Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition

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

Motivation

The last three years we have studied shallow parsing tasks applied to English (NP detection, chunking and clause identification).

Named entity recognition has received a lot of attention when applied to English but less for other languages.

Named entity recognition is a relatively open task which might require a lot of examples.

The CoNLL-2002 shared task consists of language-independent named entity recognition.

Task description

The shared task involves finding names in text. There are four categories: persons, organizations, locations and miscellaneous names. Example:

> [$_{\mathrm{PER}}$ Wolff] , ahora periodista en [$_{\mathrm{LOC}}$ Argentina] , jugó con [$_{\mathrm{PER}}$ Del Bosque] en los últimos años de los 70 en el [$_{\mathrm{ORG}}$ Real Madrid] .

Data was available for two Western European languages: Spanish and Dutch.

Participating systems were expected to contain a machine learning component.

Using additional resources, especially nonannotated text, was encouraged.

Data

- For each language there are three data files. One training file, one for testing systems during the development stage and one file for final tests.
- Data files consisted of two columns (Spanish: words and named entity tags) or three columns (Dutch: words, estimated part-of-speech tags and named entity tags).
- The Spanish data was made available by the EFE News Agency (Spain) and annotated by people from the Technical University of Catalonia and the University of Barcelona.
- The Dutch data was supplied by the newspaper De Morgen (Belgium) and annotated by people from the University of Antwerp.

Data example

Wolff	B-PER
,	O
ahora	O
periodista	O
en	O
Argentina	B-LOC
,	O
jugó	O
con	O
Del	B-PER
Bosque	I-PER
en	O
el	O
Real	B-ORG
Madrid	I-ORG
.	O

Evaluation

We register the number of completely correct phrases and compute precision, recall and $F_{\beta=1}$ rates:

Precision: number of correct phrases divided by the number of phrases found by the algorithm.

Recall: number of correct phrases divided by the number of phrases in the corpus.

$F_\beta = (\beta^2+1)$*precision*recall / β^2*precision+recall
We use $\beta = 1$.

Baseline performances have been obtained with an algorithm which only selects complete named entities which appear in the training data.

Participants

Twelve groups have participated in the CoNLL-2002 shared task. They have used boosted trees, decision trees, Hidden Markov Models, maximum entropy models, memory-based methods, statistical techniques, support vector machines and transformation-based methods.

- McNamee and Mayfield
- Black and Vasilakopoulos
- Tsukamoto, Mitsuishi and Sassano
- Jansche
- Malouf
- Patrick, Whitelaw and Munro
- Tjong Kim Sang
- Burger, Henderson and Morgan
- Cucerzan and Yarowsky
- Wu, Ngai, Carpuat, Larsen and Yang
- Florian
- Carreras, Màrques and Padró

Significance values

After the final version of the papers had been submitted, significance values have been computed for all system results.

These values have been estimated by using bootstrap resampling (Noreen, *Computer-Intensive Methods for Testing Hypotheses*, 1989).

For each output file, 1000 samples of approximately the same size have been created by randomly selecting sentences with replacement.

Results with $F_{\beta=1}$ rates outside of the center 90% of the sample group have been regarded as significantly different from the output results.

Results Spanish test data

Spanish test	precision	recall	$F_{\beta=1}$	
Carreras et al.	81.38%	81.40%	81.39	±1.5
Florian	78.70%	79.40%	79.05	±1.4
Cucerzan et al.	78.19%	76.14%	77.15	±1.4
Wu et al.	75.85%	77.38%	76.61	±1.4
Burger et al.	74.19%	77.44%	75.78	±1.4
Tjong Kim Sang	76.00%	75.55%	75.78	±1.5
Patrick et.al.	74.32%	73.52%	73.92	±1.5
Jansche	74.03%	73.76%	73.89	±1.5
Malouf	73.93%	73.39%	73.66	±1.6
Tsukamoto	69.04%	74.12%	71.49	±1.4
Black et al.*	60.53%	67.29%	63.73	±1.8
McNamee et al.	56.28%	66.51%	60.97	±1.7
baseline	26.27%	56.48%	35.86	±1.3

* results differ from those mentioned in the proceedings

Results Dutch test data

Dutch test	precision	recall	$F_{\beta=1}$	
Carreras et al.	77.83%	76.29%	77.05	±1.5
Wu et al.	76.95%	73.83%	75.36	±1.6
Florian	75.10%	74.89%	74.99	±1.5
Burger et al.	72.69%	72.45%	72.57	±1.4
Cucerzan et al.	73.03%	71.62%	72.31	±1.6
Patrick et al.	74.01%	68.90%	71.36	±1.6
Tjong Kim Sang	72.56%	68.88%	70.67	±1.6
Jansche	70.11%	69.26%	69.68	±1.7
Malouf	70.88%	65.50%	68.08	±1.9
Tsukamoto	57.33%	65.02%	60.93	±1.7
McNamee et al.	56.22%	63.24%	59.52	±2.0
Black et al.*	51.89%	47.78%	49.75	±2.2
baseline	64.38%	45.19%	53.10	±1.4

* results differ from those mentioned in the proceedings

Combining systems: method

The output of the twelve systems was combined by using majority voting.

Entities were split in starts and ends.

Only entity borders predicted by the majority of the systems were accepted.

Only consistent start and end pairs were kept:

this: [-LOC [-PER]-ORG [-MISC]-MISC]-PER

would be converted to: [-MISC]-MISC

Combining systems: results

All combinations of the twelve systems were tested on the development data and the best was applied to the test data.

Spanish

development		test		systems
82.40	-18%	83.01	-9%	1,2,3,4,7,12
81.55	-14%	81.25	0%	1-12 (all)
78.47		81.39		1 (best)

Dutch

development		test		systems
79.39	-15%	80.25	-14%	1,3,4,5,8
76.33	-3%	79.03	-9%	1-12 (all)
75.66		77.05		1 (best)

System numbers refer to their position in the result tables.

Useful techniques

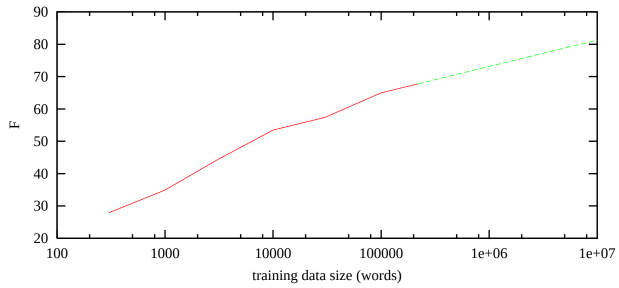
	1	2	3	4	5	6	7	8	9	10	11	12
OFE	+	+	+	+	+	+	+	+	+	+	+	+
BOO	+	+	+	+	+	+	+	−	−	+	+	−
NEC	+	+	+	−	−	+	+	+	−	−	+	+
GAZ	+	−	−	+	*	−	−	−	*	−	−	+
POS	+	−	+	+	−	−	−	−	−	−	−	−
UNA	−	−	−	−	+	−	−	−	−	−	−	−

OFE: orthographic features
BOO: boosting
NEC: separate entity border/class identification
GAZ: extra external lists of entities
POS: part-of-speech tags
UNA: unannotated data

+: used; −: did not use; *: used without effect

Numbers refer to positions in the Spanish result table.

Will more data help?



The learning curve for the Dutch data for one NER system (Tjong Kim Sang) applied to the Dutch development data seems to increase according to a straight line.

It suggests that a third of the errors of the system could be removed if we had 10 million words of training data.

Problematic sentences

Some words have been classified incorrectly by all systems (0.7% of the words in the Dutch development data):

- unknown words (*Israël*)
- booktitles (*Yo soy el Diego*)
- ambiguous entities (*Camp David*)
- neighboring entities (*Londen Swissair*)
- conjunctions (*AMP België of Zwitserland*)
- words without initial capital character (*sefardische*)
- long entities (*Centrum voor Onderzoek in de Diergeneeskunde en de Agrochemie*)

About the baseline results

The baseline system has been selected with a high precision as goal.

Yet, baseline precision is lower than that most of the systems.

The problem is that the data contained annotation errors that caused problems for the baseline system:

Spanish		Dutch		
españolas	O	vandaag	Adv	O
.	B-LOC	met	Prep	B-PER
		Frans	N	B-MISC
San	I-LOC	leven	N	B-PER
Sebastián	I-LOC	een	Art	O

Correcting the data

The three Dutch files have been checked, and inconsistencies and errors that were found have been removed.

About 300 words (0.1%) have received a different entity tag.

Dutch test	precision	recall	$F_{\beta=1}$	
Tjong Kim Sang	72.56%	68.88%	70.67	± 1.6
baseline	64.38%	45.19%	53.10	± 1.4

After correction	precision	recall	$F_{\beta=1}$	
Tjong Kim Sang	72.74%	69.55%	71.11	± 1.6
baseline	81.29%	45.42%	58.28	± 1.4

The Dutch data at the shared task web site has been replaced with the the corrected data.

Concluding remarks

- The ConLL-2002 shared task involved language-independent named entity recognition using two Western European languages.
- For both Spanish and Dutch the best results have been obtained with the boosted decision trees method used by Xavier Carreras, Lluís Màrques and Lluís Padró of the Universitat Politécnica de Catalunya (Spain).