Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition

Erik F. Tjong Kim Sang and Fien De Meulder
CNTS - Language Technology Group
University of Antwerp

Motivation

The CoNLL-2002 shared task dealt with language-independent named entity recognition (Spanish and Dutch).

Few participating systems made use of unannotated data despite an expressed interest by the shared task organizers.

The CoNLL-2003 shared task deals with language-independent named entity recognition as well (English and German).

This time the organizers have provided unannotated data.

Task description

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

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

Data was available for two Western European languages: English and German.

Data

- Three data files are available for each language: a training file, a file for testing systems during the development stage and a file for final tests.
- Data files consist of four or five columns: words, estimated lemmas (German only), estimated part-of-speech tags, estimated chunk tags and named entity tags.
- The English data comes from the Reuters Corpus and the German data has been taken from the ECI Multilingual Text cd. Data files have been annotated by people from the University of Antwerp.

Data example

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Lines contain four fields: the word, its part-of-speech tag, its chunk tag and its named entity tag. Words tagged with O are outside of named entities/chunks and words with I-XXX are inside a named entity/chunk of type XXX.

Evaluation

We register the number of 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_β = (β^2+1)*precision*recall / β^2*precision+recall (we use β = 1)

Significance values

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

For each output file, 250 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 this output file.

Baseline system

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

Longer phrases are preferred over shorter ones.

Phrases with more than one entity tag are discarded.

Participants

Sixteen systems have participated in the shared task:

Bender, Och and Ney; Carreras, Màrquez and Padró (two systems); Chieu and Ng (1); Curran and Clark (2); De Meulder and Daelemans; Florian, Ittycheriah, Jing and Zhang (6); Hammerton; Hendrickx and Van den Bosch; Klein, Smarr, Nguyen and Manning (4); Mayfield, McNamee and Piatko (3); McCallum and Li; Munro, Ler and Patrick; Whitelaw and Patrick; Wu, Ngai, and Carpuat; and Zhang and Johnson (5).

The five best performing systems for the English development data and the best of the rest for the German test data will be presented here.

Results English data

	$F_{\beta=1}$
Florian	88.8$\pm$0.7
Chieu	88.3$\pm$0.7
Klein	86.1$\pm$0.8
Zhang	85.5$\pm$0.9
Carreras (b)	85.0$\pm$0.8
Curran	84.9$\pm$0.9
Mayfield	84.7$\pm$1.0
Carreras (a)	84.3$\pm$0.9

	$F_{\beta=1}$
McCallum	84.0$\pm$0.9
Bender	83.9$\pm$1.0
Munro	82.5$\pm$1.0
Wu*	82.7$\pm$0.9
Whitelaw	79.8$\pm$1.0
Hendrickx	78.2$\pm$1.0
De Meulder	77.0$\pm$1.2
Hammerton	60.2$\pm$1.3
Baseline	59.6$\pm$1.2

Results German data

	$F_{\beta=1}$
Florian	72.4$\pm$1.3
Klein	71.9$\pm$1.2
Zhang	71.3$\pm$1.5
Mayfield	70.0$\pm$1.4
Carreras (b)	69.2$\pm$1.3
Bender	68.9$\pm$1.3
Curran	68.4$\pm$1.4
McCallum	68.1$\pm$1.4

	$F_{\beta=1}$
Munro	67.8$\pm$1.4
Carreras (a)	66.5$\pm$1.5
Wu	66.3$\pm$1.3
Chieu	65.7$\pm$1.4
Hendrickx	63.0$\pm$1.4
De Meulder	57.3$\pm$1.6
Whitelaw	54.4$\pm$1.4
Hammerton	47.7$\pm$1.5
Baseline	30.3$\pm$1.3

Techniques used

order English:	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6
AdaBoost								x				x				
Conditional Random Fields									x							
Hidden Markov Models	x		x				x						x			
Maximum Entropy Models	x	x	x			x				x						
Memory-Based Learning														x	x	
Recurrent Neural Networks																x
Robust Risk Minimization	x			x												
Support Vector Machines							x									
System Combination	x		x				x				x	x				
Transformation-Based Learning	x															
Voted Perceptrons					x											

Features used

order English:	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6
lexical information	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x
POS tags	x	x	x	x	x	x	x	x		x	x	x		x	x	x
affix information	x	x	x	x	x	x	x	x			x	x	x	x	x	
previous NE tags	x	x	x	x	x	x	x	x		x		x	x	x		
orthographic information	x	x		x	x	x	x	x	x	x		x		x	x	
gazetteers	x	x		x	x	x			x	x		x		x	x	x
chunk tags	x			x	x		x			x	x			x	x	x
orthographic patterns					x	x	x	x	x							
global case information	x					x					x		x		x	
trigger words		x		x	x						x					
bag of words					x						x					
quote information		x					x									
global document information		x														

External resources

	Gaz.	Una.	Ext.	English	German
Zhang	+	-	-	19%	15%
Florian	+	-	+	27%	5%
Hammerton	+	-	-	22%	-
Carreras (b)	+	-	-	12%	8%
Chieu	+	-	-	17%	-
Hendrickx	+	+	-	7%	5%
De Meulder	+	+	-	8%	3%
Bender	+	+	-	3%	6%
Curran	+	-	-	1%	-
McCallum	+	+	-	?	?
Wu	+	-	-	?	?

Combining systems: method

We performed a majority vote on the output tags of subsets of the systems.

The best subset for each data set was obtained with a bidirectional feature search starting from zero systems.

The performance of the majority vote was evaluated on the development data.

Combining systems: results

English

development		test		systems
94.53	-11%	90.30	-14%	1,2,3,9,13
93.87		88.76		1 (best)

German

development		test		systems
74.75	-11%	74.17	-6%	4,5,7,9,11
71.51		72.41		1 (best)

System numbers refer to positions in the English result table.

Problematic entities

Some words were classified incorrectly by all systems: 96 in the English development data and 610 in the German data. Some examples:

Long phrases: Solidarity Meeting for Sarajevo
Split phrases: Berlin BREAK Grand Prix
Included years/numbers: 1997 Fed Cup / Interstate 5
Brackets: (Jerry) Koosman
Titles of books/movies: In the Year of January / Michael Collins
Infrequent words: Jebel al-Akhdar
Mistagged words: Boxing-Bruno

Concluding remarks

- The CoNLL-2003 shared task involved language-independent named entity recognition.
- For both languages that were examined, the best results were obtained by a combined classifier presented in a paper by Florian, Ittycheriah, Jing and Zhang.
- The majority of the 16 participants has tried using some kind of extra information (gazetteers/unannotated data). However an excellent method for utilizing these resources remains to be found.

Discussion points

- What can we offer people who have a lot of raw data for language X and want to build a named entity recognition system for this language?
- What learning system should they use?
- What features should they employ?
- In what way can useful information be extracted from the data?