Morphological Disambiguation in Hebrew Using A Priori Probabilities

by

M. Levinger, U. Ornan and A. Itai

Technical Report #760
December, 1992
Morphological Disambiguation in Hebrew Using A Priori Probabilities*

Moshe Levinger
IBM Scientific Center, Haifa, Israel

Uzzi Ornan
Computer Science Department
Technion, Haifa, Israel

Alon Itai
Computer Science Department
Technion, Haifa, Israel

November 15, 1992

Abstract

This paper describes a new approach for morphological disambiguation in Hebrew using an untagged corpus. This approach demonstrates a way to extract very useful and non-trivial information from an untagged corpus, which otherwise would require laborious tagging of large corpora. The suggested method depends primarily on the following property: a lexical entry in Hebrew may have many different word forms, some of which are ambiguous while the others are not. Thus, disambiguation of a given word can be achieved using other word forms of the same lexical entry. Even though it was originally devised and implemented for dealing with the problem in Hebrew, the basic idea can be extended and used to handle similar problems in other languages with rich morphology.

*This research was partially supported by grant number 120-741 of the Israel Council for Research and Development.
1 Introduction

This paper addresses the problem of morphological disambiguation in Hebrew by extracting statistical information from an untagged corpus. Let us start with a few definitions and terminology to be used throughout this paper.

Given a language $L$, and a word $w \in L$, we can find (manually or automatically by a morphological analyzer for $L$) all the possible morphological analyses of the word $w$. If $w$ has $k$ different analyses $A_1, \ldots, A_k$, will be used to denote these $k$ analyses. A word is morphologically ambiguous if $k \geq 2$.

Given a text $T$ with $n$ words: $w_1, \ldots, w_n$, for each morphologically ambiguous word $w \in T$, with $k$ analyses: $A_1, \ldots, A_k$, there is one analysis$^1$ $- A_r$ which is the right analysis, $1 \leq r \leq k$, while all the other $k - 1$ analyses of $w$ are wrong analyses. The same word $w$, in a different text, may have, of course, a different right analysis, thus, right and wrong in this case are meaningful only with respect to the context in which $w$ appears.

Morphological disambiguation of a text $T$ is done by indicating for each ambiguous word in $T$ - which of its different analyses is the right analysis. This can be done manually, by a speaker of the language, or automatically by a computer program. When dealing with automatic disambiguation of a text it is sometimes useful to reduce the ambiguity level in the text. Reduction of the ambiguity level of an ambiguous word $w$, with $k$ morphological analyses: $A_1, \ldots, A_k$, occurs when it is not possible to indicate which of its analyses is the right one, but it is possible to select from $A_1, \ldots, A_k$, a sub-set of $l$ analyses $2 \leq l < k$, such that the right analysis of $w$ is one of these $l$ analyses. In the case where $l = 1$, we say that the word $w$ is fully disambiguated.

Since this paper suggests a method for morphological disambiguation using probabilities, the notion of a priori probabilities is also required. For a word $w$ with $k$ analyses: $A_1, \ldots, A_k$, we consider $k$ probabilities: $P_1, \ldots, P_k$, where $P_i$ is the probability that $A_i$ is the right analysis of the word $w$, independently of the context in which $w$ appears. Since $w$ has exactly $k$ different analyses: $\sum_{i=1}^{k} P_i = 1$.

Before starting, we would like to give a short review, describing the structure of this paper. Sections 2-4 include a description of the morphological ambiguity problem in Hebrew, followed by the claim that knowing the a priori probabilities of an ambiguous word can be very effective for automatic morphological disambiguation in Hebrew.

$^1$We will assume that there is only one right analysis, although, in rare cases, there might be more than one.
Then, in Sections 5–6, we present the key idea of this paper: How to acquire a good approximation for the a priori probabilities from an untagged corpus. Using this method we can find for each ambiguous word \( w \) with \( k \) analyses: \( A_1, \ldots, A_k \), probabilities \( \overline{P}_1, \ldots, \overline{P}_k \) which are an approximation to the a priori probabilities: \( P_1, \ldots, P_k \). A description of an experiment which serves in order to evaluate the approximated a priori probabilities which were calculated using an untagged corpus, will be given in Section 7.

Finally, in Section 8, a simple strategy for morphological disambiguation in Hebrew using a priori probabilities, will be described. This simple strategy was used in an experiment which was conducted in order to test the significance of the a priori probabilities as a basis for morphological disambiguation in Hebrew. The experiment shows that using our method we can significantly reduce the level of ambiguity in a Hebrew text. The usefulness of the method is measured by two parameters: discrimination factor and precision. The discrimination factor measures the proportion of analyses ruled out by the method, and the precision measures the proportion of words for which the right analysis is not ruled out. The discrimination factor we obtain using our method is 89% and the precision is 90%.

2 Morphological Ambiguity in Hebrew

Morphological ambiguity is a severe problem in modern Hebrew due to the dimension of this phenomenon. Thus, finding methods to reduce the morphological ambiguity in the language is a great challenge for researchers in the field, and for people who wish to develop natural language applications for Hebrew.

Table 1 demonstrates the dimension of the morphological ambiguity in Hebrew. The data was obtained by analyzing large texts, randomly chosen from the Hebrew press, consisting of nearly 40,000 words. According to this table, the average number of possible analyses per word was 2.1, while 55% of the words were morphologically ambiguous. The main reason for this amount of ambiguity is the standard writing system used in modern Hebrew. In this writing system not all the vowels are represented, several letters represent both consonant and different vowels, and gemination is represented by a single letter [Orn86, Orn91]. The rich morphology of the language and the fact that many prepositions and particles are attached to the word, forming a single string, further contribute to the morphological ambiguity.

In order to demonstrate the complexity of the problem, we should take a
Table 1: The dimension of morphological ambiguity in Hebrew

closer look at the Hebrew morphology. First, a morphological analysis of a word in Hebrew should extract the following information:

- lexical entry.
- category.
- tense (for verbs only).
- short words attached before the entry.
- status – indicating if a noun is in its construct or absolute form.
- gender, number, person (for nouns adjectives and verbs).
- gender, number, person (for pronoun suffixes).

For example, the morphological analysis of the Hebrew string WK$RAYTYW, written in a Latin transliteration, is as follows:

- lexical entry: RAH – the verb ‘to see’.
- category: verb.
- tense: past.
- attached words: W + K$ = ‘and when’.
- gender: feminine/masculine, singular, first person.
- object pronoun: masculine, singular, third person.
Thus, WK$RAYTYW should be translated into English as: ‘and when I saw him’.

Second, to see the nature of the morphological ambiguity in Hebrew, consider, as an example, the following string – HQPH, which has three possible analyses:

1. The determiner H + the noun QPH (‘the coffee’).
2. The noun HQPH (‘a detour’).
3. The noun HQP + the feminine possessive suffix H (‘her perimeter’).

3 Former Approaches

Eliminating or reducing the ambiguity at this early stage of automatic processing of Hebrew is very crucial for the efficiency and the success rate of parsers and other natural language applications. It should be noted that the morphological ambiguity in Hebrew makes even “simple” applications – as is often considered when dealing with other languages – complicated.

One good example for this is full-text retrieval systems [Cho80]. Such systems must handle the morphological ambiguity problem. To see that, consider, for example, the case where we look for all the texts with the word HQPH (‘a detour’). Without morphological disambiguation, we get many texts which really include the word H+QPH (‘the coffee’), or even HQP+H (‘her perimeter’) [Orn87]. Another application which is more difficult in Hebrew than in other languages is text to speech systems, which cannot be implemented in Hebrew without first solving the morphological ambiguity, since in many cases different analyses of a word imply different pronunciations.

The notion that this ambiguity problem in Hebrew is very complicated and that it can be dealt with only by using vast syntactic and semantic knowledge, led the researchers to look for solutions involving a considerable amount of human interaction.

Ornan [Orn86], for instance, developed a new writing system for Hebrew, called ‘The phonemic script’. This script enables the user to write Hebrew texts which are morphologically un-ambiguous, in order to use them later as an input for various kinds of natural language applications. However, since regular Hebrew texts are not written in this script, they first must be transcribed to phonemic texts. Choueka & Lusignan [CL85] presented a system
for morphological tagging of large texts which is based on the short context of the word but depends heavily on human interaction.

Methods using the short context of a word in order to resolve ambiguity (usually categorical ambiguity) are very common in English and other languages [DeR88, Chu88, Kar90]. A system using this approach was developed by Levinger & Ornan in order to serve as a component in their project of morphological disambiguation in Hebrew [LO]. The main resource, used by this system for disambiguation, is a set of syntactic constraints which defined manually by the authors and which followed two theoretical works which defined short context rules for Hebrew [Pin75, Alb92]. The syntactic constraints approach, which is really an extension of the short context approach, was found to be useful and reliable, but its applicability (the proportion of ambiguous words which were fully disambiguated) was very poor. Hence, the overall performance of this system is much less promising than in other languages. These results can be explained by the following properties of the ambiguity problem in Hebrew:

1. In many cases two or more alternative analyses share the same category.
2. Many short context constraints which are very useful in other languages (using determiners, prepositions etc.), are not relevant in the Hebrew case since they are exploited in the morphological analysis level.
3. Hebrew is a language with a highly free word order.

4 Our Approach

The purpose of this paper is to suggest a new approach to deal with the above-mentioned problem. This approach provides a highly useful data which can be used by systems for automatic, man-independent morphological tagging of Hebrew texts. In order to justify and to motivate people to use our approach, we must first make an important observation:

Although the Hebrew language is highly ambiguous morphologically, it seems to be that in many cases a native speaker of the language can accurately “guess” the right analysis of a word, without even being exposed to the concrete context in which it appears. The accuracy can even be enhanced if the native speaker is told from which sub-language the ambiguous word was taken.
If this observation is true, we can now suggest a simple strategy for automatic tagging of Hebrew texts:

For each ambiguous word - find the a priori probabilities of each possible analysis. If any of these analyses is substantially more frequent than the others choose it as the right analysis.

As we have already noted, by saying a priori probabilities, we mean the probability of a given analysis to be the right analysis of a word, independently of the context in which it appears. It should be emphasized that having these a priori probabilities enables us not only to use them, rather naively in the abovementioned strategy, but also to incorporate these probabilities into other systems which exploit higher kind of knowledge (syntactic, semantic etc.).

5 Acquiring the Probabilities

Adopting this approach leaves us with the problem of finding the a priori probabilities for the different analyses of every ambiguous word in the language. Since we use a large corpus for this purpose, the a priori probabilities we acquire must be considered relative to this specific corpus.

One way to acquire a priori probabilities from a corpus is to use a large tagged corpus. Given a corpus in which every word is tagged with its right analysis, we can find the a priori probabilities as reflected in the corpus. This is done by simply counting for each analysis the number of times that it was the right analysis, and using these counters to calculate the probability of each analysis being the right one. The main drawback of this solution is the need for a very large tagged corpus. The method we are about to present saves us the laborious effort of tagging a large corpus, and enables us to find a good approximation to the a priori probabilities by learning about them from an untagged corpus.

This might seem, at first sight, as an impossible mission. When we see the word HQPH in an untagged corpus we cannot decide which of its possible readings is the right one. The key idea is to shift each of the analyses of an ambiguous word in such a way that they all become distinguishable. To be more specific, for each possible analysis (lexical entry + the morphological information), we define a set of words which we call Similar Words (SW). An element in this set is another word form of the same lexical entry which has similar morphological attributes as the given analysis. These words are assumed similar to the analysis in the sense that we expect them to have
relatively the same frequency in the language as the analysis they belong to. A reasonable assumption of this kind would be, for instance, to say that the masculine form of a verb in a certain tense in Hebrew is expected to have relatively the same frequency as the feminine form of the same verb, in the same tense. To see a concrete example, consider the word RAH and one of its analyses: the verb ‘to see’, singular, masculine, third person, past form. The SW set of this analysis contains the following word:

- RATH - singular, feminine, third person, past form.

The choice of which words should be included in the SW set of a given analysis is determined preliminarily using the intuition of a native speaker. Nevertheless, the elements in the SW sets are not determined for each analysis separately, but rather are generated automatically, for each analysis, by changing the contents of one or a few morphological attributes in the morphological analysis. In the previous example the elements are generated by changing the contents of the gender attribute in the morphological analysis, while keeping all the other attributes unchanged.

By choosing the elements in the SW set carefully so that they meet the requirement of similarity, we can study the frequency of an analysis from the frequencies of the elements in its SW set. Note that we would also like to choose the words for the SW sets such that they are morphologically unambiguous. We will assume that this is the case in the following examples, and we will return to this issue in the next section.

To illustrate the whole process, let us take again the ambiguous word HQPH, which has three different analyses. The SW sets for each analysis is as follows:

- HQPH (‘a detour’)
  \[ SW = \{ HHQPH (‘the detour’) \} \]

- H+QPH (‘the coffee’)
  \[ SW = \{ QPH (‘coffee’) \} \]

- HQP+H (‘her perimeter’)
  \[ SW = \{ HQPW (‘his perimeter’), HQPM, HQPN (‘their perimeter’, masculine / feminine) \}. \]

There are some exceptions, such as YLDH – (she) gave birth.
Given the $SW$ set of each analysis we can now find out how many times each word appears in the corpus, calculate the expected frequency of each analysis and get the desired probabilities by normalizing the frequency distribution.

If the similarity assumption we made was totally correct, namely, that each word in the $SW$ set appears exactly the same number of times as the related analysis, we would expect to get a neat situation such as the following (assuming that the ambiguous word HQPH appears 200 times in the corpus):

- $SW = \{ HHQPH = 20 \}$
- $SW = \{ QPH = 180 \}$
- $SW = \{ HQPW = 0, HQPM = 0, HQPN = 0 \}$.

These counters suggest that if we manually tagged the 200 appearances of the string HQPH in the corpus, we would find out that the first analysis of HQPH is the right one 20 times out of the 200 times that the word appears in the corpus, that the second analysis is the right one 180 times, and that the third analysis is not the right analysis even once.

Using these counters we can relate the following a priori probabilities to the three analyses of HQPH: 0.10 0.90 0.00, respectively. These probabilities must be considered as an approximation to the real a priori probabilities, due to the following reasons:

1. The words in the $SW$ set are only expected to appear approximately the same number of times as the analysis they represent.
2. The number of times the ambiguous word appears in the corpus (which is really the size of the sample which we use to calculate the a priori probabilities), and the exact frequency distribution of its different analyses, should be taken into consideration in order to give a statistical evaluation for the probabilities we calculate. A statistical analysis of this kind will be given in the full version of this paper.

Following are the numbers we got for the ambiguous word HQPH, by consulting the corpus we worked with (the word HQPH appears 202 times in our corpus):

- $SW = \{ HHQPH = 3 \}$

---

3The numbers in this example are fictitious. They were chosen in order to clarify our point.
• $SW = \{ QPH = 368 \}$
• $SW = \{ HQPW = 0, HQPM = 0, HQPN = 0 \}$.

By now applying the algorithm of the next section on these counters, we can calculate the desired probabilities.

6 The Algorithm

A common case that must be handled by the algorithm is the case where a certain word appears in more than one $SW$ set. In that case, we would like to consider the counter of such a word appropriately. The algorithm takes care of this problem and works as follows:

• We start by setting the proportions between the different analyses to be equal.

• For each analysis we compute its average number of appearances, by summing up all the counters for each word in the $SW$ set, and dividing this sum by the $SW$ size. Note that in this stage we also include the ambiguous word in each of the $SW$ sets.$^4$

• If a word appears in several $SW$ sets, we calculate its contribution to the total sum according to the proportions between all those sets, using the proportion numbers of the previous iteration.

• Calculate the new proportions between the different analyses by computing the proportions between the average number of appearances of each analysis.

• This process is iterated until the new proportions calculated are sufficiently close to the proportions calculated in the previous iteration (i.e., $a^{(i)}$ the proportion of the $i$-th iteration satisfies $|a^{(i)} - a^{(i-1)}| < \varepsilon$ for some prespecified threshold $\varepsilon > 0$).

• Finally, the proportions are normalized to obtain probabilities.

Applying this algorithm to the sets and the counters extracted from the corpus (our previous example), yields the following probabilities$^5$:

$^4$This is done mainly in order to handle cases where a certain analysis has an empty $SW$ set, since it does not have a natural similar words.

$^5$In fact, the third analysis should get a zero probability, but the algorithm is intentionally designed not to assign a zero probability to any of the possible analyses.
\begin{itemize}
\item HQPH = 0.0113
\item H + QPH = 0.9870
\item HQP + H = 0.0017.
\end{itemize}

Although this method for acquiring a priori probabilities gives very good results for many ambiguous words, as will be shown in the next section, we detected two types of inherently problematic cases:

1. Due to the high degree of morphological ambiguity in Hebrew, some of the words in the $SW$ sets are also expected to be ambiguous. As long as the other possible analyses of such a word are not too frequent, it only slightly effects the final probabilities. Otherwise, we might get wrong results by erroneously crediting the high number of appearances of such a word to one of the analyses. For this reason, we try to construct the $SW$ sets from as many suitable elements as possible, in order to be able to detect "misleading" words of this sort.

2. Occasionally, the $SW$ sets defined for two different analyses are actually the same. Thus, a differentiation between those two analyses cannot be done using our method.

Another potentially problematic case is the coverage problem, which arises whenever we do not have enough data in the corpus for disambiguation of a certain word (see a discussion on this problem in [DIS91]). This problem was found to occur very rarely, and thus, when we get very small counters (less than 20) for the words in the $SW$ sets, we simply ignore the data and arbitrarily give a uniform probability to all the analyses.

7 Evaluating the Probabilities

In order to evaluate the approximated a priori probabilities acquired we need to compare these probabilities to the real a priori probabilities. Since the approximation we acquire is relative to the corpus we have been using - texts taken from the Hebrew newspaper 'Ha'aretz', we have to calculate the real a priori probabilities from texts taken from the same source. For this purpose

\footnote{\textit{Due to technical reasons, we cannot decide whether a given word is ambiguous or not when we automatically generate the words for the $SW$ sets.}}

\footnote{\textit{We would like to thank 'Ha'aretz' for the permission to use magnetic tapes from their archives.}}
we used a small corpus consisting of more than 400,000 words taken from the same newspaper. We randomly picked from this corpus 30 ambiguous words which appear more than 100 times each. These 30 words serve as our test-sample in order to evaluate the approximated a priori probabilities. The total number of different analyses for the ambiguous words in this test-sample was 97.

For each of these ambiguous words we extracted from the small corpus all the sentences in which the ambiguous word appears. We then manually tagged each ambiguous word and found for each of its analyses how many times it was the right analysis. For example, the word AWLM has the following two morphological analyses:

1. The particle – ‘but’.
2. The noun – ‘a hall’.

The word AWLM appeared 236 times in the small corpus. By manually tagging all the relevant sentences we found out that the first analysis, ‘but’, was the right analysis 232 times, and the second analysis, ‘a hall’, was the right analysis only 4 times. Given these numbers we can calculate the empirical relative weights of these two analyses: 232/236, 4/236 and the empirical probabilities: 0.983, 0.017, respectively.

In the same way, using the small corpus we found the empirical probability, \( P_{\text{emp}} \), for each of the 97 analyses in the test-sample. These probabilities should be compared now with the approximated probabilities, \( P_{\text{appr}} \), calculated using our method.

Table 2 shows the empirical probabilities and the approximated probabilities for 5 representative ambiguous words, taken out of the 30 words in our test-sample. The accuracy of the approximated probabilities as shown in this table, properly reflects the accuracy found for all the analyses in the test-sample. Furthermore, for 28 words out of 30, the analysis with the highest empirical probability was also the one with the highest approximated probability, and for more than 85% of the analyses the difference between \( P_{\text{emp}} \) and \( P_{\text{appr}} \) was very small (less than 5%).

A full statistical analysis taking into account both the fact that \( P_{\text{emp}} \) was calculated according to a small sample, and the fact that \( P_{\text{appr}} \) was calculated using the counters extracted from the untagged corpus will appear in the full paper. Nevertheless, even a superficial analysis indicates that the results are statistically significant.
In the previous section we have compared the a priori probabilities obtained by our method to the probabilities found by manually tagging a small corpus. We found that the acquired probabilities are truly a good approximation for the a priori probabilities. In this section we describe an experiment which was conducted in order to test the effectiveness of the a priori probabilities for morphological disambiguation in Hebrew.

Following are the main components in our project which were used in order to conduct the experiment:

1. A robust morphological analyzer for Hebrew which gives for each word

---

8 The morphological analyzer was developed at the IBM scientific center, Haifa, Israel. We would like to thank the center for letting us use it for research purposes.
in the language all its possible analyses. The input for our project is supplied by this module.

2. An interactive program for manually tagging Hebrew texts written in order to rapidly tag large texts. It was used to mark the right analysis for each ambiguous word in order to be used later to evaluate the performance of our method.

3. A Hebrew corpus which was build especially for our project and consists of 11 million words taken from the daily newspaper ‘Ha’aretz’.

4. A hash table which stores all the words in the corpus. Each word is accompanied by a counter indicating how many times it appears in the corpus. Since this is the only information we extract from the corpus, our algorithm needs only this hash table and is therefore very efficient.

5. A morphological generator for Hebrew which was written especially for this project. The SW sets for every analysis are generated using this module. Due to some technicalities, we are not able to use the morphological analyzer at this stage, and thus we cannot identify ambiguous words in the SW sets.

6. An implementation of the iterative algorithm which calculates the probabilities.

7. A simple selection algorithm which reduces the level of morphological ambiguity using the probabilities obtained from the corpus. The algorithm uses two thresholds, an upper threshold and a lower threshold, which serve to choose the right analysis or to rule out wrong analyses, respectively.

A set of 21 articles was randomly selected from the Hebrew press in order to test the performance of the method. The total number of words in these test-texts was 3,400 out of which nearly 50% were morphologically ambiguous.

We test the performance of the method on the test-texts from two different perspectives. First, by using the probabilities only for ambiguous words which can be fully disambiguated – a single analysis can be selected as the right analysis. Following is a table showing the applicability (the proportion of ambiguous words which were fully disambiguated) and the precision of our method for full disambiguation:
Table 3: Applicability and precision for full disambiguation

However, the a priori probabilities can also be used in order to reduce the ambiguity level in the text. We test the performance of the method in reducing the ambiguity level by looking at the analyses ruled out by the system. The performance of the method in this sense is much more interesting and important since it examines, more accurately, the quality of the probabilities as data for other, more sophisticated, systems which use higher level of information. The results are shown in the following table:

<table>
<thead>
<tr>
<th>Wrong analyses</th>
<th>Ruled out</th>
<th>Discrimination factor</th>
<th>Ambiguous words</th>
<th>Correct selections</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>3260</td>
<td>2902</td>
<td>89%</td>
<td>1613</td>
<td>1444</td>
<td>90%</td>
</tr>
</tbody>
</table>

Table 4: Discrimination factor and precision for reducing the ambiguity level in the text

These results demonstrate the effectiveness of a priori probabilities in reducing the ambiguity level in a Hebrew text, and it seems that by using this kind of information combined with other approaches for morphological disambiguation in Hebrew, we can become very close to a practical solution for this problem.

9 Conclusions

A method to acquire a priori probabilities from an untagged corpus has been described. The main idea was to use the rich morphology of the language to learn the frequency of a certain analysis from the frequency of other word forms of the same lexical entry.
The results of the experiment just described confirm the observation we made about the nature of the morphological ambiguity problem in Hebrew. It can be argued, therefore, that the computer with its pedant morphological knowledge, is facing a much complex problem than is faced by a human while reading a Hebrew text. This observation is also supported by the fact that humans are very often surprised to see the amount of possible analyses of a given ambiguous word. It may even have a significance from a psycholinguistic point of view, by suggesting that this kind of probabilities are also used by a human reader of Hebrew.

The a priori probabilities cannot serve as the only source of information for morphological disambiguation since they are imperfect by definition – they always choose the same analysis as the right one, regardless of the context in which the ambiguous word appears. Thus, as we have already mentioned, we are planning to incorporate these probabilities into an existing system which makes use of syntactic constraints.

We hope that the combination of these two methods will yield an efficient system with a very high applicability and precision.

10 Acknowledgment

We would like to thank Ayal Shiran and Ohad Zeliger for programming support of this project, and Ido Dagan for useful discussions concerning this paper.

References


