Memory-Based Named Entity Recognition

Erik Tjong Kim Sang
CNTS - Language Technology Group
University of Antwerp
Belgium

Approach

We have used the memory-based learning algorithm IB1-IG, a nearest-neighbor classifier.

Tokens have been represented by a set of features from a window of surrounding words (3 left/3 right) and substrings of the current and previous word.

en	Argentina	,	Arg	ina	e	n	B-LOC
Argentina	,	jugó	–	–	–	–	O
,	jugó	con	–	–	–	–	O
jugó	con	Del	–	–	–	–	O
con	Del	Bosque	–	el	c	n	B-PER
Del	Bosque	en	B	sque	De	l	I-PER

All training data is stored and test data is classified by taking the class of the training data item that is closest to them in the feature space.

Creating word-internal features

Word-internal features were used in order to obtain extra clues about the classes of unknown words.

Prefix and suffix strings that were part of a words inside enties in 95% of the cases and occurred 10 times or more were selected as interesting features.

Similar prefix and suffix strings were selected for words appearing before a word inside a named entity.

With these four word-internal features, the error of the unlabelled entity recognition phase reduced with almost half.

Extra techniques used

The task was split in two parts: finding entity borders and classifying entities.

The following techniques were used for improving the performance of the base learning system:

- cascading (boosting)
- feature selection
- system combination

Development results

Train	Pass 1		Pass 2	
Repr.	$F_{\beta=1}$	features used	$F_{\beta=1}$	features used
IOB1	85.86	$w_{-2..0}$ $m_{fp,fs,ps}$	88.68	$w_{-2,0,1}$ $t_{-1,1}$ $m_{fp,fs,ps}$
IOB2	82.14	$w_{-2..1}$ $m_{fs,pp,ps}$	84.39	$w_{-1..1}$ $t_{-2,-1,1}$ $m_{pp,ps}$
IOE1	85.86	$w_{-2..0}$ $m_{fp,fs,ps}$	88.76	$w_{-2,0,1}$ $t_{-1,1}$ $m_{fp,fs,ps}$
IOE2	77.18	$w_{-2..2}$ $m_{fp,fs,pp}$	83.50	$w_{-1,0,2}$ $t_{-1,1}$ $m_{fs,pp}$
O+C	80.33	O: $w_{-2..0}$ $m_{fp,pp,ps}$ C: $w_{-2..1}$ m_{ps}	84.08	O: $w_{-2..0}$ $t_{-2,-1}$ $m_{fp,pp,ps}$ C: $w_{-1..2}$ $t_{1,2}$ m_{ps}
Voting	86.10	O: all C: all	88.96	O: all C: all
Classes	72.29	$w_{-2,-1,s,f}$ $m_{s.s}$	74.34	$w_{-2,-1,s,f}$ $m_{s.s}$

The overall system was tuned to the training data only.

All three extra processing techniques helped improve performance (table shows results for Spanish).

Results Spanish

Spanish test	precision	recall	$F_{\beta=1}$
LOC	76.01%	76.01%	76.01
MISC	63.70%	50.59%	56.39
ORG	76.45%	78.36%	77.39
PER	79.57%	81.09%	80.32
overall	76.00%	75.55%	75.78

- Reasonable results except for the MISC category.
- After removing some software bugs we managed an overall F rate of 76.69.

Results Dutch

Dutch test	precision	recall	$F_{\beta=1}$
LOC	83.21%	73.28%	77.93
MISC	72.79%	64.45%	68.36
ORG	75.63%	54.50%	63.35
PER	65.72%	81.97%	72.95
overall	72.56%	68.88%	70.67

- No part-of-speech tags were used.
- A recent test with POS tags did not lead to a significant improvement (F = 70.85).
- Tests with gazetteers did not lead to improvements.

Concluding remarks

- A memory-based learning system was applied to the CoNLL-2002 shared task.
- Extra processing techniques (cascading, feature selection and system combination) all helped to improve performance.
- Using word-internal features turned out to be very useful.
- Neither POS tag information nor gazetteers helped to improve performance (yet).