Feature-Rich Named Entity Recognition for Bulgarian Using Conditional Random Fields

Georgi Georgiev[*]
georgi.georgiev@ontotext.com

Preslav Nakov[†§]
nakov@comp.nus.edu.sg

Kuzman Ganchev[‡]
kuzman@cis.upenn.ed

Petya Osenova[§]
petya@bultreebank.org

Kiril Simov[§]
kivs@bultreebank.org

Abstract

The paper presents a feature-rich approach to the automatic recognition and categorization of named entities (persons, organizations, locations, and miscellaneous) in news text for Bulgarian. We combine well-established features used for other languages with language-specific lexical, syntactic and morphological information. In particular, we make use of the rich tagset annotation of the BulTreeBank (680 morpho-syntactic tags), from which we derive suitable task-specific tagsets (local and nonlocal). We further add domain-specific gazetteers and additional unlabeled data, achieving F_1=89.4%, which is comparable to the state-of-the-art results for English.

Keywords

Named entity recognition, information extraction, conditional random fields, linear models, machine learning, morphology.

1 Introduction

The earliest work on named entity recognition (NER) was based on hand-crafted rules using pattern matching [1]. For instance, a rule could encode the knowledge that a sequence of capitalized words ending in '*Inc.*' is typically the name of an organization. An example of such a system is ANNIE in the GATE architecture [9]. Such systems could achieve very high precision, but typically suffered from low recall. They also required significant manual tuning, which was time-consuming and could be quite complicated when thousands of rules interact in complex manners.

Since the nineties, statistical models have offered a viable alternative while requiring little or no manual tuning at all. Such models typically treat NER as a sequence tagging problem, where each word is tagged with its entity type if it is part of an entity. Generative models such as Hidden Markov Models (HMM) [3, 26] have demonstrated excellent performance on the *Message Understanding Conference* (MUC) datasets [6].

Discriminative models such as locally-normalized maximum-entropy [4] and conditional random fields (CRF) [15] have also been explored for NER. Collins [7] used an HMM tagger to generate n-best outputs, which he reranked discriminatively. By using a semi-Markov CRF, [22] recast NER as a segmentation rather than a tagging problem, thus allowing for richer feature sets.

Recent research also includes semi-supervised methods, e.g., [16] use word clusters derived from large sets of unlabeled data in order to enrich their feature set.

NER can also be viewed as a two-stage process: (1) find the named entities in a sentence, and (2) classify each entity into its type, e.g, person, organization, location, etc. [7] mentions that first identifying named entities without classifying them alleviates some data sparsity issues. [8] focus on the second stage, named entity classification, assuming that the boundaries of the named entities have been found already; they use a bootstrapping approach based on co-training in order to leverage unlabeled examples. [21] use a similar bootstrapping approach for information extraction.

Using CRFs has become the dominant approach to NER [15], allowing for effective feature construction, handling very large feature sets, and modeling complex interactions across multiple levels of granularity; Thus, in the present paper, CRFs will be our learning method of choice. We will employ a rich set of features that (*i*) have been found useful for other languages, (*ii*) can handle expert knowledge in the form of gazetteers and domain-specific predicates, (*iii*) can model rich morpho-syntactic characteristics of Bulgarian, (*iv*) can represent complex predicates, (*v*) can be extracted from raw text automatically.

In the remainder of the paper, we will describe feature generation, and we will discuss the results.

2 Sequence tagging model

The identification of named entity mentions in text can be implemented using a sequence tagger, where each token is labeled with a BIO tag indicating whether it begins (B), is inside (I), or is outside (O) of a named entity mention [20]. Following CoNLL-2002 [24], we further indicate whether it is a person (PER), an organization (ORG), a location (LOC), or miscellaneous (MISC), which yields the following nine tags: B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, B-MISC, I-MISC, and O. See Figure 1 for an example.

---

[*]Ontotext AD, 135 Tsarigradsko Ch., Sofia 1784, Bulgaria

[†]Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417

[‡]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

[§]Linguistic Modelling Laboratory, Institute for Parallel Processing, Bulgarian Academy of Sciences, 25A Acad. G. Bonchev St., 1113 Sofia, Bulgaria

Георги	Първанов	е	президент	на	България	.
Georgi	*Parvanov*	*is*	*President*	*of*	*Bulgaria*	.
B-PER	I-PER	O	O	O	B-ORG	O

Fig. 1: *Sample tagging using the BIO tagset. The sentence contains two named entity mentions: one person, and one organization.*

3 Feature sets

3.1 Basic features

Feature-based models like CRFs are attractive since they reduce the problem to finding a feature set that adequately represents the target task. We used features based on individual words as well as orthographic predicates, as shown in Table 1, and character-level n-gram predicates, $2 \leq n \leq 4$. Bulgarian is morphologically rich and thus such predicates can help the system recognize informative substrings in words that were not seen in training, e.g., Slavic family name endings like *-ов* ('*-ov*') or *-ова* ('*-ova*'), and the possessive *-во* ('*-vo*') that is often used at the end of Bulgarian location names. We also included word prefix and word suffix predicates, also of lengths $2 \leq n \leq 4$. This may seem redundant, but prefix and suffix predicates also take into account the position of the n-gram in the word, which can often be informative. For example *во* ('*vo*') occurring at the end of a word is much more informative than its presence anywhere in the word, e.g., compare *вода* ('*voda*', i.e., 'water') vs. *Димитрово* ('*Dimitrovo*', a village name). We further included predicates that indicate whether the current token occurs in parentheses or inside a quotation. Finally, in addition to the current token, we also used features on the previous and on the following one.

Table 1 lists the orthographic predicates we used.

Predicate name	Regular expression
Initial capital	[А-Я].*
Capital, followed by any	[А-Я].
Initial capitals, alpha	[А-Я][а-я]*
All capitals	[А-Я]+
All lowercase	[а-я]+
Capitals mix	[А-Яа-я]+
Contains a digit	.*[0-9].*
Single digit	[0-9]
Double digit	[0-9][0-9]
Natural number	[0-9]+
Real number	[-0-9]+[\.,]?[0-9]+
Alpha-numeric	[А-Яа-я0-9]+
Roman	[ivxdlcm]+|[IVXDLCM]+
Contains dash	.*-.*
Initial dash	-.*
Ends with dash	.*-
Punctuation	[,\.;:\?!-+"]
Multidots	\.\.+
Ends with dot	.*\.
Acronym	[А-Я]+
Lonely initial	[А-Я]\.
Single character	[А-Яа-я]
Quote	["']

Table 1: ***The orthographic predicates used in our system.*** *The observation list for each token will include a predicate for each regular expression that matches it.*

Even though simple, the above feature set yielded a very good performance on the development data: see Table 2, row A. In order to add expert knowledge to the model, we used some regular expressions that generate predicates on the word by checking whether it ends with some character sequences that are common for Bulgarian names of persons, e.g., *-ска* ('*-ska*'), *-ски* ('*-ski*'), *-ов* ('*-ov*'), *-ва* ('*-va*') or locations, e.g., *-во* ('*-vo*'); see Table 2, row B for details. Another useful approach for adding domain knowledge to the model, used in previous work for named entity recognition and gene mentions tagging, is predicate generation on the basis of membership in a gazetteer.

In our case, gazetteers are lists of words, multi-token units, and acronyms. A straightforward method of integrating these knowledge sources is to create predicates indicating whether a token occurs in one of these gazetteers. For multi-token entries, we required that all entry tokens be matched. We further created similar predicates for the previous and the next tokens. Table 2, row C summarizes the effect of adding these gazetteers.

The following sections will further explore the morpho-syntactic tagset, the feature induction and using additional raw unannotated text.

3.2 Morpho-syntactic features

We made use of the rich tagset annotation of the BulTreeBank [23], from which we derived suitable task-specific tagsets (local and nonlocal).

Initially, we started with the full morpho-syntactic set of the BulTreeBank (680 morpho-syntactic tags), and we were able to achieve some improvements. However, working with so many distinct tags caused data sparsity issues, and missed opportunities for successful generation. We found the tagset was tightly coupled, thus reducing the possibility to model complex context relationships in the text sequence. Some of the tags were quite rare and apparently not very helpful. Since our NER experiments aim to be practical, we divided the tag characteristics (morpho-syntactic and part of speech) into *local* and *nonlocal*. The *local* predicates (111 tags in this set) are linguistically related to other predicates that hold on the same word, e.g., character n-grams, prefixes and suffixes, the word itself, etc. For nouns, they could be gender (e.g., masculine, feminine, neuter), number (e.g., singular, plural, count form), article (e.g., indefinite, definite). The *nonlocal* predicates (230 tags in this set) are related to predicates that hold on words in a particular context, i.e., window around the target word, e.g., the type of noun: common vs. proper.

This treatment of the BulTreeBank tagset stimulates simple and adequate treatment of the feature functions design for the NER task. The *local* characteristics are used alone and in combination with predicates holding on the current word, while the *nonlocal* ones are combined with predicates and words appearing in the local context at positions $\{-3, -2, -1, 0, +1, +2, +3\}$, where 0 is the current word, -1 is the previous one, $+1$ is the next one, etc. This approach induces many useful predicates, which is shown by the overall increase in the system performance: see Table 2, rows D and E.

	Types of predicates	NE type	Precision	Recall	F$_1$-Measure
A	Orthographic	Organization	85.50	82.73	84.10
		Person	86.05	79.86	82.84
		Location	88.34	82.97	85.57
		Miscellaneous	44.83	22.41	29.89
		OVERALL	85.67	78.89	**82.14**
B	+Domain-specific	Organization	85.35	83.81	84.57
		Person	86.46	80.40	83.32
		Location	88.51	82.48	85.39
		Miscellaneous	44.83	22.41	29.89
		OVERALL	85.86	79.20	**82.40**
C	+Gazetteers	Organization	87.89	80.94	84.27
		Person	90.70	84.17	87.31
		Location	88.45	87.59	88.02
		Miscellaneous	48.39	25.86	33.71
		OVERALL	88.26	81.96	**85.00**
D	+Local morphology	Organization	88.93	86.69	87.80
		Person	92.96	90.13	91.52
		Location	89.64	90.29	89.96
		Miscellaneous	57.14	27.12	36.78
		OVERALL	90.19	86.60	**88.36**
E	+Nonlocal morphology	Organization	87.23	88.49	87.86
		Person	90.99	92.46	91.72
		Location	90.34	90.78	90.56
		Miscellaneous	60.00	25.42	35.71
		OVERALL	89.36	88.06	**88.70**
F	+Feature induction	Organization	89.45	88.49	88.97
		Person	93.13	92.46	92.79
		Location	88.11	91.75	89.89
		Miscellaneous	60.00	25.42	35.71
		OVERALL	90.02	88.36	**89.18**
G	+Mutual information	Organization	89.89	89.57	89.73
		Person	93.13	92.46	92.79
		Location	88.89	91.26	90.06
		Miscellaneous	60.00	25.42	35.71
		OVERALL	90.38	88.44	**89.40**

Table 2: *Precision, recall and F$_1$-measure (in %s) for different feature sets on the test dataset.* *(A) Uses orthographic predicates and some simple features like token length. We define this system as our baseline. (B) Adds some simple regular expressions that match common patterns in Bulgarian personal and location names. (C) Adds predicates for gazetteer membership. (D) Adds predicates using local morpho-syntactic characteristics of the current word. (E) Adds nonlocal morpho-syntactic characteristics. (F) Adds feature induction to generate suitable combinations two of or more simple predicates. (G) Further uses unlabeled text.*

For instance, the following feature could be useful:

$$f_i(s,o) = \begin{cases} 1 & \textbf{if } 'WORD = \textit{Джина}' \in o, \\ & 'local\_tag = N - msi' \in o, \\ & \textbf{tag}_\mathbf{0}(s) = B - PER; \\ 0 & \textbf{otherwise}. \end{cases} \quad (1)$$

In the above example, the feature function will have the value of 1 if the the word is *Джина* ('*Dzhina*') and its local tag characteristics are feminine, singular, indefinite, and the named entity tag at this position is 'B-PER'; otherwise, the function value will be 0.

In contrast, we show that *nonlocal* tags would be beneficial in modeling complex context dependencies, for example:

$$f_i(s,o) = \begin{cases} 1 & \textbf{if } 'WORD_{+2} = \textit{влезе}' \in o, \\ & 'nonlocal\_tag = p' \in o, \\ & \textbf{tag}_\mathbf{0}(s) = B - PER; \\ 0 & \textbf{otherwise}. \end{cases}$$

In the above example, the function value will be 1 if the nonlocal tag describes a proper noun, the word *влезе* ('*entered*') appears at position +2, and the current tag is 'B-PER'. In all other cases, it will be 0.

The BulTreeBank tagset was further reduced by dropping information about the types of pronouns, the article was limited to indefinite and definite only, and the number and the count forms were merged into a single class. Rows D and E in Table 2 show the results when using these morpho-syntactic features.

3.3 Inducing complex features

So far, we have described features over a single predicate only, except in the design of morpho-syntactic features. However, it is often useful to create features based on the conjunction of several simple predicates:

$$f_i(s,o) = \begin{cases} 1 & \textbf{if } 'WORD = \textit{Батенберг}' \in o, \\ & 'WORD_{+1} = \textit{управлява}' \in o, \\ & \textbf{tag}_\mathbf{0}(s) = B - PER; \\ 0 & \textbf{otherwise}. \end{cases}$$

The above feature could be useful for disambiguating the token *Батенберг*[1] ('*Batenberg*'), which can be a person's name (e.g., when followed by *управлява*, '*ruled*'). However, it could be also part of a location (e.g., when preceded by *площад*, '*a square*'), and thus we might want to have a special feature for this case:

$$f_i(s, o) = \begin{cases} 1 & \textbf{if } \ 'WORD = \textit{Батенберг}' \in o, \\ & \ 'WORD_{-1} = \textit{площад}' \in o, \\ & \textbf{tag}_0(s) = B - LOC; \\ 0 & \textbf{otherwise}. \end{cases}$$

The system already uses tens of thousands of features, which makes it infeasible to create predicates for the conjunction of all pairs of simple predicates. Even if it were computationally possible, it would still be hard to gather sufficient statistics for most of them. Thus, we use the method described in [14] to limit the search space. Row F in Table 2 shows the results.

3.4 Using unlabeled text

In this section, we try using additional unlabeled text, from which we extract two kinds of additional features.

The first type is pointwise *mutual information* (MI). It is a standard measure of the strength of association between co-occurring items and has been used successfully in extracting collocations from English text [12] and for performing Chinese word segmentation [21, 13, 25], among other tasks.

The MI for two words x_1 and x_2 is defined as follows:

$$MI(x_1, x_2) = \log \frac{\Pr(x_1, x_2)}{\Pr(x_1)\Pr(x_2)}$$

where $\Pr(x)$ is the probability of observing x, and $\Pr(x_1, x_2)$ is the probability of x_2 following x_1.

Estimates of the MI are simple and cheap to compute from unlabeled data alone; this can be done in linear time on the text length.

We used 7.4M words of unlabeled newswire text, from which we extracted the top 1M word pairs, ranked according to the MI score. We then distributed these pairs into separate bins based on their MI values, where bins contained approximately equal numbers of pairs, and we created binary features of the following kind to be integrated in the CRF model:

$$f_i(s, o) = \begin{cases} 1 & \textbf{if } \ 'WORD = \textit{Батенберг}' \in o, \\ & \ 'WORD_{-1} = \textit{площад}' \in o, \\ & MI(WORD, WORD_{+1}) \in bin_x, \\ & \textbf{tag}_0(s) = B - LOC; \\ 0 & \textbf{otherwise}. \end{cases}$$

Initially, we tried using a high number of bins (50K, 100K, 200K and 500K), but did not observe improvements on the development set, probably because of the limited amount of unlabeled text and the sparsity issues resulting thereof. We thus tried smaller numbers of bins, eventually ending up with just two bins, which yielded the highest improvement on the development

set. Table 2, row G shows this also yielded a tiny improvement on the test set: from 89.18% to 89.40%.

We also tried a second kind of features based on the clustering algorithm described in [5], using (1) bottom-up agglomerative word clustering, and (2) the clustering method of [11], but were unable to achieve any performance gains on the development dataset.

4 Experiments and evaluation

In our experiments, we used the Mallet implementation of CRF. We further used manually annotated sentences from the BulTreeBank for training, development and testing:

- *training*: 8,896 sentences (147,339 tokens), including 1,563 organizations, 4,282 persons, 2,300 locations, and 353 miscellaneous named entities;

- *development*: 1,779 sentences (29,467 tokens), including 312 organizations, 856 persons, 383 locations, and 70 miscellaneous named entities;

- *testing*: 2,000 sentences (34,649 tokens), including 315 organizations, 841 persons, 438 locations, and 69 miscellaneous named entities.

In the process of system development, we did many iterations of training and evaluation on the development data, followed by predicate enhancement and new feature construction. For the final evaluation, we trained on a concatenation of the training and the development data and we tested on the unseen test data.

We were very strict in the evaluation and gave no credit for partially discovered named entities: we considered that a named entity was correctly recognized if all tokens it covers were labeled correctly, and no extra tokens were included as part of the entity.

5 Results and discussion

The evaluation results are shown in Table 2. We started with simple orthographic features in our baseline system (row A: F_1=82.14%), and we repeatedly added additional types to improve the performance.

As we can see in rows B and C, using domain-specific features in the form of simple regular expressions and gazetteers yielded about 3% absolute improvement on F_1 to 85%. Adding morpho-syntactic features resulted in additional 3% increase to 88.70% (rows D and E), and using feature induction and mutual information (rows F and G) added 1% more to F_1, which reached 89.40% for our final system (row G).

An examination of system's output on the development dataset shows that the primary source of errors were properly labeled mentions whose boundaries were off by one or more tokens. If the score was relaxed so that tagged entities were considered as true positives if and only if one or more tokens overlap with a correct entry, the performance on the development data would increase a lot. As an extreme example, consider the string *Вашку да Гама* ('*Vashku da Gama*', i.e., '*Vasco da Gama*'), which was incorrectly recognized as covering two entities of type person ('*Вашку*' and '*Гама*').

---

[1] Alexander Joseph of Battenberg (April 5, 1857 – October 23, 1893) was the first prince (knyaz) of modern Bulgaria.

We should note that our results are not directly comparable to previous publications; we are the first to try a statistical approach for Bulgarian NER, which has attracted very little research interest so far and was dominated by rule-based systems. For example, [17] describe adding manual rules for Bulgarian NER to ANNIE, but provide no formal evaluation.

It is still informative to compare our results to those achieved for other Slavic languages even if we use different kinds/amounts of training data and different sets of named entities types. For example, the best P/R/F_1 results for Russian are 79.9/63.7/70.9 (in %), which was calculated for six types of named entities [19]: persons (70.5/53.9/61.1), organizations (72.5/59.8/65.5), locations (91.2/68.7/78.4), dates (77.0/71.7/74.3), percents (87.5/87.5/87.5), and money (80.8/40.4/60.6). For Polish, the best results are the following [18]: persons (90.6/85.3/87.9), organizations (87.9/56.6/68.9), locations (88.4/43.4/58.2), time (81.3/85.9/83.5), percents (100.0/100.0/100.0), and money (97.8/93.8/95.8); overall this makes 91.0/77.5/82.4. For Czech, the best results are the following [10]: 84.0/70.0/76.0. We can see that our results are superior, especially on F_1 and recall.

The state-of-the-art F_1 scores for English at specialized competitions like the Message Understanding Conference and CoNLL-2003 has been 93.39% and 88.76%, respectively. Similarly, at CoNLL-2002[2] and CoNLL-2003[3], the best F_1 for German, Spanish and Dutch were 72.41%, 81.39% and 77.05%, respectively. The highest reported F_1 score for Arabic, which is morphologically richer than Bulgarian, is 83.5% [2]. All these systems were trained on about 200K tokens as is ours, and thus we can conclude that our F_1=89.4% is comparable to the state-of-the-art.

6 Conclusions and future work

Our experiments show that CRF models with carefully-designed features can identify mentions of named entities (organizations, persons, locations and miscellaneous) in Bulgarian text with fairly high accuracy, even without features containing domain-specific knowledge. However, such features, which in our framework take the form of membership in a gazetteer, simple common endings for personal names and location entities, and rich morpho-syntactic tagsets, can lead to improved system performance. Even on the limited training data we had available, we have shown that using external raw text could potentially help on the system performance. However, broader experiments are needed to measure the scope of influence.

We also demonstrate that proper handling of morpho-syntactic tags for morphologically rich languages like Bulgarian could lead to intelligent feature generation and huge performance gains for the named entity tagger. Still, we consider the construction of morpho-syntactic taggers that can handle the rich tagset of the BulTreeBank as a challenging but demanding task. Finally, using raw text is another promising direction we plan to pursue in future work.

---

[2] http://www.cnts.ua.ac.be/conll2002/ner/
[3] http://www.cnts.ua.ac.be/conll2003/ner/

References

[1] D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and M. Tyson. SRI International FASTUS system: MUC-6 test results and analysis. In *Proceedings of MUC-6*, 1995.

[2] Y. Benajiba, M. Diab, and P. Rosso. Arabic named entity recognition using optimal feature sets. In *EMNLP*, 2008.

[3] D. M. Bikel, R. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1):211–231, February 1999.

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

[5] P. Brown, V. Pietra, P. deSouza, J. Lai, and R. Mercer. Class-based n-gram models of natural language. computational linguistics. *Proceedings of the IEEE*, 18(4):467479, 1992.

[6] N. A. Chinchor. Overview of MUC-7/MET-2. In *Proceedings of MUC-7*, 1998.

[7] M. Collins. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron, 2002.

[8] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proceedings of EMNLP-VLC*, 1999.

[9] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of ACL*, 2002.

[10] J. Kravalova and Z. Zabokrtsky. Czech named entity corpus and SVM-based recognizer. In *Proceedings of the Named Entities Workshop*, 2009.

[11] P. Liang. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science., 2005.

[12] D. Lin. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, 1998.

[13] S. Maosong, S. Dayang, and B. Tsou. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of ACL*, 1998.

[14] A. McCallum. Efficiently inducing features of conditional random fields. In *Proceedings of UAI*, 2003.

[15] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Seventh Conference on Natural Language Learning (CoNLL)*, 2003.

[16] S. Miller, J. Guinness, and A. Zamanian. Name tagging with word clusters and discriminative training. In *Proceedings of HLT-NAACL*, pages 277–296, 2004.

[17] E. Paskaleva, G. Angelova, M.Yankova, K. Bontcheva, H. Cunningham, and Y. Wilks. Slavonic named entities in GATE. Technical Report 03, University of Sheffield, 2002.

[18] J. Piskorski, P. Homola, M. Marciniak, A. Mykowiecka, A. Przepirkowski, and M. Wolinski. Information extraction for Polish using the SProUT platform. In *Intelligent Information Systems*, Advances in Soft Computing. Springer, 2004.

[19] B. Popov, A. Kirilov, D. Maynard, and D. Manov. Creation of reusable components and language resources for named entity recognition in Russian. In *Proceedings of LREC*, 2004.

[20] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*. ACL, 1995.

[21] E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479, 1999.

[22] S. Sarawagi and W. Cohen. Semi-Markov conditional random fields for information extraction. In *Proceedings of NIPS*, 2004.

[23] K. Simov, P. Osenova, and M. Slavcheva. BTB-TR03: BulTreeBank morphosyntactic tagset. Technical Report CS-02-01, 2004.

[24] E. F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, 2002.

[25] J. Zhang, J. Gao, and M. Zhou. Extraction of Chinese compound words: An experimental study on a very large corpus. In *Proceedings of the ACL Chinese Processing Workshop*, 2000.

[26] G. Zhou and J. Su. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of ACL-2001*, 2001.