Using Wikipedia Links to Construct Word Segmentation Corpora

David Gabay and **Ziv Ben-Eliahu** and **Michael Elhadad**
Ben Gurion University of the Negev
Department of Computer Science
POB 653 Be'er Sheva, 84105, Israel
(gabayd|ben-elia|elhadad)@cs.bgu.ac.il

Abstract

Tagged corpora are essential for evaluating and training natural language processing tools. The cost of constructing large enough manually tagged corpora is high, even when the annotation level is shallow. This article describes a simple method to automatically create a partially tagged corpus, using Wikipedia hyperlinks. The resulting corpus contains information about the correct segmentation of 523,599 non-consecutive words in 363,090 sentences. We used our method to construct a corpus of Modern Hebrew (which we have made available at http://www.cs.bgu.ac.il/~nlpproj). The method can also be applied to other languages where word segmentation is difficult to determine, such as East and South-East Asian languages.

Word Segmentation in Hebrew

Automatic detection of word boundaries is a non-trivial task in a number of languages. Ambiguities arise in writing systems that do not contain a word-end mark, most notably East-Asian logographic writing systems and South-East Asian alphabets. Ambiguities also appear in alphabets that contain a word-end mark, but sometimes allow agglutination of words or insertion of a word-end mark inside a single word. A discussion of the definition of "word" in general can be found, for example, in (Sciullo and Williams 1987). We focus in this work on word segmentation in unvocalized Modern Hebrew. According to common definitions (see (Adler 2007) Chapter 2 for a recent review), a Hebrew word may consist of the following elements: a proclitic, a stem, an inflectional suffix and a pronominal suffix (enclitic). In the official standard defined by the Mila Knowledge Center for Hebrew,[1], as well as other work in parts of speech (POS) tagging and morphological analysis of Hebrew, inflectional suffixes are referred to as attributes of the stem. The problem of word segmentation in Hebrew concerns, therefore, the identification of the proclitics, stem and enclitics of a word, while POS tagging refers to assigning the correct part of speech to each part. Morphological disambiguation refers, one step further, to the complete analysis of all the morphological attributes of each word part. Proclitics include conjunctions, prepositions, complementizers and the definite article. They are composed of one or two letters and follow a strict order. The segmentation of a given word is often ambiguous. In a corpus of 40 million tokens, we found that there are, on average, 1.26 different possible segmentations per token, even when only proclitics are being considered. For example, the word[2] *$btw* may be segmented, among other options, as:
$-b-tw, meaning "that in a note"
$-bt-w, meaning "that his daughter"
$btw, meaning "(they) went on strike"
The major cause for ambiguity is proclitics, as enclitics are rare in Modern Hebrew. When performing POS-tagging or full morphological analysis, word segmentation can be performed as a separate first step (Bar-Haim, Sima'an, and Winter 2005), alongside POS-tagging (Adler and Elhadad 2006) or even in joint inference with syntactic analysis (Cohen and Smith 2007), following (Tsarfati 2006). Word segmentation may also be considered a separate task, easier than full morphological analysis but still far from trivial. As a separate task, it has practical value on its own - narrowing search results, for example. Current work in POS tagging and morphological analysis reports success rate of 97.05 percent in word segmentation for supervised learning (Bar-Haim, Sima'an, and Winter 2008). In the case of unsupervised learning, 92.32 percent accuracy is reported by (Adler and Elhadad 2006) in segmentation and simple POS tagging, without full morphological analysis. The lack of annotated corpora is one of the problems in assessing NLP tools for modern Hebrew. In this work, we propose an original method that exploits Wikipedia data to obtain high-quality word segmentation data.

Wikipedia Links and Word Segmentation

Using Wikipedia as a data source for NLP and AI tasks has become common in recent years, as work in different fields makes use of the attractive Wikipedia qualities: it is easily accessible, large and constantly growing, multilingual, highly structured, and deals with a considerable number of topics. In this work, we focus on the form of hyperlinks in Wikipedia. Wikipedia hyperlinks, together with a man-

[1] http://www.mila.cs.technion.ac.il

[2] For the sake of simplicity, we use only transliterations in this article. Translitatetion follows ISO standard.

ual mapping of article names into WordNet labels, have already been used by (Mihalcea 2007) to generate sense-tagged corpora. We follow a similar intuition to address word segmentation. We make use of the structure of hyperlinks within Wikipedia, that is, hyperlinks between different articles within Wikipedia. Internal Wikipedia hyperlinks consist of a surface form, visible to the reader, and the name of the Wikipedia article to which the hyperlink leads, which is a unique identifier. The syntax of internal hyperlinks in Wikitext - the markup language in which Wikipedia is written - is as follows: hyperlinks are surrounded by double brackets. They may contain a pipe (|) sign, in which case the text preceding the pipe is the name of the linked article, and the part following it is the text that is visible to the reader.

If no pipe appears, then the name of the linked article will be the visible text. For example, the Wikitext sentence (taken from the English Wikipedia): *"During the [[Great Depression in the United States|Great Depression]] of the 1930s, Roosevelt created the [[New Deal]]"* will be parsed to generate the sentence *"During the Great Depression of the 1930s, Roosevelt created the New Deal"*. The first hyperlink will lead to the article *"Great Depression in the United States"* and the second to the article *"New Deal"*. Text that is adjacent immediately before or after the double brackets will be adjacent to the visible text of the hyperlink.

Constructing a Word Segmentation Corpus from the Hebrew Wikipedia

The format of internal hyperlinks in Wikipedia makes it a source of information on word segmentation. We describe how we exploit the form of Wikipedia hyperlinks to construct a corpus in which some of the words are (fully or partially) segmented. For our task, the tags in the corpus will denote the existence or absence of proclitics.[3] We identified five types of internal hyperlinks and text combinations relevant to word segmentation: 1. [[A]], 2. p[[A]], 3. [[A|pA]], 4. p[[B|A]], 5. [[A|B]]; where p is a sequence of one or more proclitics, and A and B are different text segments, consisting of one or more words (note that types 2 and 3 are equivalent in terms of Wiki syntax). Hyperlinks of the first three forms provide reliable data on the correct segmentation of the first word in the text A (provided that A, which is an article name, does not begin with a proclitic, an issue discussed in the next subsection). For example, the hyperlink *[[lwndwn]] (London)* of type 1 indicates that the token *lwndwn* does not contain proclitics, and the hyperlinks *l[[lwndwn]]* of type 2 and *[[lwndwn|llwndwn]]* of type 3 both indicate that the word *llwndwn* should be segmented into *l+lwndwn (to-London)*. Hyperlinks of types 4 and 5 may also contain information on word segmentation, but this information is not consistent, since prepositional letters may appear both in and out of the hyperlink. In the first step of the construction of our word segmentation corpus, we removed all Wikitext syntax from the article code, except for internal hyperlinks of types 2 and 3, and hyperlinks of the first type in which the first word in the brackets is not ambiguous in terms of word segmentation. In the second step, the double brackets format was replaced by a unified XML setting. The result was a corpus of Hebrew text in which some of the tokens are tagged for word segmentation. Since automatic taggers for Hebrew work at the sentence level and do not make use of higher context, we also constructed a more succinct corpus containing only complete sentences (and not list items, or other isolated text units) in which one or more tokens contain segmentation data. The resulting corpus includes 363,090 sentences with 523,599 tagged words (out of 8.6 million tokens altogether), taken from 54,539 Wikipedia articles. Each tagged word in the resulting corpus is potentially ambiguous, meaning that its beginning matches some sequence of proclitics. (other than *h*, for reasons specified in the next subsection).

Accuracy

The method described in this section relies on two assumptions: that the vast majority of article names in Wikipedia do not start with a proclitic and that almost all hyperlinks in Wikipedia are written correctly, according to Wikitext syntax. The first assumption does not hold for one proclitic - *h*, the definite article. A non-negligible amount of articles do start with this proclitic (*e.g., hmhpkh hcrptit, The-revolution the-French, the French revolution*). Due to this fact, and to the fact that *h*, unlike any other proclitic, may be covert, we decided not to include any data regarding the proclitic *h* in the corpus. With the *h* excluded, we manually checked a random sample of 600 tags generated by our corpus and found all of them to be correct. Our assumption on article names seems to hold in this case since article names are given according to a naming policy, and are usually names of people, places, organizations, objects or concepts, that do not start with a proclitic, except for the definitive article case. However, a small fraction of article names do start with other proclitics. This group includes titles of works of art (such as movies and books) and expressions. There are several ways to address this issue. The first is to ignore it, as it corresponds to a very small number of tags: a randomly picked sample of 600 article names with a possible segmentation ambiguity (note that this check is not the same as picking hyperlinks at random) contained just one title that begins with a proclitic - *lgnawlwgih $l hmwsr (to-Genealogy of the-Morality, On the Genealogy of Morality)*, the name of a book. This low amount of noise can be considered bearable. A second option of dealing with the article names problem is to use the fact that Wikipdeia articles are categorized, and that almost all 'bad' article names are concentrated in few categories. We could omit the tags in the case of possibly ambiguous article names from categories such as movies, books or expressions. A third option is to review the list of possibly ambiguous article names, and remove from the corpus all tags that are based on hyperlinks to articles whose names start with a prepositional letter. The first option will result in some noise in the corpus, the second will not allow full usage of the data, and the third requires some manual labor, although considerably less than the effort needed to tag the corpus manually. (Note that most articles have several

[3] We did not deal with enclitics, as they are quite rare in Modern Hebrew.

hyperlinks leading to them).

Our second assumption - that hyperlinks are almost always correct - seems to hold since great effort is made by Wikipedia editors to remove syntax errors. Still, it may be wise to use the articles' versions history to filter out new articles, or articles that were only edited a few times, or by too few people, before generating a tagged corpus out of a snapshot of Wikipedia.

Using the Word Segmentation Corpus

An immediate usage of the "Wikipedia segmentation corpus" generated by our method is to evaluate Hebrew parts-of-speech taggers. Large-scale tagged corpora are needed to properly evaluate the quality of any NLP tool, and are of particular importance at the current stage of research in Hebrew POS-tagging. In the last decade, significant progress has been made in this field. Further improvements in existing systems require fine tuning, for example, redefining the tagset (Netzer et al. 2007) in a way that affects a small percentage of the words. The impact of such changes is difficult to measure using currently available annotated corpora. Since Hebrew is highly inflectional, the number of POS categories and word-forms is high, and a large corpus is needed so that a sufficient number of combinations would be encountered. Constructing large manually tagged corpora is an expensive and laborious task, and lack of tagged corpora remains a bottleneck. We first used the corpus obtained by our method to evaluate the performance of Adler's HMM POS tagger (Adler and Elhadad 2006). The tagger gave the correct segmentation on 74.8 percent of the tagged words. This result suggests that word segmentation in Hebrew is not yet fully resolved, at least in the case of unsupervised taggers. The accuracy rate is significantly lower than that reported in previous results, as can be explained by the high rate of out-of-vocabulary words and by the fact that every token in the Wikipedia segmentation corpus is ambiguous in terms of segmentation, while previous success rates refer to unambiguous words as well. Any POS-tagger or morphological disambiguation tool must also provide word segmentation: it is reasonable to assume that the results on segmentation are a good indicator of overall performance, as failure to tag a word correctly is likely to lead to segmentation errors further on in the sentence. In all reported results, heuristics that improve segmentation also improve tagging accuracy. Thus, a very large segmentation corpus may be used, together with a small fully-annotated corpus, to evaluate overall performance of tools that do more than segmentation. The Wikipedia corpus cannot be used to effectively train supervised segmentors or taggers on its own, since it only contains partial, non-representative data on the segmentation of words in the corpus. It can be used to improve the training done by a manually tagged corpus. It may also be used to tune the initial conditions in unsupervised methods (see (Goldberg, Adler, and Elhadad 2008) for details on the importance of initial conditions). Experiments are currently being performed to evaluate the effectiveness of this approach.

Future work

The Hebrew Wikipedia Word Segmentation corpus can be extended to include enclitics. With some manual labor, it can also be extended to deal with the proclitic *h*. The general approach described in this paper may be applied to any language in which word segmentation is a problem and where a large enough Wikipedia (or other sources written in Wikitext) exists, while large-scale manually annotated corpora do not. Thai, with more than 34,000 articles in its version of Wikipedia (as of March 2008), is an example of such a language.

References

[Adler and Elhadad 2006] Adler, M., and Elhadad, M. 2006. An unsupervised morpheme-based hmm for hebrew morphological disambiguation. In *ACL06*, 665–672. Morristown, NJ: Association for Computational Linguistics.

[Adler 2007] Adler, M. 2007. *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. Ph.D. Dissertation, Ben-Gurion University of the Negev, Beer-Sheva, Israel.

[Bar-Haim, Sima'an, and Winter 2005] Bar-Haim, R.; Sima'an, K.; and Winter, Y. 2005. Choosing an optimal architecture for segmentation and pos-tagging of modern Hebrew. In *Proceedings of ACL-05 Workshop on Computational Approaches to Semitic Languages*.

[Bar-Haim, Sima'an, and Winter 2008] Bar-Haim, R.; Sima'an, K.; and Winter, Y. 2008. Part-of-speech tagging of modern hebrew text. *Journal of Natural Language Engineering* 14:223–251.

[Cohen and Smith 2007] Cohen, S. B., and Smith, N. A. 2007. Joint morphological and syntactic disambiguation. In *EMNLP07*, 208–217. Prague, Czech: Association for Computational Linguistics.

[Goldberg, Adler, and Elhadad 2008] Goldberg, Y.; Adler, M.; and Elhadad, M. 2008. Em can find pretty good hmm pos-taggers (when given a good start). In *Proceedings of ACL 2008*.

[Mihalcea 2007] Mihalcea, R. 2007. Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL HLT 2007*, 196–203. Rochester, NY: Association for Computational Linguistics.

[Netzer et al. 2007] Netzer, Y.; Adler, M.; Gabay, D.; and Elhadad, M. 2007. Can you tag the modal? you should! In *ACL07 Workshop on Computational Approaches to Semitic Languages*, 57–65. Prague, Czech: Association for Computational Linguistics.

[Sciullo and Williams 1987] Sciullo, A. M. D., and Williams, E. 1987. *On the Definition of Word*. Cambridge, MA: MIT Press.

[Tsarfati 2006] Tsarfati, R. 2006. Integrated morphological and syntactic disambiguation for modern hebrew. In *Proceedings of CoLing/ACL-06 Student Research Workshop*, 49–54. Association for Computational Linguistics.