Word-Based or Morpheme-Based? Annotation Strategies for Modern Hebrew Clitics

Introduction

Background

In morphologically rich languages a single word may carry different sorts of information and the different morphs may indicate a relation to other elements in the syntactic tree.

The Question

Should we analyze a word as a sequence of morphological units or should we treat orthographic space-delimited words as the primitive units of our analyses?

Our Investigation

We discuss and evaluate the adequacy of Morpheme-based and Word-based annotation strategies for the development of statistical parsers for Modern Hebrew.

The Data

Pronominal Clitics in Modern Hebrew

(1) Prepositions

He saw her

He saw ACC she

(2) Possessive Markers

Our children

Children our

(3) Accusative Markers

He saw her

Saw he ACC her

(4) The Dative Shift in Modern Hebrew

I gave him a present

Gave.1p.sing  to.3p.masc.sing  a-present

I gave him 3 singular dative

I gave him a present

I gave him 3 singular dative

(5) Coordinated Structures in Modern Hebrew

Our children

Children of we

He came to me

Came he to.1p.sing

He came to me

He came to.1p.sing

He came to me

He came to.1p.sing


Theoretical and Methodological Background


We thank the following for their encouragement and support: Meni Adler, James P. Blevins, Michael Elhadad, Tal Frankel, Miki Frankel, Tomer Gophna, Remko Scha, and Jelle Zuidema for comments and discussion. We acknowledge the Netherlands Organisation for Scientific Research (NWO) and the Lynne and William Frankel Center for Computer Sciences in support of the work of R. Tsarfaty and Y. Goldberg respectively.

References


ATTCL

Ambiguity Tolerant Treebank Construction Language.

ACKNOWLEDGMENTS

Our Investigation

We discuss and evaluate the adequacy of Morpheme-based and Word-based annotation strategies for the development of statistical parsers for Modern Hebrew.

Experimental Design

Goal

Compare and contrast the adequacy of the MB and WB annotation strategies for Modern Hebrew pronominal clitics.

Methodology

Evaluate parsing performance of different treebank PCGs corresponding to different annotation strategies.

Data

The Modern Hebrew Treebank version 1.0 (Sima‘an et al., 2001), 5000 sentences from the daily newspaper ‘Ha’aretz’. The Test-Set constitutes the first 500 non-empty sentences.

Evaluation

Problem: Different strategies result in different sentence length

Quantitative: compare PARSEVAL measures of sentences without clitics

Qualitative: analyze differences in the parse-trees that include clitics

Results and Analysis

Quantitative Analysis

For sentences without pronominal clitics the conversion does not have a significant influence on the disambiguation capacity of the resulting grammars.

Qualitative Analysis

Sentences with pronominal clitics the WB analysis is always as good or better than the original MB analysis.

Conclusion

The WB analyses are more faithful to the surface forms thus avoiding the need for preceding segmentation.

The WB resulting treebank grammars provide better PP attachment disambiguation capacity.

The Word-Based (WB) annotation strategy is more adequate than the Morpheme-Based (MB) strategy for training statistical parsers on the Modern Hebrew Treebank.

The original annotation strategy in the Modern Hebrew Treebank.

Advantage | Capture correctly word- and constituent-boundaries discontinuities.

Disadvantage | Does not correspond to the yields of syntactic parse trees.

A Naive Word-Based Strategy

A naive proposal for representing words as the yields of syntactic parse trees.

Advantage | Yields correspond directly to surface forms.

Disadvantage | Grammatically incorrect.

Our Word-Based Strategy

We propose an alternative Word-Based (WB) analysis as inflectional features on top of specialized categories of prepositions / markers. The special categories capture membership in distinct syntactic classes, and the features indicate agreement with a definite phrases, Prepositional Possessive Phrases)

The main source of errors for the WB strategy is its tendency to learn attachment high-prepositions that originate from cliticized elements.

The Word-Based (WB) annotation strategy is more adequate than the Morpheme-Based (MB) strategy for training statistical parsers on the Modern Hebrew Treebank.

The WB resulting treebank grammars provide better PP attachment disambiguation capacity.

The Word-Based (WB) theories consider words the atomic units of the language, and morphological considerations predict generalizations concerning the syntactic behavior of morphologically similar words (Bloomfield 2006).

The WB analyses are more faithful to the surface forms thus avoiding the need for preceding segmentation.

The WB resulting treebank grammars provide better PP attachment disambiguation capacity.

The Word-Based (WB) annotation strategy is more adequate than the Morpheme-Based (MB) strategy for training statistical parsers on the Modern Hebrew Treebank.

The main source of errors for the MB strategy is its tendency to learn attachment high-prepositions that originate from cliticized elements.

References


Published by Elsevier B.V. All rights reserved.