaches were tested only on a small number of pre-selected words, further evaluation is necessary to predict the actual cost of their application to broad domains. The methods described below, on the other hand, rely on resources which were already available on a large scale, and it is therefore possible to estimate the expected cost of their broad application.

Word classification. Yarowsky (1992) proposes a method which completely avoids manual tagging of the training corpus. This is achieved by estimating parameters for classes of words rather than for individual word senses. In his work, Yarowsky considered the semantic categories defined in Roget's thesaurus as classes. He then mapped (manually) each of the senses of an ambiguous word to one or several of the categories under which this word is listed in the thesaurus. The task of sense disambiguation, thus, becomes the task of selecting the appropriate category for each occurrence of an ambiguous word.[18]

When estimating the parameters of a category,[19] any occurrence of a word that belongs to that category is counted as an occurrence of the category. This means that each occurrence of an ambiguous word is counted as an occurrence of all the categories to which the word belongs, and not just the category which corresponds to the specific occurrence. A substantial amount of noise is introduced by this training method, which

17 Each occurrence is represented as a context vector, and the vectors are then clustered.
18 In some cases, the Roget index was found to be incomplete, and a missing category had to be added to the list of possibilities for a word.
19 Yarowsky uses statistics on occurrences of specific words in the global context of the category, but the method can be used to collect other types of statistics, such as the cooccurrence of the category with other categories.

is a consequence of the circularity problem: To avoid the noise, it would be necessary to tag each occurrence of an ambiguous word with the appropriate category. As explained by Yarowsky, however, this noise can usually be tolerated. The "correct" parameters of a certain class are acquired from all its occurrences, while the "incorrect" parameters are distributed through occurrences of many different classes, and usually do not produce statistically significant patterns. To further reduce the noise, Yarowsky uses a system of weights which assigns lower weights to frequent words, since such words may introduce more noise.[20] The word class method thus overcomes the circularity problem by mapping word senses to classes of words. However, because of this mapping the method cannot distinguish between senses which belong to the same class, and also introduces some level of noise.

Bilingual corpus. Brown et al. (1991) were concerned with sense disambiguation for machine translation. Having a large aligned bilingual corpus available, they noticed that the target word which corresponds to an occurrence of an ambiguous source word can serve as a tag of the appropriate sense. This kind of tagging provides sense distinctions when different senses of a source word translate to different target words. For the purpose of translation, these are exactly the cases for which sense distinction is required. Conceptually, the use of a bilingual corpus does not eliminate (or reduce) manual tagging of the training corpus. Such a corpus is a result of manual translation, and it is the translator who provides tagging of senses as a side effect of the translation process. Practically, whenever a bilingual corpus is available, it provides a useful source of a sense tagged corpus. Gale et al. (1992a) have also exploited this resource for achieving large amounts of testing and training materials.

Bilingual lexicon and a monolingual corpus. The method of the current paper also exploits the fact that different senses of a word are usually mapped to different words in another language. However, our work shows that the differences between languages enable us to avoid any form of manual tagging of the corpus (including translation). This is achieved by a bilingual lexicon which maps a source language word to all its possible equivalents in the target language. This approach has practical advantages for the purpose of machine translation, where a bilingual lexicon needs to be constructed in any case, while very large bilingual corpora are not usually available. From a theoretical point of view, the difference between the two methods can be made clear if we assume that the bilingual lexicon contains exactly all the different translations of a word which occur in a bilingual corpus. For a given set of senses that need to be disambiguated, our method requires a bilingual corpus of size k, in which each sense occurs at least once, in order to establish its mapping to a target word. In addition, a larger *monolingual* corpus, of size n, is required, to provide enough training examples of typical contexts for each sense. On the other hand, using a bilingual corpus for training the disambiguation model would require a bilingual corpus of size n, which is significantly larger than k. The savings in resources is achieved since the mapping between the languages is done at the level of *single* words. The larger amount of information about word *combinations*, on the other hand, is acquired from an untagged monolingual corpus, after the mapping has been performed. Our results show that the precision of the selection algorithm is high

20 The method of acquiring parameters from ambiguous occurrences in a corpus, relying on the "spreading" of noise, can be used in many contexts. For example, it was used for acquiring statistics for disambiguating prepositional phrase attachments, counting ambiguous occurrences of prepositional phrases as representing both noun-PP and verb-PP constructs (Hindle and Rooth, 1991).

despite the additional noise which is introduced by mapping single words independently of their context. As mentioned in section 6.3, an optimal method may combine the two methods.

In some sense, the use of a bilingual lexicon resembles the use of a thesaurus in Yarowsky's approach. Both rely on a manually established mapping of senses to other concepts (classes of words or words in another language), and collect information about the target concepts from an untagged corpus. In both cases, ambiguous words in the corpus introduce some level of noise: counting an occurrence of a word as an occurrence of all the classes to which it belongs, or counting an occurrence of a target word as an occurrence of all the source words to which it may correspond (a smaller amount of noise is introduced in the latter case, as a mapping to target words is much more fine grained than a mapping to Roget's categories). Also, both methods can distinguish only between senses which are distinguished by the mappings they use: either senses that belong to different classes, or senses that correspond to different target words. An interesting difference, though, relates to the feasibility of implementing the two methods for a new domain of texts (in particular technical domains). The construction of a bilingual lexicon for a new domain is relatively straightforward, and is often carried out for translation purposes. The construction of an appropriate classification for the words of a new domain is more complex, and, furthermore, it is not clear that in every domain it is possible to construct a classification which is sufficient for the purpose of sense disambiguation.

7.3 The Computational Decision Model

Sense disambiguation methods require a decision model which evaluates the relevant statistics. Sense disambiguation thus resembles many other decision tasks, and, not surprisingly, several common decision algorithms were employed in different works. These include a Bayesian classifier (Gale, Church, and Yarowsky, 1993) and a distance metric between vectors (Schütze, 1993), both inspired by methods in information retrieval; the use of the flip-flop algorithm for ordering possible informants about the preferred sense, trying to maximize the mutual information between the informant and the ambiguous word (Brown et al., 1991); and the use of confidence intervals to establish the degree of confidence in a certain preference, combined with a constraint propagation algorithm (the current paper). At the current stage of research on sense disambiguation, it is difficult to judge whether a certain decision algorithm is significantly superior than others.[21] Yet, these decision models can be characterized by several criteria, which clarify the similarities and differences between them. As will be explained below, many of the differences are correlated with the different information sources employed by these models.

- Combining several informants.
  The methods in (Brown et al., 1991) and in the current paper combine several informants (i.e., statistics about several context words) by choosing the informant which seems most indicative for the selection. The effect of other, less significant, informants is then discarded. The Bayesian classifier and the vector distance metric combine all informants simultaneously, in a

21 Once the important information sources for sense selection have been identified, it is possible that different decision algorithms would achieve comparable results.

multiplicative or additive manner, possibly assigning a certain weight to each informant.

- Reducing the number of parameters.
  Since sense disambiguation relies on statistics about lexical cooccurrence, the number of relevant parameters is very high, especially when cooccurrence in the global context is considered. For this reason, Schutze uses two compaction methods: First, 5000 "informative" four-grams are used instead of words. Second, the 5000 dimensions are decomposed to 97 dimensions, using singular value decomposition. This method reduces the number of parameters significantly, but has the disadvantage that it is neither possible to trace the meaning of the entries in the resulting vectors nor to associate them directly with the original cooccurrence statistics. Gale et al. (1992b, pp. 58-59) propose another approach, and reduce the number of parameters by selecting the most informative context words for each sense. The selection of context words is based on a theoretically motivated criterion, borrowed from Mosteller and Wallace (1964, pp. 55-56). Finally, Yarowsky's method further reduces the number of parameters, as it records cooccurrences between individual words and word classes.

- Statistical significance of the selection.
  In the current paper, we use confidence intervals to test whether the statistical preference for a certain sense is significant. In a simple multiplicative preference score, on the other hand, it is not possible to distinguish whether preferences rely on small or large counts. The method of Gale et al. remedies this problem indirectly (in most cases) by introducing a sophisticated interpolation between the actual counts of the cooccurrence parameters and the frequency counts of the individual words (see (Gale, Church, and Yarowsky, 1993) for details). In Schutze's method it is not possible to trace the statistical significance of the parameters since they are the result of extensive processing and compaction of the original statistical data.

- Resolving all ambiguities simultaneously.
  In the current paper, the selection of a sense for one word affects the selection for another word through a constraint propagation algorithm. This property is absent in most other methods.

The differences between various disambiguation methods correlate with the difference in information sources, in particular whether they use local or global context. When local context is used, only few syntactically related informants may provide reliable information about the selection. It is therefore reasonable to base the selection on only one, the most informative informant, and it is also important to test the statistical significance of that informant. The problem of parameter explosion is less severe, and the number of parameters is comparable to that of a bigram language model (and even smaller). When using global context, on the other hand, the number of potential parameters is significantly larger, but each of them is usually less informative. It is therefore important to take into account as many parameters as possible in each ambiguous case, but it is less important to test for detailed statistical significance, or to worry about the mutual effects of sense selections for adjacent words.

7.4 Performance Evaluation

In most of the abovementioned papers, experimental results are reported for a small set of up to 12 pre-selected words, usually with 2 or 3 senses per word. In the current paper we have evaluated our method using a random set of example sentences, with no a priori selection of the words. This standard evaluation method, which is commonly used for other natural language processing tasks, provides a direct prediction for the expected success rate of the method when employed in a practical application.

In order to compare results on different test data, it is useful to compare the precision of the disambiguation method with some a priori figure that reflects the degree of ambiguity in the text. Reporting the number of senses per example word corresponds to the expected success rate of random selection. A more informative figure is the success rate of a naive method that always selects the most frequent sense (the Word Frequencies method in our evaluations). The success rate of this naive method is higher than that of random selection, and thus provides a tighter lower bound for the desired precision of a proposed disambiguation method.

An important practical issue in evaluation is how to get the test examples, which should be tagged with the correct sense. In most papers (including ours) the tagging of the test data was done by hand, which limits the size of the test set. Preparing one test set by hand may still be reasonable, though time consuming. However, it is useful to have more than one set, such that results will be reported on a new, unseen, set, while another set is used for developing and tuning the system. One useful source of tagged examples is an aligned bilingual corpus, which can be used for testing any sense disambiguation method, including methods that do not use bilingual material for training. Gale proposes to use "pseudo-words" as another practical source of testing examples (Gale, Church, and Yarowsky, 1992b) (equivalently, Schutze (Schütze, 1992) uses "artificial ambiguous words"). Pseudo-words are constructed artificially as a union of several different words (say, w_1, w_2 and w_3 define three "senses" of the pseudo-word x). The disambiguation method is presented with texts where all occurrences of w_1, w_2 and w_3 are considered as occurrences of x, and should then select the original word ("sense") for each occurrence. Though testing with this method does not provide results for real ambiguities that occur in the text, it can be very useful while developing and tuning the method (Gale shows high correlation between the performance of his method on real sense ambiguities and pseudo-words).

8. Conclusions

The method presented in this paper takes advantage of two linguistic phenomena, both proven to be very useful for sense disambiguation: (a) the different mapping between words and word senses among different languages and (b) the importance of lexical cooccurrence within syntactic relations. The first phenomenon provides the solution for the circularity problem in acquiring sense disambiguation data. Using a bilingual lexicon and a monolingual corpus of the target language we can acquire statistics on word senses automatically, without manual tagging. As explained in Section 7, this method has significant practical and theoretical advantages over the use of aligned bilingual corpora. We pay for these advantages by introducing an additional level of noise, in mapping individual words independently to the other language. Our results show, however, that the precision of the selection algorithm is high despite this additional noise.

This work also emphasizes the importance of lexical cooccurrence within syntactic relations for the resolution of lexical ambiguity. Cooccurrences found in a large corpus reflect a huge amount of semantic knowledge, which was traditionally constructed by hand. Moreover, frequency data for such cooccurrences reflect both linguistic and domain specific preferences, thus indicating not only what is *possible*, but also what is *probable*. It is important to notice that frequency information on lexical cooccurrence was found much more predictive than single word frequency. In the three experiments we reported, there were 61 cases in which the two types of information contradicted each other, favoring different target words. In 56 of these cases (92%), it was the most frequent lexical cooccurrence, and not the most frequent word, that predicted the correct translation. This result may raise relevant hypotheses for psycholinguistic research, which has indicated the relevance of word frequencies to human sense disambiguation (e.g. (Simpson and Burgess, 1988)).

We suggest that the high precision achieved in the experiments relies on two characteristics of the ambiguity phenomena, namely the *sparseness* and *redundancy* of the disambiguating data. By sparseness we mean that within the large space of alternative interpretations produced by ambiguous utterances, only a small portion is commonly used. Therefore, the chance that an inappropriate interpretation is observed in the corpus (in other contexts) is low. Redundancy relates to the fact that different informants (such as different lexical relations or deep understanding) tend to support rather than contradict one another, and therefore the chance of picking a "wrong" informant is low.

It is interesting to compare our method with some aspects of the statistical machine translation system of Brown et al. (1990). As mentioned in the introduction, this system also incorporates target language statistics in the translation process. In order to translate a French sentence, f, they choose the English sentence, e, that maximizes the term $\Pr(e) \cdot \Pr(f|e)$. The first factor in this product, which represents the target language model, may thus affect any aspect of the translation, including target word selection.

It seems, however, that Brown et al. expect that target word selection would be mainly determined by translation probabilities (the second factor in the above term), which should be derived from a bilingual corpus (Brown et al., 1990, pp. 79). This view is reflected also in their elaborate method for target word selection (Brown et al., 1991), where better estimates of translation probabilities are achieved as a result of word sense disambiguation. Our method, on the other hand, incorporates only target language probabilities, and ignores any notion of translation probabilities. It thus demonstrates a possible tradeoff between these two types of probabilities: using more informative statistics of the target language may compensate for the lack of translation probabilities. For our system, the more informative statistics are achieved by syntactic analysis of both the source and target languages, instead of the simple tri-gram model used by Brown et al.. In a broader sense, this can be viewed as a tradeoff between the different components of a translation system: having better analysis and generation models may reduce some burden from the transfer model.

In our opinion, the method proposed in this paper may have immediate practical value, beyond its theoretical aspects. As we argue below, we believe that the method is feasible for practical machine translation systems, and can provide a cost effective improvement on target word selection methods. The identification of syntactic relations in the source sentence is available in any MT system which uses some form of syntactic parsing. Trivially, a bilingual lexicon is available. A parser for the target language becomes common in many systems which offer bidirectional translation capabilities, requiring parsers for several languages (see (Miller, 1993) for available language pairs in several commercial MT systems). If a parser for the target language corpus is not available, it is possible to approximate the statistics using word cooccurrence in a window, as was demonstrated by a variant of our method (Dagan, Marcus, and Markovitch, 1993) (see Section 5.1). In both cases, the statistical model was shown to successfully handle the noise produced in automatic acquisition of the data. Substantial effort may be required for collecting a sufficiently large target language corpus. We have not studied the relationship between the corpus size and the performance of the algorithm, but it is our impression that a corpus of several hundred thousand words will prove useful for translation in a well defined domain. With current availability of texts in electronic form,[22] a corpus of this size is feasible in many domains. The effort of assembling this corpus should be compared to the effort of manual coding of sense disambiguation information. Finally, our method was evaluated by simulating realistic MT lexicons, on randomly selected examples, and yielded high performance in two different broad domains (foreign news articles and a software manual). It is therefore expected that the results reported here will be reproduced in other domains and systems.

To further improve the performance of target word selection, our method may be combined with other sense disambiguation methods. As discussed in Section 6.2, it is possible to increase the applicability (coverage) of the selection method by considering word cooccurrence in a limited context and/or by using similarity-based methods which reduce the problem of data sparseness. To a lesser extent, the use of a bilingual corpus may further increase the precision of the selection (see Section 6.3). A practical strategy may be to use a bilingual corpus for enriching the bilingual lexicon, while relying mainly on cooccurrence statistics from a larger monolingual corpus for disambiguation.

In a broader context, this paper promotes the combination of statistical and linguistic models in natural language processing. It provides an example of how a problem can be first defined in detailed linguistic terms, using an implemented linguistic tool (a syntactic parser, in our case). Then, having a well defined linguistic scenario, we apply a suitable statistical model to highly informative linguistic structures. According to this view, a complex task, such as machine translation, should be first decomposed on a linguistic basis. Then, appropriate statistical models can be developed for each sub-problem. We believe that this approach provides a beneficial compromise between two extremes in natural language processing: either using linguistic models which ignore quantitative information, or using statistical models which are linguistically ignorant.

Acknowledgements

Special thanks are due to Ulrike Schwall for her fruitful collaboration. We are grateful to Mori Rimon, Peter Brown, Ayala Cohen, Ulrike Rackow, Herb Leass and Bill Gale for their help and comments. We also thank the anonymous reviewers for their detailed

22 Optical character recognition (OCR) can also be used to acquire relevant texts in electronic form. In this case, it may be necessary to approximate the statistics using word cooccurrence in a window, since parsing noisy OCR output is difficult.

comments, which resulted in additional discussions and clarifications. This research was partially supported by grant number 120-741 of the Israel Council for Research and Development.

A. Approximating $var\left[\ln\left(\frac{\hat{p_1}}{\hat{p_2}}\right)\right]$

In order to approximate $var\left[\ln\left(\frac{\hat{p_1}}{\hat{p_2}}\right)\right]$ we first approximate $\ln\left(\frac{\hat{p_1}}{\hat{p_2}}\right)$ by the first order derivatives (the first term of the Taylor series):

$$\begin{aligned}
\ln\left(\frac{\hat{p_1}}{\hat{p_2}}\right) &\approx \ln\left(\frac{p_1}{p_2}\right) + (\hat{p_1} - p_1)\left[\frac{\partial}{\partial x_1}\ln\left(\frac{x_1}{x_2}\right)\right]_{p_1,p_2} \\
&\quad + (\hat{p_2} - p_2)\left[\frac{\partial}{\partial x_2}\ln\left(\frac{x_1}{x_2}\right)\right]_{p_1,p_2} \\
&= \ln\left(\frac{p_1}{p_2}\right) + \frac{\hat{p_1} - p_1}{p_1} - \frac{\hat{p_2} - p_2}{p_2} \\
&= \ln\left(\frac{p_1}{p_2}\right) + \frac{\hat{p_1}}{p_1} - \frac{\hat{p_2}}{p_2}. \qquad (5)
\end{aligned}$$

We use the following equations (see (Agresti, 1990)):

$$\begin{aligned}
var(x+c) &= var(x), \\
var(x_1 - x_2) &= var(x_1) + var(x_2) - 2 \cdot covariance(x_1, x_2), \\
var(\hat{p}) &= \frac{p(1-p)}{n}, \\
var\left(\frac{x}{c}\right) &= \frac{var(x)}{c^2}, \\
covariance(\hat{p_i}, \hat{p_j}) &= -\frac{p_i p_j}{n}, \\
covariance(\frac{x_1}{c_1}, \frac{x_2}{c_2}) &= \frac{covariance(x_1, x_2)}{c_1 c_2}.
\end{aligned}$$

Using (5) we get

$$\begin{aligned}
var\left[\ln\left(\frac{\hat{p_1}}{\hat{p_2}}\right)\right] &\approx var\left[\ln\left(\frac{p_1}{p_2}\right) + \frac{\hat{p_1}}{p_1} - \frac{\hat{p_2}}{p_2}\right] \\
&= var\left[\frac{\hat{p_1}}{p_1} - \frac{\hat{p_2}}{p_2}\right] \\
&= var\left[\frac{\hat{p_1}}{p_1}\right] + var\left[\frac{\hat{p_2}}{p_2}\right] - 2 \cdot covariance\left[\frac{\hat{p_1}}{p_1}, \frac{\hat{p_2}}{p_2}\right] \\
&= \frac{1}{p_1^2}\frac{p_1(1-p_1)}{n} + \frac{1}{p_2^2}\frac{p_2(1-p_2)}{n} + 2\frac{p_1 p_2}{n p_1 p_2} \\
&= \frac{1}{np_1} + \frac{1}{np_2} \approx \frac{1}{n\hat{p_1}} + \frac{1}{n\hat{p_2}} = \frac{1}{n_1} + \frac{1}{n_2}.
\end{aligned}$$

References

Agresti, Alan. 1990. *Categorical Data Analysis.* John Wiley & Sons.

Alcalay, R. 1990. *The Complete Hebrew-English Dictionary.* Massada.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to language translation. *Computational Linguistics*, 16(2):79–85.

Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer. 1991. Word sense disambiguation using statistical methods. In *Proc. of the Annual Meeting of the ACL*, pages 264–270.

Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. But dictionaries are data too. In *Proc. of the ARPA Workshop on Human Language Technology*, pages 202–205.

Chodorow, M. S., R. J. Byrd, and G. E. Heidron. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proc. of the Annual Meeting of the ACL*, pages 299–304.

Church, Kenneth W. and William A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Dagan, Ido, Kenneth Church, and William Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8.

Dagan, Ido and Alon Itai. 1990. Automatic acquisition of constraints for the resolution of anaphora references and syntactic ambiguities. In *Proc. of COLING*, volume 3, pages 330–332.

Dagan, Ido and Alon Itai. 1991. A statistical filter for resolving pronoun references. In Y. A. Feldman and A. Bruckstein, editors, *Artificial Intelligence and Computer Vision*. Elsevier Science Publishers B.V., pages 125–135. (The Proc. of the 7th Israeli Sym. on Artificial Intelligence and Computer Vision, 1990).

Dagan, Ido, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proc. of the Annual Meeting of the ACL*, pages 130–137.

Dagan, Ido, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proc. of the Annual Meeting of the ACL*, pages 164–171.

Gale, William, Kenneth Church, and David Yarowsky. 1992a. Using bilingual materials to develop word sense disambiguation methods. In *Proc. of the International Conference on Theoretical and Methodolgical Issues in Machine Translation*, pages 101–112.

Gale, William, Kenneth Church, and David Yarowsky. 1992b. Work on statistcal methods for word sense disambiguation. In *Working Notes, AAAI Fall Symposium Series, Probabilistic Approaches to Natural Language*, pages 54–60.

Gale, William, Kenneth Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Golan, Igal, Shalom Lappin, and Mori Rimon. 1988. An active bilingual lexicon for machine translation. In *Proc. of COLING*, pages 205–211.

Grishman, R., L. Hirschman, and Ngo Thanh Nhan. 1986. Discovery procedures for sublanguage selectional patterns – initial experiments. *Computational Linguistics*, 12:205–214.

Hearst, Marti. 1991. Noun homograph disambiguation using local context in large text corpora. In *Proc. of the Annual Conference of the UW Center for the New OED and Text Research*, pages 1–22.

Hindle, D. 1990. Noun classification from predicate-argument structures. In *Proc. of the Annual Meeting of the ACL*, pages 268–275.

Hindle, D. and M. Rooth. 1991. Structural ambiguity and lexical relations. In *Proc. of the Annual Meeting of the ACL*, pages 229–236.

Hornby, A. S., C. Ruse, J. A. Reif, and Y. Levy. 1986. *Oxford Student's Dictionary for Hebrew Speakers*. Kernerman Publishing Ltd., Lonnie Kahn & Co. Ltd.

Jelinek, Frederick. 1990. Self-organized language modeling for speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*. Morgan Kaufmann Publishers, Inc., San Mateo, California, pages 450–506.

Jelinek, Frederick, Robert L. Mercer, and Salim Roukos. 1992. Principles of lexiacal language modeling for speech recognition. In Sadaoki Furui and M. Mohan Sondhi, editors, *Advances in Speech Signal Processing*. Mercer Dekker, Inc., pages 651–699.

Katz, Slava M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, speech, and Signal Processing*, 35(3):400–401.

Lenat, D. B., R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. 1990. Cyc: Toward programs with common sense. *Communications of the ACM*, 33(8):30–49.

McCord, M. C. 1990. Slot grammar: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural Language and Logic: International Scientific Symposium*, Lecture Notes in Computer Science. Springer Verlag, Berlin, pages 118–145.

McCord, M. C. 1991. The slot grammar system. Technical Report RC 17313, IBM Research Report. To appear in J. Wedekind and C. Rohrer (Eds.), *Unification in Grammar*, MIT Press.

McRoy, Susan W. 1992. Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1–30.

Miller, L. Chris. 1993. Bableware for the desktop. *BYTE*, January:177–183.

Mosteller, Frederick and David Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Nirenburg, S., editor. 1987. *Machine Translation*. Cambridge University Press.

Nirenburg, S., I. Monarch, T. Kaufmann, I. Nirenburg, and J. Carbonell. 1988. Acquisition of very large knowledge bases: Methodology, tools and applications. Technical Report CMU-CMT-88-108, Center for Machine Translation, Carnegie-Mellon.

Sadler, V. 1989. *Working with analogical semantics: Disambiguation techniques in DLT*. Foris Publications.

Schütze, Hinrich. 1992. Dimensions of meaning. In *Proc. of Supercomputing*, pages 787–796.

Schütze, Hinrich. 1993. Word space. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*. Morgan Kaufman Publishers, San Mateo CA, pages 895–902.

Simpson, Greg B. and Curt Burgess. 1988. Implications of lexical ambiguity resolution for word recognition. In G. W. Cotrell Small, S. L. and M. K. Tanenhaus, editors, *Lexical Ambiguity Resolution*. Morgan Kaufman Publishers, pages 271–288.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Yarowsky, David. 1992. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proc. of COLING*, pages 454–460.

Zernik, U. and P. Jacobs. 1990. Tagging for learning: Collecting thematic relations from corpus. In *Proc. of COLING*, volume 1, pages 34–39.