Proceedings Paper:

Yuan, Y., Gao, J. orcid.org/0000-0002-3610-8748 and Zhang, Y. (2018) Supervised Learning for Robust Term Extraction. In: Asian Language Processing (IALP), 2017 International Conference on. International Conference on Asian Language Processing (IALP 2017), 5 - 7 December, 2017, Singapore. IEEE .

https://doi.org/10.1109/IALP.2017.8300603

Supervised Learning for Robust Term Extraction

Yu Yuan
School of Languages and Cultures
Nanjing Uni. of Info. Sci. & Tech.
Email:hittle.yuan@gmail.com

Jie Gao
Department of Computer Science
University of Sheffield
Email: j.gao@sheffield.ac.uk

Yue Zhang
Information Systems Technology & Design
Singapore Uni. of Tech. & Design
Email:yue_zhang@sutd.edu.sg

***Abstract*—We propose a machine learning method to automatically classify the extracted ngrams from a corpus into terms and non-terms. We use 10 common statistics in previous term extraction literature as features for training. The proposed method, applicable to term recognition in multiple domains and languages, can help 1) avoid the laborious work in the post-processing (e.g. subjective threshold setting); 2) handle the skewness and demonstrate noticeable resilience to domain-shift issue of training data. Experiments are carried out on 6 corpora of multiple domains and languages, including GENIA and ACLRD-TEC(1.0) corpus as training set and four TTC subcorpora of wind energy and mobile technology in both Chinese and English as test set. Promising results are found, which indicate that this approach is capable of identifying both single word terms and multiword terms with reasonably good precision and recall.**

***Keywords*-term extraction; supervised learning; classification; n-gram**

I. Introduction

Automatic Term Extraction (ATE) (also known as Term Recognition) has many potential applications, such as human or machine translation, document indexing, lexicography, knowledge engineering, etc.[1]. There have been a plethora of studies into ATE. [2] have identified among them that rule-based systems, purely statistical systems and hybrid systems are three predominant approaches to Automatic Term Recognition (ATR).

Rule-based Approach is heavily language-dependent with low portability and extensibility to a different language. Additionally, PoS based rule system suffers from low recall due to erroneous PoS tagging. Moreover, complex structures using modifiers always pose parsing challenges for most simple PoS tagging algorithms.

Purely **statistical systems** are commonly achieved by means of frequency, significance and degree of association and heuristics measures in order to determine the termhood of words and the unithood for multiple terminology units. However, studies have shown that quantity and quality of the dataset have been identified as the important factors influencing statistical approaches [3].

For the predominant **hybrid approach**, it exploits the advantages of both rule-based and statistical methods. Statistical steps are applied to the narrowed-down list of candidate terms identified by various domain-specific linguistic heuristics so as to further improve the accuracy. Nevertheless, the combination of linguistic filters and statistical ranking would lead to a degenerated precision with the increase of recall, as reported in [4].

In general, these approaches are heuristic, unsupervised in nature. Supervised learning method has been proved superior to unsupervised methods in many NLP tasks, including but not limited to sentiment analysis [5], named entity recognition [6], event detection [7], and coreference resolution [8]. These studies show that supervised learning often produces a state-of-the-art system that outperforms systems built with complex models.

In contrast to other machine learning based methods [9], [10], our approach does not restrict itself on a limited set of certain patterns or unigram/bigram terms. The more difficult challenge of Multi-Word Terms (MWTs) extraction is also tackled instead. In addition, unlike that most of the current studies work only on monolingual data and single domain, the effectiveness of our proposed features and model across multiple domains and languages are examined too. For cross-language processing, we adopt no features that require domain-specific heuristics (e.g. term length).

II. Methodology

In the following we briefly describe our proposed method.

A. Supervised Learning Method

We treat the process of identifying terms as a supervised learning task. Our assumption is that statistical features applicable to both single words and multiword lexemes can be employed to train supervised classifiers, given sufficient annotated data of different domains. This approach is domain independent and could minimize the negative impacts of previous heuristic-based and language dependent methods.

For the purpose of comparison, we select six learning algorithms, including Random Forest (RF), Linear Support Vector Machine (LinearSVC), Radial Basis Function Support Vector Machine (SVC RBF), Multinomial Naive Bayes (MNB), Linear model (Logistic Regression, SLR) and Linear model (SGDClassifier, SGD), in the wish to test whether the proposed approach is robust enough in different types of classifiers and estimate the optimal performance. For the model selection, stratified ten-fold cross-validation is used and repeated grid-search is employed for parameter tuning.

Table I
FEATURES USED FOR TRAINING

Feature	Algorithm
TTF	Total Term Freq.
ATTF	Average TTF
TTF-IDF	TTF with Inverse doccument Freq.
RIDF	Residual IDF
C-Value	C-Value
RAKE	Rapid Keyword Extraction
χ^2	Chi-square
Weirdness	Weirdness
GlossEx	Glossary Extraction
TermEx	Term Extraction

Table II
TRAINING AND TESTING CORPORA

Corpus	# of documents	Size(tokens)	RTL
GENIA	1,999	420,000	35,800
ACL RD-TEC	10,900	36,729,513	22,013
TTC-W (EN)	172	750,855	188
TTC-M (EN)	37	308,263	143
TTC-W (ZH)	178	4,263,336	204
TTC-M (ZH)	92	2,435,232	150

B. Features

Our study is based on the assumption that domain-specific terms has morphological feature, distribution feature, context feature, domain-specific feature and so forth, which distinguish them from common words. Identifying and leveraging those features, to indicate the term's termhood or unithood for MWTs, serve as a basis for the methods of ATR.

We take some conventional measures for candidate terms as our feature input obtained from JATE 2.0 [11] (listed in Table I).

III. EXPERIMENTS

A. Corpora

6 corpora are selected in our experiment, covering 4 different domains and 2 different languages (ranging from small to large size). The GENIA corpus [12],ACL RD-TEC(Version 1.0) [13] are used as training and development data, while TTC subcorpora of wind energy (TTC-W) and mobile technology (TTC-M) in English and Chinese [14] are used as test sets for evaluation. Detailed information of all 6 corpora we used are presented in Table II.

B. Dataset Pre-processing

Both English and Chinese datasets are tokenized. Next, 1-5 grams candidates are extracted and further filtered by stop words.

In the training stage, two methods of feature scaling are applied respectively, namely Min-Max scaling and Mean and Standard deviation scaling. To address the low proportion of true terms in unbalanced data set (see details in Table III), under-sampling method [15] for the majority non-terms is applied.

Table III
TERMS AND NON-TERMS IN NGRAM DATASETS

Ngram Datasets	# of terms	# of non-terms	# recall
GENIA	4,240	45,350	38%
ACL RD-TEC	9,057	858,544	45.1%
TTC-W (EN)	120	30,925	76.5%
TTC-M (EN)	149	20,505	98%
TTC-W (ZH)	125	132,407	41.8%
TTC-M (ZH)	168	105,599	57.1%

All training sets (i.e., GENIA and ACL RD-TEC) are split proportionally (75% for training and 25% for held-out development). All 4 TTC test datasets generated and used in our experiments are labeled data based on the public available Reference Term List (RTL) [16], which contains annotated terms, their inflected forms, and synonymous variants.

C. Evaluation

For our experiment, the performance of 7 classifiers trained on two train sets ('GENIA' and 'ACL RE-TEC') is evaluated on the held-out set and the other 5 separate test sets. Additionally, the contribution of each feature is studied. We assume that all the features are independent from each other, and therefore Pearsons correlation coefficient is employed to evaluate statistical correlation between individual feature and the label (i.e. term vs. non-term). GENIA dataset is employed to study the feature correlation. Pearsons score is computed by Weka tool [17]. The performance variance with Top N features are examined based on the SLR classifier.

Although the task is treated as a binary classification problem, we only focus on the evaluation results corresponding to 'term' class. The standard Precision (P), Recall (R) and F-measure (F1) is adopted to measure the output of the model. These measures are defined as:

$$precison = \frac{tp}{tp + fp} \tag{1}$$

$$recall = \frac{tp}{tp + fn} \tag{2}$$

$$F1 = \frac{2{*}tp}{2{*}tp + fp + fn} \tag{3}$$

where tp stands for true positive (terms), fp stands for false positive (non-terms misclassified as terms) and fn stands for false negative (terms misclassified as non-terms).

Table IV presents the previous state-of-the-art methods on four English corpora. Firstly, TTC TermSuite v2.2[1][18] is used in our experiment as the primary baseline for four English dataset. At the time of writing, it does not support Chinese processing. PoS based C-Value implementation in JATE 2.0 [11] is also chosen as baseline for ACL RD-TEC and GENIA corpus. [19]'s system was the best performed system in the shared task of BioNLP/NLPBA 2004 which used GENIA as dataset. It is worth noting that except for

[1] http://termsuite.github.io/

[19], since the goal of predominant ATR systems focus on term ranking, these results are not directly comparable with our results. Thus, we only report and compare our results with their Top N subset performance. For all test sets, we further compare results between classifiers trained with two different train sets.

Table IV
BASELINES PERFORMANCE ON FOUR ENGLISH CORPORA

		Precision										Recall
Baselines	Dataset	Top 50	Top 100	Top 300	Top 500	Top 800	Top 1000	Top 1500	Top 2000	Top 10000	Overall	Overall
TermSuite v2.2	ACL RD-TEC	0.12	0.09	0.14	0.15	0.12	0.11	-	-	0.06	-	0.15
	GENIA	0.48	0.46	0.48	0.43	0.43	0.44	-	-	0.46	-	0.1
	TTC-W(EN)	0.4	0.29	0.18	0.12	0.08	0.08	-	-	0.01	-	0.44
	TTC-M(EN)	0.32	0.24	0.15	0.12	0.45	0.07	-	-	0.01	-	0.62
JATE 2.0 CValue (PoS)	ACL RD-TEC	0.46	0.41	0.37	0.36	0.35	0.35	0.35	0.36	0.28	-	0.74
	GENIA	0.94	0.91	0.9	0.86	0.84	0.82	0.79	0.77	-	-	0.1
Zhou & Su (2004)	GENIA	-	-	-	-	-	-	-	-	-	0.76	0.69

IV. RESULTS AND DISCUSSION

The performance of 6 classifiers on 6 datasets is presented in Table V. The classifers with best F1 score are considered as best models in our experiment. With regards to the overall recall, baseline results of four English corpora overall are relatively lower than those of our classifiers trained on either train set, except that the result of [19] on GENIA is about 25% higher than that of our optimal model (LinearSVC) trained with ACL RD-TEC dataset.

The recalls of optimal models with ACL RD-TEC train set on two TTC English test sets are relatively higher than the results of those on GENIA train set by 1% and 4% respectively, while the results in two TTC Chinese test sets are much lower than those of GENIA based optimal models by 16% and 11% respectively. More obviously, the optimal model (SVC RBF) with GENIA train set has a 48% higher recall on ACL RD-TEC test set over the ACL RD-TEC based optimal model (LinearSVC) on GENIA test set.

As expected, the Top N precisions of statistic based baselines (TermSuite v2.2 and JATE 2.0 CValue) decease gradually with the increase of recall. The overall precisions of all optimal models trained with either GENIA or ACL RD-TEC dataset obtained much higher precisions than all the Top N subset precisions of TermSuite baselines on two English TTC datasets. In addition, the overall precisions of GENIA based optimal models in ACL RD-TEC test set are much higher than all the top N precisions of JATE 2.0 CValue baseline for ACL RD-TEC corpus (by 26%, 31%, 35%, 36%, 37%, 37%, 37%, 36% and 44% respectively). However, the overall precision (79%) of ACL RD-TEC based optimal model in GENIA test set is relatively lower than all subsets of Top 1500 precisions of JATE 2.0 Cvalue baselines by (by 15%, 12%, 11%, 7%, 5% and 3% respectively), despite that the result is still slightly higher than previous best performed system [19] by 3% and much higher than all Top N precisions of TermSuite baseline. In terms of precision, ACL RD-TEC train set

Table V
MODEL PERFORMANCE ON 6 TESTING DATASETS

		GENIA			ACL RD-TEC		
Classifier	Testing Dataset	Precision	Recall	F1	Precision	Recall	F1
Random Forest	GENIA/ACL(held-out)	0.80	0.84	0.82	0.84	0.88	0.86
	TTC-W(EN)	**0.79**	**0.71**	**0.75**	0.84	0.51	0.64
	TTC-M(EN)	**0.77**	**0.74**	**0.75**	0.83	0.68	0.75
	TTC-W(ZH)	**0.58**	**0.69**	**0.63**	**0.67**	**0.53**	**0.60**
	TTC-M(ZH)	0.57	0.60	0.58	**0.69**	**0.51**	**0.59**
	ACL RD-TEC(1.0)/GENIA	0.51	0.99	0.67	0.82	0.26	0.40
LinearSVC	GENIA/ACL(held-out)	0.70	0.69	0.70	0.82	0.81	0.82
	TTC-W(EN)	0.66	0.79	0.72	0.78	0.55	0.65
	TTC-M(EN)	0.67	0.76	0.71	0.74	0.56	0.63
	TTC-W(ZH)	0.56	0.51	0.53	0.63	0.36	0.46
	TTC-M(ZH)	0.54	0.56	0.55	0.65	0.42	0.51
	ACL RD-TEC(1.0)/GENIA	0.71	0.93	0.81	**0.79**	**0.44**	**0.57**
SVC RBF	GENIA/ACL(held-out)	0.73	0.73	0.73	0.83	0.83	0.83
	TTC-W(EN)	0.69	0.82	0.75	0.76	0.68	0.71
	TTC-M(EN)	0.70	0.82	0.75	**0.79**	**0.78**	**0.78**
	TTC-W(ZH)	0.51	0.53	0.52	0.62	0.42	0.50
	TTC-M(ZH)	0.59	0.65	0.62	0.64	0.44	0.52
	ACL RD-TEC(1.0)/GENIA	**0.72**	**0.92**	**0.81**	0.81	0.41	0.55
MultinomialNB	GENIA/ACL(held-out)	0.64	0.59	0.61	0.79	0.73	0.76
	TTC-W(EN)	0.51	0.89	0.65	0.66	0.75	0.70
	TTC-M(EN)	0.53	0.97	0.69	0.64	0.95	0.76
	TTC-W(ZH)	0.74	0.49	0.59	0.68	0.20	0.31
	TTC-M(ZH)	**0.66**	**0.62**	**0.64**	0.76	0.36	0.49
	ACL RD-TEC(1.0)/GENIA	0.69	0.82	0.75	0.78	0.22	0.35
SGD	GENIA/ACL(held-out)	0.70	0.69	0.70	0.83	0.80	0.82
	TTC-W(EN)	0.69	0.79	0.74	0.73	0.55	0.63
	TTC-M(EN)	0.67	0.82	0.73	0.76	0.55	0.64
	TTC-W(ZH)	0.60	0.49	0.54	0.61	0.34	0.43
	TTC-M(ZH)	0.58	0.59	0.58	0.62	0.38	0.47
	ACL RD-TEC(1.0)/GENIA	**0.72**	**0.92**	**0.81**	0.79	0.43	0.56
SLR	GENIA/ACL(held-out)	0.70	0.70	0.70	0.82	0.81	0.82
	TTC-W(EN)	0.68	0.81	0.74	0.73	0.57	0.64
	TTC-M(EN)	0.70	0.81	0.75	0.73	0.56	0.63
	TTC-W(ZH)	0.58	0.51	0.54	0.60	0.35	0.44
	TTC-M(ZH)	0.59	0.59	0.59	0.65	0.37	0.47
	ACL RD-TEC(1.0)/GENIA	0.71	0.93	0.80	0.78	0.44	0.57

Table VI
SLR MODEL PERFORMANCE ON TOP FEATURES

		GENIA			ACL RD-TEC		
Classifier	Testing Dataset	Precision	Recall	F1	Precision	Recall	F1
Top 1 Feature	GENIA/ACL(held-out)	0.70	0.37	0.48	0.83	0.64	0.73
	TTC-W(EN)	0.75	0.69	0.72	**0.72**	**0.69**	**0.70**
	TTC-M(EN)	**0.77**	**0.79**	**0.78**	**0.74**	**0.79**	**0.77**
	TTC-W(ZH)	0.72	0.57	0.64	**0.77**	**0.54**	**0.63**
	TTC-M(ZH)	0.71	0.53	0.61	**0.73**	**0.53**	**0.61**
	ACL RD-TEC(1.0)/GENIA	0.82	0.67	0.74	0.72	0.33	0.45
Top 2 Feature	GENIA/ACL(held-out)	0.64	0.67	0.65	0.79	0.76	0.78
	TTC-W(EN)	0.74	0.69	0.72	0.70	0.58	0.63
	TTC-M(EN)	0.71	0.78	0.74	0.76	0.71	0.73
	TTC-W(ZH)	0.71	0.53	0.61	0.71	0.47	0.56
	TTC-M(ZH)	0.68	0.53	0.59	0.69	0.51	0.59
	ACL RD-TEC(1.0)/GENIA	**0.75**	**0.81**	**0.78**	**0.76**	**0.41**	**0.53**
Top 3 Feature	GENIA/ACL(held-out)	0.63	0.67	0.65	0.80	0.76	0.78
	TTC-W(EN)	0.70	0.71	0.70	0.70	0.56	0.62
	TTC-M(EN)	0.70	0.78	0.74	0.74	0.69	0.71
	TTC-W(ZH)	0.64	0.54	0.59	0.68	0.45	0.54
	TTC-M(ZH)	0.70	0.53	0.60	0.65	0.50	0.56
	ACL RD-TEC(1.0)/GENIA	0.71	0.82	0.76	0.76	0.41	0.53
Top 4 Feature	GENIA/ACL(held-out)	0.63	0.65	0.64	0.80	0.76	0.78
	TTC-W(EN)	0.68	0.74	0.71	0.74	0.56	0.64
	TTC-M(EN)	0.71	0.79	0.75	0.74	0.69	0.71
	TTC-W(ZH)	0.59	0.53	0.56	0.69	0.45	0.55
	TTC-M(ZH)	0.52	0.62	0.57	0.71	0.50	0.59
	ACL RD-TEC(1.0)/GENIA	0.70	0.84	0.76	0.76	0.41	0.53
Top 5 Feature	GENIA/ACL(held-out)	0.65	0.60	0.62	0.80	0.76	0.78
	TTC-W(EN)	**0.76**	**0.75**	**0.76**	0.73	0.55	0.63
	TTC-M(EN)	0.67	0.82	0.74	0.78	0.69	0.73
	TTC-W(ZH)	**0.69**	**0.61**	**0.65**	0.66	0.43	0.52
	TTC-M(ZH)	**0.68**	**0.57**	**0.62**	0.69	0.48	0.57
	ACL RD-TEC(1.0)/GENIA	0.71	0.83	0.76	0.76	0.40	0.52

based models apparently perform better than those trained on GENIA dataset for the latter four test sets (by 2%, 9%, 3% and 7% respectively), although the result for the first test set (TTC-W(EN)) is 7% lower. Therefore, the current experiment indicates that although larger train set

(ACL RD-TEC) does not necessarily perform better than a smaller (but with a good quality) train set (GENIA) in terms of overall performance (F1), it can be leveraged to boost precision for specific situations (typically in ATR), which precision is a top priority concern. The results of optimal models trained separately with GENIA and ACL RD-TEC on 5 test sets are highlighted in Table V.

V. Conclusion

In this study, we propose a machine learning method that automatically discriminates terms from the large amounts of ngram candidates extracted from textual corpus cross domains and languages. This method exploits 10 commonly used ATE ranking algorithms available in JATE2 library as features for machine learning methods. Our cross-domain and cross-language evaluation presents its robustness and efficiency in generic ATE task.

This approach is advantageous in that it can save the steps of candidate term ranking and subjective threshold setting as seen in conventional ATE methods, and can work across languages and domains. Making use of features computed and extracted by using an open-source ATE library, term classifiers trained for (a) domain(s) can be directly applied to a different domain or language with acceptable accuracy. In the future, we may consider researching into bilingual term extraction with the integration of word and phrase alignment.

References

[1] K. Kageura and B. Umino, "Methods of automatic term recognition: A review," *Terminology*, vol. 3, no. 2, pp. 259–289, 1996.

[2] B.-K. Kang, B.-B. Chang, Y.-R. Chen, and S.-W. Yu, "Extracting terminologically relevant collocations in the translation of chinese monograph," in *International Conference on Natural Language Processing*, R. Dale, K.-F. Wong, J. Su, and O. Y. Kwong, Eds. Jeju Island, Korea: Springer, 2005, pp. 1017–1028.

[3] L. Li, Y. Dang, J. Zhang, and D. Li, "Domain term extraction based on conditional random fields combined with active learning strategy," *Journal of Information &Computational Science*, vol. 9, no. 7, pp. 1931–1940, 2012.

[4] M. T. Pazienza, M. Pennacchiotti, and F. M. Zanzotto, "Terminology extraction: An analysis of linguistic and statistical approaches," in *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference*, S. Sirmakessis, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, ch. 3, pp. 255–279.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, ser. EMNLP '02. Philadelphia, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.

[6] J. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, 2016.

[7] Y. Qin, Y. Zhang, M. Zhang, and D. Zheng, "Feature-rich segment-based news event detection on twitter," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, October 2013, pp. 302–310.

[8] E. Bengtson and D. Roth, "Understanding the value of features for coreference resolution," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Honolulu, Hawaii: Association for Computational Linguistics, 2008, pp. 294–303.

[9] J. Foo and M. Merkel, "Using machine learning to perform automatic term recognition," in *LREC 2010 Workshop on Methods for Automatic Acquisition of Language Resources and Their Evaluation methods, 23 May 2010, Valletta, Malta*, N. Bel, B. Daille, and A. Vasiljevs, Eds. Marrakech, Morocco: FlaReNet, 2010, pp. 49–54.

[10] D. G. Fedorenko, N. Astrakhantsev, and D. Turdakov, "Automatic recognition of domain-specific terms: An experimental evaluation." *SYRCoDIS*, vol. 1031, pp. 15–23, 2013.

[11] Z. Zhang, J. Gao, and F. Ciravegna, "Jate 2.0: Java automatic term extraction with apache solr," *The LREC 2016 Proceedings*, 2016.

[12] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus–a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl 1, pp. i180–i182, 2003.

[13] S. Handschuh and B. QasemiZadeh, "The acl rd-tec: A dataset for benchmarking terminology extraction and classification in computational linguistics," in *COLING 2014: 4th International Workshop on Computational Terminology*, 2014.

[14] H. Blancafort, B. Daille, T. Gornostay, U. Heid, C. Méchoulam, and S. Sharoff, "Ttc: Terminology extraction, translation tools and comparable corpora," in *Proceedings of the 14th EuraLex International Congress*, Leeuwarden, Netherlands, July 2010, pp. 263–268.

[15] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *CoRR*, vol. abs/1609.06570, 2016. [Online]. Available: http://arxiv.org/abs/1609.06570

[16] E. Loginova, A. Gojun, H. Blancafort, M. Guégan, T. Gornostay, and U. Heid, "Reference lists for the evaluation of term extraction tools," in *Proceedings of the Terminology and Knowledge Engineering Conference (TKE'2012)*. Citeseer, 2012.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[18] J. Rocheteau and B. Daille, "Ttc termsuite: A uima application for multilingual terminology extraction from comparable corpora," in *5th International Joint Conference on Natural Language Processing (IJCNLP)*, 2011, pp. 9–12.

[19] G. Zhou and J. Su, "Exploring deep knowledge resources in biomedical name recognition," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, ser. JNLPBA '04. Geneva, Switzerland: Association for Computational Linguistics, 2004, pp. 96–99.