Memory-Based Named Entity Recognition

Erik Tjong Kim Sang
CNTS - Language Technology Group
University of Antwerp
Belgium
erikt@uia.ua.ac.be

Task description

Language-independent named entity recognition is the shared task of CoNLL-2002 (August/September 2002, Taiwan).

Named entity recognition is the process of finding names in text. Example:

[PER Wolff] , currently a journalist in
[LOC Argentina] , played with
[PER Del Bosque] in the final years
of the seventies in [ORG Real Madrid] .

The CoNLL-2002 data sets contain entities concerning persons, locations, organizations and miscellaneous names. There is data available for Spanish and Dutch.

Tjong Kim Sang 19/06/2002

Approach

We will use a memory-based learner (Timbl) for building a named entity recognizer.

In order to improve the recognizer, we will apply some other techniques: cascading, feature selection and system combination.

The basic features available to the learner are the focus word and words from its immediate context. Additionally we will examine the usability of morphological features (character prefixes and suffixes) that are relevant for this particular task.

Data format (1)

For both target languages, three data sets are available: training data, development data and test data.

All data files contain texts with one word per line. A tag at the end of each line defines whether the word is part of a named entity or not.

in	O
Argentina	B-LOC
,	O
played	O
with	O
Del	B-PER
Bosque	I-PER

In the files, sentence breaks are encoded by empty lines.

Data format (2)

The training, development and test data sets for Spanish contain about 270,000, 50,000 and 50,000 words respectively. For Dutch these numbers are 220,000, 40,000 and 70,000.

The Dutch files had part-of-speech tag information available that was generated by a POS tagger (MBT). However, we have not used it.

The Spanish data consists of news wire texts from the Spanish EFE News Agency from May 2000. The Dutch text contains articles from the Belgian newspaper De Morgen (June-September 2000).

Cascading

A standard approach to classifier task like these is to use words and their context as features:

Del Bosque in $\rightarrow$ I-PER

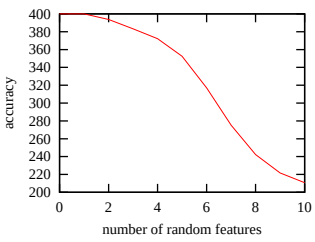
It would be interesting to know if the words in the context could be part of a named entity as well. For this purpose we performing cascading: processing the output of one classifier with another:

Del Bosque in B-PER O $\rightarrow$ I-PER

The classification for the previous and the next word have been computed by another classifier. Note that we do not use the classification of the focus word.

Feature selection (1)

The memory-based classifier is not very good with dealing with a large group of irrelevant features. Its performance on an easy binary classification task (XOR) dropped when random features were added to the data:



After adding ten random features the performance is not much better than guessing (211/400).

Feature selection (2)

The solution for this is feature selection: evaluate the performance of the system for each subset of features.

For this particular problem there are too many feature subsets to perform a complete search. Therefore we use an heuristic search algorithm to find the best feature sub set: bidirectional hill-climbing.

It starts from a particular subset of features (for example the empty set) and evaluates all subsets that have an extra feature or one feature less. Then it selects the best feature subset and repeats the search until no improvement can be made.

System combination (1)

Suppose we use five learning algorithms for predicting whether a named entity starts at a certain position or not.

	c_1	c_2	c_3	c_4	c_5	correct
$word_1$	.	.	.	.	.	.
$word_2$	[	[	[	[	[	[
$word_3$	.	.	.	.	.	.
$word_4$	[	.	[	[	[	[
$word_5$	.	.	[	.	.	.
$word_6$	[	[	[	[	.	[
$word_7$	[	.	.	.	.	.
$word_8$	[	[	[	.	[	[

We can combine the results with majority voting: choose the result that has been predicted most often.

System combination (2)

We generate different results for the named entity task by using different representations of the output tags. Here is example for noun phrase chunking:

	IOB1	IOB2	IOE1	IOE2	O	C
In	O	O	O	O	.	.
early	I	B	I	I	[	.
trading	I	I	I	E	.	]
in	O	O	O	O	.	.
Hong	I	B	I	I	[	.
Kong	I	I	E	E	.	]
Monday	B	B	I	E	[	]

The learner will make different errors while using a different representation. With majority voting we attempt to remove some of these errors.

Morphological features

These features contain word-interal information: prefixes and suffixes. Not all of them are interesting for this task. We have selected the following from the training data:

- Prefixes and suffixes of capitalized words that occur 10 times or more and are part of a named entity in 95% or more of the cases.
- Prefixes and suffixes of words before a capitalized word that occur 10 times or more while the next word is part of a named entity in 95% or more of the cases.

Some examples from the Dutch data: Q- (Qatar, Quattro, Quentin), -ert (Albert, Gert, Jalabert, Stevaert), a- (aandeel, advocaat, aldus, avonturier), and -r (broer, CVP'er, professor, Walter).

Dutch	Pass 1	Pass 2	
Repr.	$F_{\beta=1}$	$F_{\beta=1}$	features used
IOB1	86.46	86.83	$w_{-1..2}$ t_1 $m_{fs,pp,ps}$
IOB2	84.93	86.44	$w_{-2..1}$ $t_{-2,-1,1}$ $m_{fs,pp,ps}$
IOE1	86.68	87.13	$w_{-1..2}$ t_{-1} $m_{fp,fs,ps}$
IOE2	77.30	85.91	$w_{-1..2}$ $t_{1,2}$ $m_{fp,fs,pp}$
O+C	78.49	83.57	O: $w_{-2..2}$ $t_{-2,-1}$ $m_{fp,pp,ps}$ C: $w_{-1..1}$ $t_{-2,1}$ m_{ps}
Majority	86.84	88.94	O: IOB1 IOB2 IOE1 IOE2 O C: IOB2 IOE1 IOE2
classes	72.29	74.34	$w_{-2,-1,\mathrm{start,final}}$

Table 1: Results of a 10-fold cross-validation experiment on the Dutch training data. All results except those on the final line deal with entity recognition without classification.

Test set results

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

Dutch test	precision	recall	$F_{\beta=1}$
LOC	81.66%	75.10%	78.24
MISC	74.10%	60.49%	66.60
ORG	70.62%	57.01%	63.09
PER	66.92%	79.53%	72.68
overall	72.34%	67.91%	70.05

The best CoNLL-2002 system obtains more than $F_{\beta=1}$=80 for Spanish and about $F_{\beta=1}$=75 for Dutch.

Techniques that did not work (yet)

Adding binary features like IsACapitalizedWord and OccursInDictionary did not improve performance.

Generating the morphological features from raw text did not work. Here we selected candidate named entities with character strings which occurred in or around named entities in 100% of the cases.

Concluding remarks

We have presented a method for finding named entities in text.

Apart from memory-based learning, the method uses cascading, feature selection and system combination for improving performance.

The system uses morphological features which have been derived from the training data with statistical measures.

The results obtained on the two CoNLL-2002 data sets are reasonable.