Statistical and Learning Methods in Natural Language Processing

Ido Dagan and Shuly Wintner

Assignment 1

Objectives

The final projects you will have to submit all deal with one particular problem: part-of-speech tagging for Hebrew. Each team will implement a different machine learning method for solving the same problem. This will enable us to compare the performance of different ML algorithms on this task.

In order to give each implementation a fair chance, we prefer to have a unique data representation model for all projects. In other words, we prefer that all projects use, as much as possible, the same feature set for training and representing test examples.

Data representation is an art, and is very much dependent on the classification task and the classification model. In order to select the features correctly one must understand what instances of the problem look like and which attributes are useful in order to tell them apart. The objective of this assignment is to define a set of features for POS tagging of Hebrew. We will then discuss the union of the features suggested by all students and select the ones we will think are best.

Task definition

A Hebrew text is given as a sequence of words, one word per line. To simplify processing, words are represented using ASCII characters according to the following transliteration table:

A B G D H W Z X @ I K L M N S & P C Q R $ T

An example text follows:

```
&$RWT AN$IM MGI&IM MTAILND LI$RAL K$HM NR$MIM KMTNDBIM , AK LM&$H M$M$IM
&WBDIM $KIRIM ZWLIM . TWP&H ZW HTBRRH ATMWL BWW&DT H&BWDH WHRWWXH $L
HKNST , $DNH BNW$A H&SQT &WBDIM ZRIM .
```

In addition to the text itself, each word is associated with a list of one or more morphological analyses, produced by an automatic analyzer. Each analysis, which occupies a single line, is a record of morphological information including the part of speech. The format of the morphological record is complex, but the part of speech in the following examples is always the character which immediately follows the first '*' in the morphological record. Also, before listing the possible analyses of each word, the number of the analyses is printed.

As an example, consider the morphological analysis of the first sentence of the above text:

```
2
&$R*M**LZRF*** #&$RWT (&$R, MSPR LA˜MIWD& ZKR RBIM NPRD )
&$R*M**LZRS*** #&$RWT- (&$R, MSPR LA˜MIWD& ZKR RBIM NSMK )
 2
AI$*E**LZRF*** AI$-IM (AI$, &CM LA˜MIWD& ZKR RBIM NPRD )
HN$IM*P*BY1T ANI-AN$IM (HN$IM, PW&L &TID ZW"N IXID MDBR )
 1
HGI&*P*ZRAH HM-&K$W-MGI&IM (HGI&, PW&L HWWH ZKR RBIM  )
 1
TAILND*p*mZ M-'TAILND' (TAILND, $M˜PR@I ZKR )
 1
I$RAL*p*lB L-'I$RAL' (I$RAL, $M˜PR@I ZW"N )
 1
HWA*gK*LZR3 K$-HMg (HWA, MILT˜GWP LA˜MIWD& ZKR RBIM NSTR )
 1
NR$M*P*ZRAH HM-&K$W-NR$MIM (NR$M, PW&L HWWH ZKR RBIM  )
 2
MTNDB*E*kKZRF*** KH-MTNDB-IM (MTNDB, &CM MIWD& ZKR RBIM NPRD )
MTNDB*E*kLZRF*** K-MTNDB-IM (MTNDB, &CM LA˜MIWD& ZKR RBIM NPRD )
 1
,*x* , (,, MILT˜XIBWR )
 1
AK*x* AKx (AK, MILT˜XIBWR )
4
LM&$H*t* LM&$Ht (LM&$H, TWAR˜PW&L )
M&$*E*lLZYFNY3 L-H-M&$-$LH (M&$, &CM LA˜MIWD& ZKR IXID NPRD  (SIOMT: NQBH IXID N
M&$H*E*lKZYF*** LH-M&$H (M&$H, &CM MIWD& ZKR IXID NPRD )
M&$H*E*lLZYF*** L-M&$H (M&$H, &CM LA˜MIWD& ZKR IXID NPRD )
 3
M$M$*E**LZRF*** M$M$-IM (M$M$, &CM LA˜MIWD& ZKR RBIM NPRD )
$IM$*P*ZRAH HM-&K$W-M$M$IM ($IM$, PW&L HWWH ZKR RBIM  )
$M$*E*mLZRF*** M-$M$-IM ($M$, &CM LA˜MIWD& ZKR RBIM NPRD )
 2
&WBD*E**LZRF*** &WBD-IM (&WBD, &CM LA˜MIWD& ZKR RBIM NPRD )
&BD*P*ZRAH HM-&K$W-&WBDIM (&BD, PW&L HWWH ZKR RBIM  )
 2
$KIR*E**LZRF*** $KIR-IM ($KIR, &CM LA˜MIWD& ZKR RBIM NPRD )
$KIR*T**LZRF %$KIRIM ($KIR, TWAR LA˜MIWD& ZKR RBIM NPRD )
 1
ZWL*T**LZRF %ZWLIM (ZWL, TWAR LA˜MIWD& ZKR RBIM NPRD )
 1
.*x* . (., MILT˜XIBWR )
```

Thus, the second word in the text, AN$IM, has two analyses, one a masculine plural noun (“men”) and the other a future first person verb (“I will resuscitate”). The analyses include the following information:

```
AI$*E**LZRF*** AI$-IM (AI$, &CM LA˜MIWD& ZKR RBIM NPRD )
```

First, the base AI$ is given. The E indicates that it is a noun; L stands for indefinite, Z for masculine, R for plural and F for the absolute (rather construct) state. This information is further printed in the parentheses.

```
HN$IM*P*BY1T ANI-AN$IM (HN$IM, PW&L &TID ZW"N IXID MDBR )
```

Here, the base is HN$IM, a verb (P). It is further specified as present tense (B), singular (Y), first person (1) and unspecified for gender (T).

A definition of the various part of speech tags follows:

E	Noun
P	Verb
T	Adjective
t	Adverb
y	Preposition
g	Pronoun
p	Proper name
z	Auxiliary verb
x	Conjunction
j	Question word
m	Other

A detailed description of the morphological analysis and the output format is available at:
http://www.cs.technion.ac.il/˜erelsgl/bxi/hmntx/tqstim/teud_nitux_curni.html

Finally, for evaluation purposes you will also obtain a file in which each word is associated with exactly one morphological record, from which the part of speech can be easily extracted. Your final project will aim to produce this output given the input delineated above. Thus, for the example sentence, the correct analysis is:

```
&$RWT           &$R*M**LZRS***
AN$IM           AI$*E**LZRF***
MGI&IM          HGI&*P*ZRAH
MTAILND         TAILND*p*mZ
LI$RAL          I$RAL*p*lB
K$HM            HWA*gK*LZR3
NR$MIM          NR$M*P*ZRAH
KMTNDBIM        MTNDB*E*kLZRF***
,               ,*x*
AK              AK*x*
LM&$H           LM&$H*t*
M$M$IM          $IM$*P*ZRAH
&WBDIM          &WBD*E**LZRF***
$KIRIM          $KIR*T**LZRF
ZWLIM           ZWL*T**LZRF
.               .*x*                .
```

Submission

Define a set of features which you believe can be instrumental for part of speech tagging in Hebrew. Observe that the problem is reminiscent of (but simpler than) full morphological disambiguation. Justify your choice of features and motivate them by examples of ambiguous words which can be disambiguated using the features you propose. There is no limit on the number of features (some learning algorithms perform perfectly well with thousands of features).

Deadline: next class (after Pessax).

Good luck!