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
{erikt,fien.demeulder}@uia.ua.ac.be

Abstract

We describe the CoNLL-2003 shared task: language-independent named entity recognition. We give background information on the data sets (English and German) and the evaluation method, present a general overview of the systems that have taken part in the task and discuss their performance.

1 Introduction

Named entities are phrases that contain the names of persons, organizations and locations. Example:

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

This sentence contains three named entities: *Ekeus* is a person, *U.N.* is a organization and *Baghdad* is a location. Named entity recognition is an important task of information extraction systems. There has been a lot of work on named entity recognition, especially for English (see Borthwick (1999) for an overview). The Message Understanding Conferences (MUC) have offered developers the opportunity to evaluate systems for English on the same data in a competition. They have also produced a scheme for entity annotation (Chinchor et al., 1999). More recently, there have been other system development competitions which dealt with different languages (IREX and CoNLL-2002).

The shared task of CoNLL-2003 concerns language-independent named entity recognition. We will concentrate on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups. The shared task of CoNLL-2002 dealt with named entity recognition for Spanish and Dutch (Tjong Kim Sang, 2002). The participants of the 2003 shared task have been offered training and test data for two other European languages: English and German. They have used the data for developing a named-entity recognition system that includes a machine learning component. The shared task organizers were especially interested in approaches that made use of resources other than the supplied training data, for example gazetteers and unannotated data.

2 Data and Evaluation

In this section we discuss the sources of the data that were used in this shared task, the preprocessing steps we have performed on the data, the format of the data and the method that was used for evaluating the participating systems.

2.1 Data

The CoNLL-2003 named entity data consists of eight files covering two languages: English and German[1]. For each of the languages there is a training file, a development file, a test file and a large file with unannotated data. The learning methods were trained with the training data. The development data could be used for tuning the parameters of the learning methods. The challenge of this year's shared task was to incorporate the unannotated data in the learning process in one way or another. When the best parameters were found, the method could be trained on the training data and tested on the test data. The results of the different learning methods on the test sets are compared in the evaluation of the shared task. The split between development data and test data was chosen to avoid systems being tuned to the test data.

The English data was taken from the Reuters Corpus[2]. This corpus consists of Reuters news stories

[1] Data files (except the words) can be found on http://lcg-www.uia.ac.be/conll2003/ner/

[2] http://www.reuters.com/researchandstandards/

English data	Articles	Sentences	Tokens
Training set	946	14,987	203,621
Development set	216	3,466	51,362
Test set	231	3,684	46,435

German data	Articles	Sentences	Tokens
Training set	553	12,705	206,931
Development set	201	3,068	51,444
Test set	155	3,160	51,943

Table 1: Number of articles, sentences and tokens in each data file.

English data	LOC	MISC	ORG	PER
Training set	7140	3438	6321	6600
Development set	1837	922	1341	1842
Test set	1668	702	1661	1617

German data	LOC	MISC	ORG	PER
Training set	4363	2288	2427	2773
Development set	1181	1010	1241	1401
Test set	1035	670	773	1195

Table 2: Number of named entities per data file

between August 1996 and August 1997. For the training and development set, ten days' worth of data were taken from the files representing the end of August 1996. For the test set, the texts were from December 1996. The preprocessed raw data covers the month of September 1996.

The text for the German data was taken from the ECI Multilingual Text Corpus[3]. This corpus consists of texts in many languages. The portion of data that was used for this task, was extracted from the German newspaper Frankfurter Rundshau. All three of the training, development and test sets were taken from articles written in one week at the end of August 1992. The raw data were taken from the months of September to December 1992.

Table 1 contains an overview of the sizes of the data files. The unannotated data contain 17 million tokens (English) and 14 million tokens (German).

2.2 Data preprocessing

The participants were given access to the corpus after some linguistic preprocessing had been done: for all data, a tokenizer, part-of-speech tagger, and a chunker were applied to the raw data. We created two basic language-specific tokenizers for this shared task. The English data was tagged and chunked by the memory-based MBT tagger (Daelemans et al., 2002). The German data was lemmatized, tagged and chunked by the decision tree tagger Treetagger (Schmid, 1995).

Named entity tagging of English and German training, development, and test data, was done by hand at the University of Antwerp. Mostly, MUC conventions were followed (Chinchor et al., 1999). An extra named entity category called MISC was added to denote all names which are not already in the other categories. This includes adjectives, like *Italian*, and events, like *1000 Lakes Rally*, making it a very diverse category.

[3]http://www.ldc.upenn.edu/

2.3 Data format

All data files contain one word per line with empty lines representing sentence boundaries. At the end of each line there is a tag which states whether the current word is inside a named entity or not. The tag also encodes the type of named entity. Here is an example sentence:

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

Each line contains 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 and the I-XXX tag is used for words inside a named entity of type XXX. Whenever two entities of type XXX are immediately next to each other, the first word of the second entity will be tagged B-XXX in order to show that it starts another entity. The data contains entities of four types: persons (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC). This tagging scheme is the IOB scheme originally put forward by Ramshaw and Marcus (1995). We assume that named entities are non-recursive and non-overlapping. When a named entity is embedded in another named entity, usually only the top level entity has been annotated.

Table 2 contains an overview of the number of named entities in each data file.

2.4 Evaluation

The performance in this task is measured with $F_{\beta=1}$ rate:

$$F_\beta = \frac{(\beta^2 + 1) * precision * recall}{(\beta^2 * precision + recall)} \quad (1)$$

	lex	pos	aff	pre	ort	gaz	chu	pat	cas	tri	bag	quo	doc
Florian	+	+	+	+	+	+	+	-	+	-	-	-	-
Chieu	+	+	+	+	+	+	-	-	-	+	-	+	+
Klein	+	+	+	+	-	-	-	-	-	-	-	-	-
Zhang	+	+	+	+	+	+	+	-	-	+	-	-	-
Carreras (a)	+	+	+	+	+	+	+	+	-	+	+	-	-
Curran	+	+	+	+	+	+	-	+	+	-	-	-	-
Mayfield	+	+	+	+	+	-	+	+	-	-	-	+	-
Carreras (b)	+	+	+	+	+	-	-	+	-	-	-	-	-
McCallum	+	-	-	-	+	+	-	+	-	-	-	-	-
Bender	+	+	-	+	+	+	+	-	-	-	-	-	-
Munro	+	+	+	-	-	-	+	-	+	+	+	-	-
Wu	+	+	+	+	+	+	-	-	-	-	-	-	-
Whitelaw	-	-	+	+	-	-	-	-	+	-	-	-	-
Hendrickx	+	+	+	+	+	+	+	-	-	-	-	-	-
De Meulder	+	+	+	-	+	+	+	-	+	-	-	-	-
Hammerton	+	+	-	-	-	+	+	-	-	-	-	-	-

Table 3: Main features used by the the sixteen systems that participated in the CoNLL-2003 shared task sorted by performance on the English test data. Aff: affix information (n-grams); bag: bag of words; cas: global case information; chu: chunk tags; doc: global document information; gaz: gazetteers; lex: lexical features; ort: orthographic information; pat: orthographic patterns (like Aa0); pos: part-of-speech tags; pre: previously predicted NE tags; quo: flag signing that the word is between quotes; tri: trigger words.

with β=1 (Van Rijsbergen, 1975). Precision is the percentage of named entities found by the learning system that are correct. Recall is the percentage of named entities present in the corpus that are found by the system. A named entity is correct only if it is an exact match of the corresponding entity in the data file.

3 Participating Systems

Sixteen systems have participated in the CoNLL-2003 shared task. They employed a wide variety of machine learning techniques as well as system combination. Most of the participants have attempted to use information other than the available training data. This information included gazetteers and unannotated data, and there was one participant who used the output of externally trained named entity recognition systems.

3.1 Learning techniques

The most frequently applied technique in the CoNLL-2003 shared task is the Maximum Entropy Model. Five systems used this statistical learning method. Three systems used Maximum Entropy Models in isolation (Bender et al., 2003; Chieu and Ng, 2003; Curran and Clark, 2003). Two more systems used them in combination with other techniques (Florian et al., 2003; Klein et al., 2003). Maximum Entropy Models seem to be a good choice for this kind of task: the top three results for English and the top two results for German were obtained by participants who employed them in one way or another.

Hidden Markov Models were employed by four of the systems that took part in the shared task (Florian et al., 2003; Klein et al., 2003; Mayfield et al., 2003; Whitelaw and Patrick, 2003). However, they were always used in combination with other learning techniques. Klein et al. (2003) also applied the related Conditional Markov Models for combining classifiers.

Learning methods that were based on connectionist approaches were applied by four systems. Zhang and Johnson (2003) used robust risk minimization, which is a Winnow technique. Florian et al. (2003) employed the same technique in a combination of learners. Voted perceptrons were applied to the shared task data by Carreras et al. (2003a) and Hammerton used a recurrent neural network (Long Short-Term Memory) for finding named entities.

Other learning approaches were employed less frequently. Two teams used AdaBoost.MH (Carreras et al., 2003b; Wu et al., 2003) and two other groups employed memory-based learning (De Meulder and Daelemans, 2003; Hendrickx and Van den Bosch, 2003). Transformation-based learning (Florian et al., 2003), Support Vector Machines (Mayfield et al., 2003) and Conditional Random Fields (McCallum and Li, 2003) were applied by one system each.

Combination of different learning systems has proven to be a good method for obtaining excellent results. Five participating groups have applied system combination. Florian et al. (2003) tested different methods for combining the results of four systems and found that robust risk minimization worked best. Klein et al. (2003) employed a stacked learning system which contains Hidden Markov Models, Maximum Entropy Models and Conditional Markov Models. Mayfield et al. (2003) stacked two learners and obtained better performance. Wu et al. (2003) applied both stacking and voting to three learners. Munro et al. (2003) employed both voting and bagging for combining classifiers.

3.2 Features

The choice of the learning approach is important for obtaining a good system for recognizing named entities. However, in the CoNLL-2002 shared task we found out that choice of features is at least as important. An overview of some of the types of features chosen by the shared task participants, can be found in Table 3.

All participants used lexical features (words) except for Whitelaw and Patrick (2003) who implemented a character-based method. Most of the systems employed part-of-speech tags and two of them have recomputed the English tags with better taggers (Hendrickx and Van den Bosch, 2003; Wu et al., 2003). Othographic information, affixes, gazetteers and chunk information were also incorporated in most systems although one group reports that the available chunking information did not help (Wu et al., 2003) Other features were used less frequently. Table 3 does not reveal a single feature that would be ideal for named entity recognition.

3.3 External resources

Eleven of the sixteen participating teams have attempted to use information other than the training data that was supplied for this shared task. All included gazetteers in their systems. Four groups examined the usability of unannotated data, either for extracting training instances (Bender et al., 2003; Hendrickx and Van den Bosch, 2003) or obtaining extra named entities for gazetteers (De Meulder and Daelemans, 2003; McCallum and Li, 2003). A reasonable number of groups have also employed unannotated data for obtaining capitalization features for words. One participating team has used externally trained named entity recognition systems for English as a part in a combined system (Florian et al., 2003).

Table 4 shows the error reduction of the systems with extra information compared to while using only the available training data. The inclusion of extra named entity recognition systems seems to have worked well (Florian et al., 2003). Generally the systems that only used gazetteers seem to gain more than systems that have used unannotated data for other purposes than obtaining capitalization information. However, the gain differences between the two approaches are most obvious for English, for which better gazetteers are available. With the exception of the result of Zhang and Johnson (2003), there is not much difference in the German results between the gains obtained by using gazetteers and those obtained by using unannotated data.

	G	U	E	English	German
Zhang	+	-	-	19%	15%
Florian	+	-	+	27%	5%
Hammerton	+	-	-	22%	-
Carreras (a)	+	-	-	12%	8%
Chieu	+	-	-	17%	-
Hendrickx	+	+	-	7%	5%
De Meulder	+	+	-	8%	3%
Bender	+	+	-	3%	6%
Curran	+	-	-	1%	-
McCallum	+	+	-	?	?
Wu	+	-	-	?	?

Table 4: Error reduction for the two development data sets when using extra information like gazetteers (G), unannotated data (U) or externally developed named entity recognizers (E). The lines have been sorted by the sum of the reduction percentages for the two languages.

3.4 Performances

A baseline rate was computed for the English and the German test sets. It was produced by a system which only identified entities which had a unique class in the training data. If a phrase was part of more than one entity, the system would select the longest one. All systems that participated in the shared task have outperformed the baseline system.

For all the $F_{\beta=1}$ rates we have estimated significance boundaries by using bootstrap resampling (Noreen, 1989). From each output file of a system, 250 random samples of sentences have been chosen and the distribution of the $F_{\beta=1}$ rates in these samples is assumed to be the distribution of the performance of the system. We assume that performance A is significantly different from performance B if A is not within the center 90% of the distribution of B.

The performances of the sixteen systems on the two test data sets can be found in Table 5. For English, the combined classifier of Florian et al. (2003) achieved the highest overall $F_{\beta=1}$ rate. However, the difference between their performance and that of the Maximum Entropy approach of Chieu and Ng (2003) is not significant. An important feature of the best system that other participants did not use, was the inclusion of the output of two externally trained named entity recognizers in the combination process. Florian et al. (2003) have also obtained the highest $F_{\beta=1}$ rate for the German data. Here there is no significant difference between them and the systems of Klein et al. (2003) and Zhang and Johnson (2003).

We have combined the results of the sixteen system in order to see if there was room for improvement. We converted the output of the systems to the same IOB tagging representation and searched for the set of systems from which the best tags for the development data could be obtained with majority voting. The optimal set of systems was determined by performing a bidirectional hill-climbing search (Caruana and Freitag, 1994) with beam size 9, starting from zero features. A majority vote of five systems (Chieu and Ng, 2003; Florian et al., 2003; Klein et al., 2003; McCallum and Li, 2003; Whitelaw and Patrick, 2003) performed best on the English development data. Another combination of five systems (Carreras et al., 2003b; Mayfield et al., 2003; McCallum and Li, 2003; Munro et al., 2003; Zhang and Johnson, 2003) obtained the best result for the German development data. We have performed a majority vote with these sets of systems on the related test sets and obtained $F_{\beta=1}$ rates of 90.30 for English (14% error reduction compared with the best system) and 74.17 for German (6% error reduction).

4 Concluding Remarks

We have described the CoNLL-2003 shared task: language-independent named entity recognition. Sixteen systems have processed English and German named entity data. The best performance for both languages has been obtained by a combined learning system that used Maximum Entropy Models, transformation-based learning, Hidden Markov Models as well as robust risk minimization (Florian et al., 2003). Apart from the training data, this system also employed gazetteers and the output of two externally trained named entity recognizers. The performance of the system of Chieu et al. (2003) was not significantly different from the best performance for English and the method of Klein et al. (2003) and the approach of Zhang and Johnson (2003) were not significantly worse than the best result for German.

Eleven teams have incorporated information other than the training data in their system. Four of them have obtained error reductions of 15% or more for English and one has managed this for German. The resources used by these systems, gazetteers and externally trained named entity systems, still require a lot of manual work. Systems that employed unannotated data, obtained performance gains around 5%. The search for an excellent method for taking advantage of the vast amount of available raw text, remains open.

Acknowledgements

Tjong Kim Sang is financed by IWT STWW as a researcher in the ATraNoS project. De Meulder is supported by a BOF grant supplied by the University of Antwerp.

References

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. Maximum Entropy Models for Named Entity Recognition. In *Proceedings of CoNLL-2003*.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, New York University.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003a. Learning a Perceptron-Based Named Entity Chunker via Online Recognition Feedback. In *Proceedings of CoNLL-2003*.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003b. A Simple Named Entity Extractor using AdaBoost. In *Proceedings of CoNLL-2003*.

Rich Caruana and Dayne Freitag. 1994. Greedy Attribute Selection. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 28–36. New Brunswick, NJ, USA, Morgan Kaufman.

Hai Leong Chieu and Hwee Tou Ng. 2003. Named Entity Recognition with a Maximum Entropy Approach. In *Proceedings of CoNLL-2003*.

Nancy Chinchor, Erica Brown, Lisa Ferro, and Patty Robinson. 1999. *1999 Named Entity Recognition Task Definition*. MITRE and SAIC.

James R. Curran and Stephen Clark. 2003. Language Independent NER using a Maximum Entropy Tagger. In *Proceedings of CoNLL-2003*.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2002. *MBT: Memory-Based Tagger, version 1.0, Reference Guide*. ILK Technical Report ILK-0209, University of Tilburg, The Netherlands.

Fien De Meulder and Walter Daelemans. 2003. Memory-Based Named Entity Recognition using Unannotated Data. In *Proceedings of CoNLL-2003*.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named Entity Recognition through Classifier Combination. In *Proceedings of CoNLL-2003*.

James Hammerton. 2003. Named Entity Recognition with Long Short-Term Memory. In *Proceedings of CoNLL-2003*.

Iris Hendrickx and Antal van den Bosch. 2003. Memory-based one-step named-entity recognition: Effects of seed list features, classifier stacking, and unannotated data. In *Proceedings of CoNLL-2003*.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named Entity Recognition with Character-Level Models. In *Proceedings of CoNLL-2003*.

James Mayfield, Paul McNamee, and Christine Piatko. 2003. Named Entity Recognition using Hundreds of Thousands of Features. In *Proceedings of CoNLL-2003*.

Andrew McCallum and Wei Li. 2003. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In *Proceedings of CoNLL-2003*.

Robert Munro, Daren Ler, and Jon Patrick. 2003. Meta-Learning Orthographic and Contextual Models for Language Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text Chunking Using Transformation-Based Learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pages 82–94. Cambridge, MA, USA.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of EACL-SIGDAT 1995*. Dublin, Ireland.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

C.J. van Rijsbergen. 1975. *Information Retrieval*. Buttersworth.

English test	Precision	Recall	$F_{\beta=1}$
Florian	88.99%	88.54%	88.76±0.7
Chieu	88.12%	88.51%	88.31±0.7
Klein	85.93%	86.21%	86.07±0.8
Zhang	86.13%	84.88%	85.50±0.9
Carreras (b)	84.05%	85.96%	85.00±0.8
Curran	84.29%	85.50%	84.89±0.9
Mayfield	84.45%	84.90%	84.67±1.0
Carreras (a)	85.81%	82.84%	84.30±0.9
McCallum	84.52%	83.55%	84.04±0.9
Bender	84.68%	83.18%	83.92±1.0
Munro	80.87%	84.21%	82.50±1.0
Wu	82.02%	81.39%	81.70±0.9
Whitelaw	81.60%	78.05%	79.78±1.0
Hendrickx	76.33%	80.17%	78.20±1.0
De Meulder	75.84%	78.13%	76.97±1.2
Hammerton	69.09%	53.26%	60.15±1.3
Baseline	71.91%	50.90%	59.61±1.2

German test	Precision	Recall	$F_{\beta=1}$
Florian	83.87%	63.71%	72.41±1.3
Klein	80.38%	65.04%	71.90±1.2
Zhang	82.00%	63.03%	71.27±1.5
Mayfield	75.97%	64.82%	69.96±1.4
Carreras (b)	75.47%	63.82%	69.15±1.3
Bender	74.82%	63.82%	68.88±1.3
Curran	75.61%	62.46%	68.41±1.4
McCallum	75.97%	61.72%	68.11±1.4
Munro	69.37%	66.21%	67.75±1.4
Carreras (a)	77.83%	58.02%	66.48±1.5
Wu	75.20%	59.35%	66.34±1.3
Chieu	76.83%	57.34%	65.67±1.4
Hendrickx	71.15%	56.55%	63.02±1.4
De Meulder	63.93%	51.86%	57.27±1.6
Whitelaw	71.05%	44.11%	54.43±1.4
Hammerton	63.49%	38.25%	47.74±1.5
Baseline	31.86%	28.89%	30.30±1.3

Table 5: Overall precision, recall and $F_{\beta=1}$ rates obtained by the sixteen participating systems on the test data sets for the two languages in the CoNLL-2003 shared task.

Casey Whitelaw and Jon Patrick. 2003. Named Entity Recognition Using a Character-based Probabilistic Approach. In *Proceedings of CoNLL-2003*.

Dekai Wu, Grace Ngai, and Marine Carpuat. 2003. A Stacked, Voted, Stacked Model for Named Entity Recognition. In *Proceedings of CoNLL-2003*.

Tong Zhang and David Johnson. 2003. A Robust Risk Minimization based Named Entity Recognition System. In *Proceedings of CoNLL-2003*.