Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists

Zornitsa Kozareva
Dept. de Lenguajes y Sistemas Informáticos
University of Alicante
Alicante, Spain
zkozareva@dlsi.ua.es

Abstract

Current Named Entity Recognition systems suffer from the lack of hand-tagged data as well as degradation when moving to other domain. This paper explores two aspects: the automatic generation of gazetteer lists from unlabeled data; and the building of a Named Entity Recognition system with labeled and unlabeled data.

1 Introduction

Automatic information extraction and information retrieval concerning particular person, location, organization, title of movie or book, juxtaposes to the Named Entity Recognition (NER) task. NER consists in detecting the most silent and informative elements in a text such as names of people, company names, location, monetary currencies, dates. Early NER systems (Fisher et al., 1997), (Black et al., 1998) etc., participating in Message Understanding Conferences (MUC), used linguistic tools and gazetteer lists. However these are difficult to develop and domain sensitive.

To surmount these obstacles, application of machine learning approaches to NER became a research subject. Various state-of-the-art machine learning algorithms such as Maximum Entropy (Borthwick, 1999), AdaBoost(Carreras et al., 2002), Hidden Markov Models (Bikel et al.,), Memory-based Based learning (Tjong Kim Sang, 2002b), have been used[1]. (Klein et al., 2003), (Mayfield et al., 2003), (Wu et al., 2003), (Kozareva et al., 2005c) among others, combined several classifiers to obtain better named entity coverage rate.

Nevertheless all these machine learning algorithms rely on previously hand-labeled training data. Obtaining such data is labor-intensive, time consuming and even might not be present for languages with limited funding. Resource limitation, directed NER research (Collins and Singer, 1999), (Carreras et al., 2003), (Kozareva et al., 2005a) toward the usage of semi-supervised techniques. These techniques are needed, as we live in a multilingual society and access to information from various language sources is reality. The development of NER systems for languages other than English commenced.

This paper presents the development of a Spanish Named Recognition system based on machine learning approach. For it no morphologic or syntactic information was used. However, we propose and incorporate a very simple method for automatic gazetteer[2] construction. Such method can be easily adapted to other languages and it is low-costly obtained as it relies on n-gram extraction from unlabeled data. We compare the performance of our NER system when labeled and unlabeled training data is present.

The paper is organized in the following way: brief explanation about NER process is represented in Section 2. In Section 3 follows feature extraction. The experimental evaluation for the Named Entity detection and classification tasks with and without labeled data are in Sections 4 and 5. We conclude in Section 6.

2 The NER how to

A Named Entity Recognition task can be described as composition of two subtasks, entity de-

[1] For other machine learning methods, consult http://www.cnts.ua.ac.be/conll2002/ner/ http://www.cnts.ua.ac.be/conll2003/ner/

[2] specialized lists of names for location and person names, e.g. Madrid is in the location gazetteer, Mary is in the person gazetteer

tection and entity classification. Entity delimitation consist in determining the boundaries of the entity (e.g. the place from where it starts and the place it finishes). This is important for tracing entities composed of two or more words such as "Presidente de los Estados Unidos "[3], "Universidad Politecnica de Cataluña"[4]. For this purpose, the BIO scheme was incorporated. In this scheme, tag B denotes the start of an entity, tag I continues the entity and tag O marks words that do not form part of an entity. This scheme was initially introduced in CoNLL's (Tjong Kim Sang, 2002a) and (Tjong Kim Sang and De Meulder, 2003) NER competitions, and we decided to adapt it for our experimental work.

Once all entities in the text are detected, they are passed for classification in a predefined set of categories such as location, person, organization or miscellaneous[5] names. This task is known as entity classification. The final NER performance is measured considering the entity detection and classification tasks together.

Our NER approach is based on machine learning. The two algorithms we used for the experiments were instance-based and decision trees, implemented by (Daelemans et al., 2003). They were used with their default parameter settings. We selected the instance-based model, because it is known to be useful when the amount of training data is not sufficient.

Important part in the NE process takes the location and person gazetteer lists which were automatically extracted from unlabeled data. More detailed explanation about their generation can be found in Section 3.

To explore the effect of labeled and unlabeled training data to our NER, two types of experiments were conducted. For the supervised approach, the labels in the training data were previously known. For the semi-supervised approach, the labels in the training data were hidden. We used bootstrapping (Abney, 2002) which refers to a problem setting in which one is given a small set of labeled data and a large set of unlabeled data, and the task is to induce a classifier.

- Goals:
  - utilize a minimal amount of supervised examples;
  - obtain learning from many unlabeled examples;
- General scheme:
  - initial supervision seed examples for training an initial model;
  - corpus classification with seed model;
  - add most confident classifications to training data and iterate.

In our bootstrapping, a newly labeled example was added into the training data L, if the two classifiers C_1 and C_2 agreed on the class of that example. The number n of iterations for our experiments is set up to 25 and when this bound is reached the bootstrapping stops. The scheme we follow is described below.

1. `for` $iteration = 0 \ldots n$ `do`
2. pool 1000 examples from unlabeled data;
3. annotate all 1000 examples with classifier C_1 and C_2;
4. for each of the 1000 examples compare classes of C_1 and C_2;
5. add example into L only if classes of C_1 and C_2 agree;
6. train model with L;
7. calculate result
8. `end for`

Bootstrapping was previously used by (Carreras et al., 2003), who were interested in recognizing Catalan names using Spanish resources. (Becker et al., 2005) employed bootstrapping in an active learning method for tagging entities in an astronomic domain. (Yarowsky, 1995) and (Mihalcea and Moldovan, 2001) utilized bootstrapping for word sense disambiguation. (Collins and Singer, 1999) classified NEs through co-training, (Kozareva et al., 2005a) used self-training and co-training to detect and classify named entities in news domain, (Shen et al., 2004) conducted experiments with multi-criteria-based active learning for biomedical NER.

The experimental data we work with is taken from the CoNLL-2002 competition. The Spanish

[3]"President of the United States"
[4]"Technical University of Cataluña"
[5]book titles, sport events, etc.

corpus[6] comes from news domain and was previously manually annotated. The train data set contains 264715 words of which 18798 are entities and the test set has 51533 words of which 3558 are entities.

We decided to work with available NE annotated corpora in order to conduct an exhaustive and comparative NER study when labeled and unlabeld data is present. For our bootstrapping experiment, we simply ignored the presence of the labels in the training data. Of course this approach can be applied to other domain or language, the only need is labeled test data to conduct correct evaluation.

The evaluation is computed per NE class by the help of *conlleval*[7] script. The evaluation measures are:

$$Precision = \frac{number\ of\ correct\ answers\ found\ by\ the\ system}{number\ of\ answers\ given\ by\ the\ system} \quad (1)$$

$$Recall = \frac{number\ of\ correct\ answers\ found\ by\ the\ system}{number\ of\ correct\ answers\ in\ the\ test\ corpus} \quad (2)$$

$$F_{\beta=1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

3 Feature extraction

Recently diverse machine learning techniques are utilized to resolve various NLP tasks. For all of them crucial role plays the feature extraction and selection module, which leads to optimal classifier performance. This section describes the features used for our Named Entity Recognition task.

Feature vectors ϕ_i=$\{f_1$,...,$f_n\}$ are constructed. The total number of features is denoted by n, and ϕ_i corresponds to the number of examples in the data. In our experiment features represent contextual, lexical and gazetteer information. Here we number each feature and its corresponding argument.

f_1: all letters of w_0[8] are in capitals;

f_2-f_8: $w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}$ initiate in capitals;

f_9: position of w_0 in the current sentence;

f_{10}: frequency of w_0;

f_{11}-f_{17}: word forms of w_0 and the words in $[-3, +3]$ window;

f_{18}: first word making up the entity;

f_{19}: second word making up the entity, if present;

f_{20}: w_{-1} is trigger word for location, person or organization;

f_{21}: w_{+1} is trigger word for location, person or organization;

f_{22}: w_0 belongs to location gazetteer list;

f_{23}: w_0 belongs to first person name gazetteer list;

f_{24}: w_0 belongs to family name gazetteer list;

f_{25}: 0 if the majority of the words in an entity are locations, 1 if the majority of the words in an entity are persons and 2 otherwise.

Features f_{22}, f_{23}, f_{24} were automatically extracted by a simple pattern validation method we propose below.

The corpus from where the gazetteer lists were extracted, forms part of Efe94 and Efe95 Spanish corpora provided for the CLEF[9] competitions. We conducted a simple preprocessing, where all sgml documents were merged in a single file and only the content situated among the text tags was extracted and considered for further processing. As a result, we obtained 1 Gigabyte of unlabeled data, containing 173468453 words. The text was tokenized and the frequency of all unigrams in the corpus was gathered.

The algorithm we propose and use to obtain location and person gazetteer lists is very simple. It consists in finding and validating common patterns, which can be constructed and utilized also for languages other than Spanish.

The location pattern $\langle prep_i,\ w_j \rangle$, looks for preposition i which indicates location in the Spanish language and all corresponding right capitalized context words w_j for preposition i. The dependency relation between $prep_i$ and w_j, conveys the semantic information on the selection restrictions imposed by the two related words. In a walk through example the pattern $\langle en, * \rangle$, extracts all right capitalized context words w_j as $\{$Argentina, Barcelona, Madrid, Valencia$\}$ placed next to preposition "en". These words are taken as location candidates. The selection restriction implies searching for words appearing after the preposition "en" (e.g. en Madrid) and not before the preposition (e.g. Madrid en).

The termination of the pattern extraction $\langle en, * \rangle$, initiates the extraction phase for the next prepositions in $prep_i$ = $\{$en, En, desde, Desde, hacia, Hacia$\}$. This processes is repeated until the complete set of words in the preposition set are validated. Table 1 represents the number of entities extracted

[6] http://www.cnts.ua.ac.be/conll2002/ner/data/

[7] http://www.cnts.ua.ac.be/conll2002/ner/bin/

[8] w_0 indicates the word to be classified.

[9] http://www.clef-campaign.org/

by each one of the preposition patterns.

p_i	en	En	desde	Desde	hacia	Hacia
w_j	15567	2381	1773	320	1336	134

Table 1: *Extracted entities*

The extracted capitalized words are passed through a filtering process. Bigrams "$prep_i$ $Capitalized\_word_j$" with frequency lower than 20 were automatically discarded, because we saw that this threshold removes words that do not tend to appear very often with the location prepositions. In this way misspelled words as Bacelona instead of Barcelona were filtered. From another side, every capitalized word composed of two or three characters, for instance "La, Las" was initiated in a trigram $\langle prep_i, Capitalized\_word_j, Capitalized\_word_{j+1}\rangle$ validation pattern. If these words were seen in combination with other capitalized words and their trigram frequency was higher then 20 they were included in the location gazetteer file. With this trigram validation pattern, locations as "Los Angeles", "Las Palmas", "La Coruña" ,"Nueva York"[10] were extracted.

In total 16819 entities with no repetition were automatically obtained. The words represent countries around the world, European capitals and mostly Spanish cities. Some noisy elements found in the file were person names, which were accompanied by the preposition "en". As person names were capitalized and had frequency higher than the threshold we placed, it was impossible for these names to be automatically detected as erroneous and filtered. However we left these names, since the gazetteer attributes we maintain are mutually nonexclusive. This means the name "Jordan" can be seen in location gazetteer indicating the country Jordan and in the same time can be seen in the person name list indicating the person Jordan. In a real NE application such case is reality, but for the determination of the right category name entity disambiguation is needed as in (Pedersen et al., 2005).

Person gazetteer is constructed with graph exploration algorithm. The graph consists of:

1. two kinds of nodes:
   - First Names
   - Family Names
2. undirected connections between First Names and Family Names.

The graph connects Family Names with First Names, and vice versa. In practice, such a graph is not necessarily connected, as there can be unusual first names and surnames which have no relation with other names in the corpus. Though, the corpus is supposed to contain mostly common names in one and the same language, names from other languages might be present too. In this case, if the foreign name is not connected with a Spanish name, it will never be included in the name list.

Therefore, starting from some common Spanish name will very probably place us in the largest connected component[11]. If there exist other different connected components in the graph, these will be outliers, corresponding to names pertaining to some other language, or combinations of both very unusual first name and family name. The larger the corpus is, the smaller the presence of such additional connected components will be.

The algorithm performs an uninformed breadth-first search. As the graph is not a tree, the stop condition occurs when no more nodes are found. Nodes and connections are found following the pattern $\langle First\_name, Family\_name\rangle$. The node from which we start the search can be a common Spanish first or family name. In our example we started from the Spanish common first name *José*.

The notation $\langle i, j\rangle \in C$ refers to finding in the corpus C the regular expression[12]

```
[A-Z][a-z]* [A-Z][a-z]*
```

This regular expression indicates a possible relation between first name and family name. The scheme of the algorithm is the following:

Let C be the corpus, F be the set of first names, and S be the set of family names.

1. $F = \{$*"José"*$\}$
2. $\forall i \in F$ do
   $S_{new} = S_{new} \cup \{j\}\ , \forall j \mid \langle i, j\rangle \in C$
3. $S = S \cup S_{new}$
4. $\forall j \in S$ do
   $F_{new} = F_{new} \cup \{i\}\ , \forall i \mid \langle i, j\rangle \in C$

[10] New York

[11] A connected component refers to a maximal connected subgraph, in graph theory. A connected graph, is a graph containing only one connected component.

[12] For Spanish some other characters have to be added to the regular expression, such as *ñ* and accents.

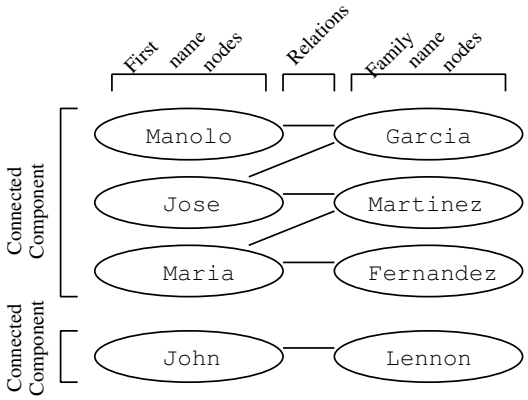

Figure 1: An example of connected components.

5. $F = F \cup F_{new}$

6. `if` $(F_{new} \neq \emptyset) \vee (S_{new} \neq \emptyset)$
   `then goto 2.`
   `else finish.`

Suppose we have a corpus containing the following person names: {"José García", "José Martínez", "Manolo García", "María Martínez", "María Fernández", "John Lennon"} $\subset C$.
Initially we have $F = \{$"José"$\}$ and $S = \emptyset$. After the 3rd step we would have $S = \{$"García", "Martínez"$\}$, and after the 5th step: $F = \{$"José", "Manolo", "María"$\}$. During the next iteration "Fernández" would also be added to S, as "María" is already present in F. Neither "John", nor "Lennon" are connected to the rest of the names, so these will never be added to the sets. This can be seen in Figure 1 as well.

In our implementation, we filtered relations appearing less than 10 times. Thus rare combinations like "Jose Madrid, Mercedes Benz" are filtered. Noise was introduced from names related to both person and organization names. For example the Spanish girl name Mercedes, lead to the node Benz, and as "Mercedes Benz" refers also to the car producing company, noisy elements started to be added through the node "Benz". In total 13713 fist names and 103008 surnames have been automatically extracted.

We believe and prove that constructing automatic location and person name gazetteer lists with the pattern search and validation model we propose is a very easy and practical task. With our approach thousands of names can be obtained, especially given the ample presence of unlabeled data and the World Wide Web.

The purpose of our gazetteer construction was not to make complete gazetteer lists, but rather generate in a quick and automatic way lists of names that can help during our feature construction module.

4 Experiments for delimitation process

In this section we describe the conducted experiments for named entity detection. Previously (Kozareva et al., 2005b) demonstrated that in supervised learning only superficial features as context and ortografics are sufficient to identify the boundaries of a Named Entity. In our experiment the superficial features $f_1 \div f_{10}$ were used by the supervised and semi-supervised classifiers. Table 2 shows the obtained results for **B**egin and **I**nside tags, which actually detect the entities and the total BIO tag performance.

experiment	B	I	BIO
Supervised	94.40	85.74	91.88
Bootstrapped	87.47	68.95	81.62

Table 2: *F-score of detected entities.*

On the first row are the results of the supervised method and on the second row are the highest results of the bootstrapping achieved in its seventeenth iteration. For the supervised learning 91.88% of the entity boundaries were correctly identified and for the bootstrapping 81.62% were correctly detected. The lower performance of bootstrapping is due to the noise introduced during the learning. Some examples were learned with the wrong class and others didn't introduce new information in the training data.

Figure 2 presents the learning curve of the bootstrapping processes for 25 iterations. On each iteration 1000 examples were tagged, but only the examples having classes that coincide by the two classifiers were later included in the training data. We should note that for each iteration the same amount of B, I and O classes was included. Thus the balance among the three different classes in the training data is maintained.

According to z' statistics (Dietterich, 1998), the highest score reached by bootstrapping cannot outperform the supervised method, however if both methods were evaluated on small amount of data the results were similar.

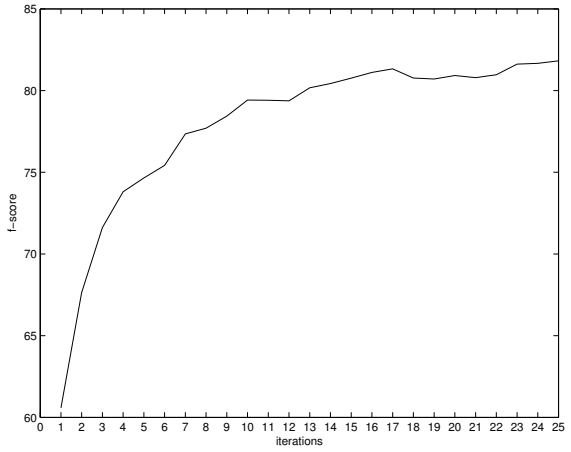

Figure 2: Bootstrapping performance

5 Experiments for classification process

In a Named Entity classification process, to the previously detected Named Entities a predefined category of interest such as name of person, organization, location or miscellaneous names should be assigned. To obtain a better idea of the performance of the classification methods, several experiments were conducted. The influence of the automatically extracted gazetteers was studied, and a comparison of the supervised and semi-supervised methods was done.

experiment	PER	LOC	ORG	MISC
$NoGazetteerSup.$	80.98	71.66	73.72	49.94
$GazetteerSup.$	84.32	75.06	77.83	53.98
$Bootstrapped$	62.59	51.19	50.18	33.04

Table 3: *F-score of classified entities.*

Table 3 shows the obtained results for each one of the experimental settings. The first row indicates the performance of the supervised classifier when no gazetteer information is present. The classifier used f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_{18}, f_{19}, f_{20}, f_{21} attributes. The performance of the second row concerns the same classifier, but including the gazetteer information by adding f_{22}, f_{23}, f_{24} and f_{25} attributes. The third row relates to the bootstrapping process. The attributes used for the supervised and semi-supervised learning were the same.

Results show that among all classes, miscellaneous is the one with the lowest performance. This is related to the heterogeneous information of the category. The other three categories performed above 70%. As expected gazetteer information contributed for better distinction of person and location names. Organization names benefitted from the contextual information, the organization trigger words and the attribute validating if an entity is not a person or location then is treated as an organization. Bootstrapping performance was not high, due to the previously 81% correctly detected named entity boundaries and from another side to the training examples which were incorrectly classified and included into the training data.

In our experiment, unlabeled data was used to construct in an easy and effective way person and location gazetteer lists. By their help supervised and semi-supervised classifiers improved performance. Although one semi-supervised method cannot reach the performance of a supervised classifier, we can say that results are promising. We call them promising in the aspect of constructing NE recognizer for languages with no resources or even adapting the present Spanish Named Entity system to other domain.

6 Conclusions and future work

In this paper we proposed and implemented a pattern validation search in an unlabeled corpus though which gazetteer lists were automatically generated. The gazetteers were used as features by a Named Entity Recognition system. The performance of this NER system, when labeled and unlabeled training data was available, was measured. A comparative study for the information contributed by the gazetteers in the entity classification process was shown.

In the future we intend to develop automatic gazetteers for organization and product names. It is also of interest to divide location gazetteers in subcategories as countries, cities, rivers, mountains as they are useful for Geographic Information Retrieval systems. To explore the behavior of named entity bootstrapping, other domains as bioinformatics will be explored.

Acknowledgements Many thanks to the three anonymous reviewers for their useful comments and suggestions.

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-0664-C02-02 and PROFIT number FIT-340100-2004-14 and by the Valencia Government under project numbers GV04B-276 and GV04B-268.

References

Steven P. Abney. 2002. Bootstrapping. In *Proceedings of Association of Computational Linguists*, pages 360–367.

Markus Becker, Ben Hachey, Beatrice Alex, and Claire Grover. 2005. Optimising selective sampling for bootstrapping named entity recognition. In *Proceedings of the Workshop on Learning with Multiple View, ICML*, pages 5–10. Bonn, Germany.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of Conference on Applied Natural Language Processing*.

William J Black, Fabio Rinaldi, and David Mowatt. 1998. Facile: Description of the ne system used for muc-7. In *Proceedings of MUC-7*.

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

Xavier Carreras, Lluís Màrques, and Lluís Padró. 2002. Named entity extraction using adaboost. In *Proceedings of CoNLL-2002*, pages 167–170. Taipei, Taiwan.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003. Named entity recognition for catalan using only spanish resources and unlabelled data. In *EACL*, pages 43–50.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. Timbl: Tilburg memory-based learner. Technical Report ILK 03-10, Tilburg University, November.

Thomas G. Dietterich. 1998. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

David Fisher, Stephen Soderland, Joseph McCarthy, Fangfang Feng, and Wendy Lehnert. 1997. Description of the umass system as used for muc-6. In *Proceedings of MUC-6*.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 180–183. Edmonton, Canada.

Zornitsa Kozareva, Boyan Bonev, and Andres Montoyo. 2005a. Self-training and co-training for spanish named entity recognition. In *4th Mexican International Conference on Artificial Intelligence*, pages 770–780.

Zornitsa Kozareva, Oscar Ferrández, Andres Montoyo, and Rafael Muñoz. 2005b. Using language resource independent detection for spanish named entity recognition. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, pages 279–283.

Zornitsa Kozareva, Oscar Ferrández, Andrés Montoyo, Rafael Muñoz, and Armando Suárez. 2005c. Combining data-driven systems for improving named entity recognition. In *NLDB*, pages 80–90.

James Mayfield, Paul McNamee, and Christine Piatko. 2003. Named entity recognition using hundreds of thousands of features. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 184–187. Edmonton, Canada.

Rada Mihalcea and Dan I. Moldovan. 2001. A highly accurate bootstrapping algorithm for word sense disambiguation. *International Journal on Artificial Intelligence Tools*, 10(1-2):5–21.

Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. 2005. Name discrimination by clustering similar contexts. In *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City*, pages 226–237.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of Association of Computational Linguists*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Erik F. Tjong Kim Sang. 2002a. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

Erik F. Tjong Kim Sang. 2002b. Memory-based named entity recognition. In *Proceedings of CoNLL-2002*, pages 203–206. Taipei, Taiwan.

Dekai Wu, Grace Ngai, and Marine Carpuat. 2003. A stacked, voted, stacked model for named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 200–203. Edmonton, Canada.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*, pages 189–196.