A Corpus Based Morphological Analyzer for Unvocalized Modern Hebrew

Alon Itai and Erel Segal
Department of Computer Science
Technion—Israel Institute of Technology, Haifa, Israel

The Problem

Written Hebrew texts are ambiguous.
The reasons

- The vowels and gemination are omitted.
  קופה = קוּפָּה קו פָּה QWPH
- small words are prepended.
  = וכשתלך WK$TLK
  and when you will go = ו + כש + תלך
- Hebrew morphology is complex
  I loved her = אהב + תי +ה = AHBTIH אהבתיה

The structure of a Hebrew word

morphemes

- the lexical lemma,
- short words such as determiners, prepositions and conjunctions prepended to the word,
- suffixes for possessives and object clitics.

linguistic features

- The linguistic features mark part-of-speech (POS), tense, person etc.

Example

- $QMTI שקמתי
- $iqmati – שִׁקְמָתִי my sycamore

  noun sg possessive-1sg

- $e-qamti – שֶׁקַּמְתִּי that I got up

  connective+verb 1sg past

- $e-qammati – שֶׁקַּמַּתִּי that my hey

  connective + noun sg possessive-1sg

Previous work

Morphological Analyzers

- Rav Millim Choueka
- AVGAD—IBM Haifa Scientific Center Ben-Tur et al. 1992
- Segal 2001

Morphological Disamiguation

- Choueka and Lusingnan 1985
- Albeck 1992
- Levinger, Ornan, Itai 1992; Levinger, 1992
- Carmel and Maarek 1999

Three stages

1. Word stage – find the most probable reading of a word regardless of its context.
2. Pair stage –correct the analysis of a word based on the analysis of its immediate neighbors.
3. Sentence stage – use a syntactic parser to rule out improbable analyses.

Combining all three stages yielded the best results

The Word Stage

- Give each word its most probable analysis.
- How to estimate the probability of each analysis?
- Estimate the probability of each analysis from a large analyzed corpus.
- A large enough corpus does not exist.
- Since each word has many forms, the number of word tokens is so large that many word forms won't appear even in 10M word corpus.

The Word Stage

- Following the "Similar Words Method", (Levinger, Ornan and Itai 1995) estimate the probability of each analysis of an ambiguous word by changing a (single) feature of each analysis, and comparing the occurrences of the resultant words in a large corpus.
- Example: HQPH הקפה
- the coffee: definite to indefinite QPH
- encirclement: indefinite to definite HHQPH
- her perimeter: feminine possessive to masculine possessive HQPW.
- Distribution: QPH=180, HHQFH=18, HQPW=2

and the winner is הקפה (the coffee)

Our variation of the SW method

- To overcome sparseness, we assumed that the lemma and the other morphemes/linguistic features are statistically independent.
  Namely,
  $$P(\text{the coffee}) = P(\text{the}) \times P(\text{coffee}).$$
- Even though the assumption is not valid, the resultant ranking is correct.

Evaluation and Complexity

- Errors 36% → 14.5%

- Complexity of algorithm O(*c*), where *c* is the size of the corpus.
- Keeping a copy of the corpus as an inverse file reduces the complexity to linear in the number of different similar words.

The pair stage

- Following Brill, we learned correction rules from a corpus.
- The initial *morphological score* of an analysis is its probability as obtained at the word stage.
- Correction rules modify the scores by considering pairs of adjacent words, checking if the rule applies, and if so modify the scores.

Example of a correction rule

If the POS of the current tag of w_1 is a proper-noun

and the POS of the current tag of w_2 is a noun

and w_2 has an analysis as a verb that matches w_1 by gender and number,

then add 0.5 to the morphological score of w_2 as a verb, and normalize the scores .

Example

YWSP &DR יוסף עדר

- YWSP = proper noun masc.(Joseph)
- &DR = noun masc. sg. abs indef
  (herd) score=~~0.7~~ 0.467

  normalization
- &DR = verb past 3sg masc.
  (hoed) score=~~0.3~~ ~~0.8~~ 0.533

Learning the Rules from a training corpus

Input: A training corpus, where each word is correctly analyzed.

- Run the word stage on the training corpus.
- Generate all possible rules.
- For each rule, set the correction factor to be the minimum value that does more good than damage.
- Choose the rule that does the maximum benefit.
- Repeat until no rule improves the overall analyses of the training corpus.

Evaluation and Complexity

- Training corpus 4892 word tokens learned 93 rules.
  errors 14.5% → 6.2%

- Complexity of the learning algorithm $O(c^3)$, where c = size of the training corpus.
- Complexity of the correction $O(r \cdot n)$, where r = number of rules,
  n = size of trial text.

The sentence stage

- Use a syntactic parser to rule out improbable analyses.
- The pair stage – adjacent words, the sentence stage – long term dependencies.

Example

- מורה הכיתה הנמוכה נכנס לכיתה

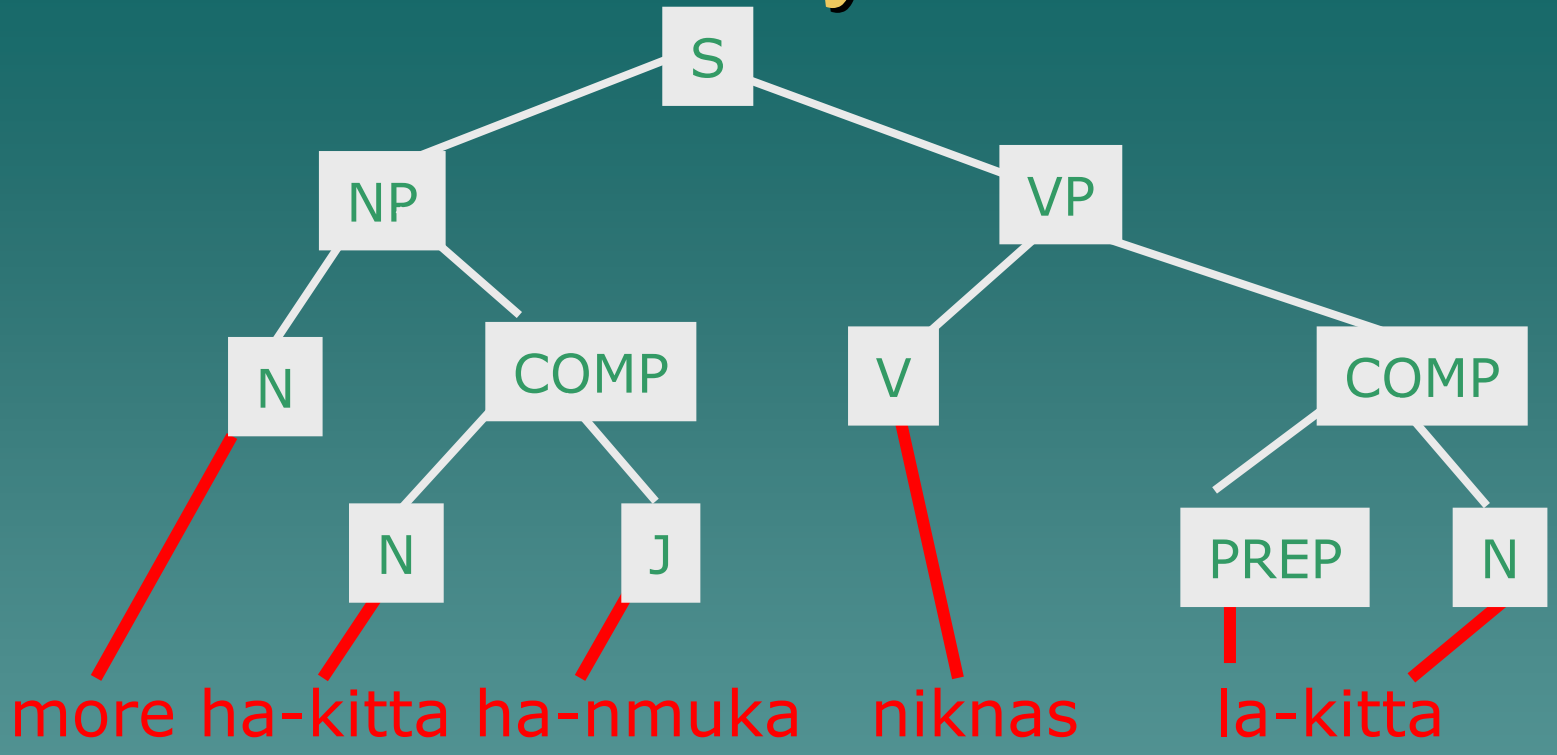
MWRH HKITH HNMWKH NKNS LKITH

more/mora ha-kitta ha-nmuka niknas …

masc/fem verb-masc

more ha-kitta ha-nmuka niknas …

Score of a syntax tree



score(s)
= score(more)×score(ha-kitta) × … ×score(la-kitta)

The challenge: calculate the score of all syntax trees without enumerating all trees

Dynamic Programming

- *Table*[*i,j,A*] = the maximum score of all parses $A \xrightarrow{*} w_i \cdots w_j$
- Fill table by incrasing values of $\ell = j - i$
- $\underline{\ell = 0}$

$$Table[i,i,A] = \max\{s_{im} : A \to t_{im} \in G \text{ and } t_{im} \in T_i\}$$

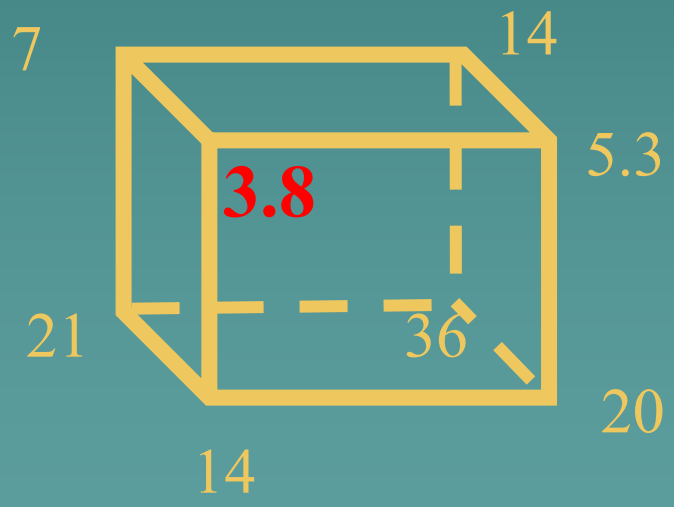
- $\underline{\ell > 0:}$

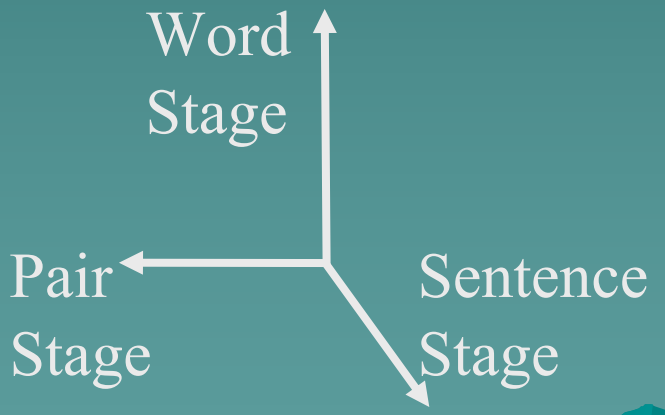
$$Table[i,j,A] = \max_{\substack{A \to BC \in G \\ i \le k < j}} \{Table[i,k,B] \times Table[k+1,j,C]\}$$

Time complexity $O(|G|n^3)$

Evaluation

error rate





Conclusions and Future Work

- We used statistical methods to obtain a 96% accurate morphological disambiguator.

Error Analysis

- Idioms -- $R HPNIM = ministry of the face / interior ministry
- Proper names
- The limits of statistical methods 2%?