Selecting Features for Domain-Independent Named Entity Recognition

Maksim Tkachenko
St Petersburg State University
St Petersburg, Russia
maksim.tkachenko@math.spbu.ru

Andrey Simanovsky
HP Labs Russia
St Petersburg
andrey.simanovsky@hp.com

Abstract

We propose a domain adaptation method for supervised named entity recognition (NER). Our NER uses conditional random fields and we rank and filter out features of a new unknown domain based on the means of weights learned on known domains. We perform experiments on English texts from OntoNotes version 4 benchmark and see a statistically significant better performance on a small number of features and a convergence of performance to the maximum F_1-measure faster than conventional feature selection (information gain). We also compare with using the weights learned on a mixture of known domains.

1 Introduction

Nowadays the majority of text analytics techniques require that named entities (people, companies, products, etc) are recognized in text. While domain-specific named entity recognition (NER), e.g. in newswire, can be quite precise (Ratinov and Roth, 2009), the accuracy of NER systems is significantly degraded in the presence of several domains (Evans, 2003), especially unknown ones. The generalization of this problem is known as a domain-adaptation (DA) problem. DA is a hard problem(Jiang, 2008). Another issue with NER systems is efficiency. Feature selection can address the efficiency issue in supervised NER systems but regular methods of fast feature selection under-perform in the presence of multiple domains (Satpal and Sarawagi, 2007).

In this paper we consider the problem of feature selection for a supervised NER system that works with texts from multiple domains. We take a large set of feature types that we analyzed on CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) data and compare common feature selection methods (information gain) with methods that rank features based on the weights learned by a machine learning algorithm on the known domains (Jiang and Zhai, 2006). We perform experiments on OntoNotes version 4 (Hovy et al., 2006). We demonstrate that proposed by us feature ranking by domain weights mean is a better feature selection method than information gain and it is competitive against ranking by the weight learned over a mixture of domains.

The main contribution of this paper is that we propose the mean of feature weights learned by the algorithm on known domains as a feature ranking criterion for unknown domains. On large new OntoNotes benchmark, we observe that thresholding on the suggested ranking is a more effective feature selection method.

2 Related work

Named entity recognition is a task that is actively pursued in industry (for example, www.opencalais.com) and in academia (Ratinov and Roth, 2009) since the 6-th Message Understanding Conference (Grishman and Sundheim, 1996). A good overview of the area is given by (Nadeau and Sekine, 2007). We use supervised NER based on conditional random fields (CRF) that were first proposed in (McCallum and Li, 2003). The feature types that we consider come from various previous works (Ratinov and Roth, 2009) etc. We also consider novel feature types.

Features
Context features (token windows and bi-grams)
Token-based features (shape, affixes)
Part-of-speech tags
Brown clusters (Brown et al., 1992)
Clark clusters (Clark, 2003)
Phrasal clusters (Lin et al., 2010) (novel in NER)
Wikipedia gazetteers (Tkatchenko et al., 2011)
DBPedia gazetteers (novel)
Context aggregation (Ratinov and Roth, 2009)
2-stage prediction (Krishnan and Manning, 2006)
Date and hyphenation features

Table 1: Evaluated features.

Several works specifically focus on adapting NER to new domains. Jiang and Zhai (Jiang and Zhai, 2006) explored non-uniform Gaussian prior on the weights of the features to penalize domain specific features. In (Jiang and Zhai, 2007) the authors used the same idea and proposed a two-stage approach to DA. The first stage is recognition of generalizable features. The second stage is learning an appropriate weight with the use of a modified logistic regression framework. A team from Bombay (Satpal and Sarawagi, 2007) proposed an approach to choosing a feature space in which source and target distributions are close. Ando and Zhang (Ando and Zhang, 2005) proposed a semi-supervised multi-task learning framework which is used to identify generalizable features (Ben-david et al., 2007). Klinger and Friedrich (Klinger and Friedrich, 2009) explored information gain and iterative feature pruning in application to feature selection in NER but they did not consider DA perspective.

3 Domain adaptation method

We consider a supervised named entity recognition with a sequential labeling algorithm. The machine learning model uses a comprehensive set of features that represent tokens which are classified using appropriate common labelling schemes like BILOU. Table 1 contains the set of features that we evaluated in this work.

Our primary target is the case of domain adaptation, when there are several known domains but a new document comes from an unknown domain. In this setup we assume that the NER system has a lot of information on the known domains including recognizers that have been trained on the domains or their mixtures. At the same time limited information is available on the unknown domain apart from the document in which named entities are being recognized. Two options were suggested in the literature. One can map the problem dimension set into a higher dimensional space reserving one set of dimensions for each domain and a separate set of dimensions for a combination of all domains (Daume III, 2007). A recognizer for the target document is learned in this higher dimensional space. Alternatively, the case can be addressed by applying a machine learning algorithm that starts not from the default weights for the target document but from the weights that are functions of the known domains. The expectation is that the weights learned for the target document would be biased towards the weights learned on the known domains. The latter approach can also be interpreted as a feature selection strategy — features are ranked according to their weights learned on a mixture of known domains and filtered out based on a threshold value.

We propose to enhance the latter approach with the use of the mean of weights learned across several known domains. The intuition behind this proposition is that a feature that has big weights in several domains is more likely to be important in a new domain than a feature that has a big weight in only one large domain. Thus, macro-averaging makes more sense than micro-averaging that was applied in other works. We also claim and show that the weight-based method works significantly better than a naive feature selection by a common fast feature ranking like information gain.

Our method implies the following steps. NER recognizers are trained on known domains and feature weights produced by them are remembered. When a new domain is encountered, its features are ranked according to the means of the remembered feature weights. Features that do not appear in a domain have a zero weight in it. Previously unseen features are ranked lowest (smoothing can be applied). The obtained ranking is used as a feature utility metrics and top N features are selected.

	Training		Test	
	Range	Size (KB)	Range	Size (KB)
nw-xinhua	0-260	4336	261-325	883
mz-sinorama	0-1062	6047	1063-1078	1519
wb-eng	0-13	1934	14-17	778
bc-msnbc	0-5	2228	6-7	813
bn-cnn	0-375	2989	376-437	716

Table 2: The size of training and test sets for the subcorpora. The file ranges refer to the numbers within the names of the original OntoNotes files.

4 Experiments

We performed experiments on English texts from OntoNotes version 4.0 benchmark. It is a large set of mainly newswire texts of various genres. We used the CoNLL 2003 task NER classes. We compared our feature selection method to information gain and to a feature selection algorithm based on ranking features in accordance with weights learned on a mixture of domains. Five OntoNotes corpora from different domains were used. In each experiment one corpus was withheld as an unknown domain; the rest were used as known domains. Each corpus was split into training and test sets using the document ranges presented in Table 2. The subcorpora are MSNBC (broadcast conversation), CNN (broadcast news), Sinorama (magazine), Xinhua (newswire), and wb (web data).

Figure 1 shows feature selection results for five experiments on *sinorama* subcorpus; the results on the test sets of other subcorpora are similar. In the presented experiment the feature ranking was the same in each of the five *sinorama* experiments and was built using the means of the feature weights learned on the training sets of the four other subcorpora. In each of the experiments a training set of a different OntoNotes subcorpus was used. The most interesting setup is experiment (c), where the training set of *sinorama* was used to collect features for feature selection, since it is the case closest to the appearance of a new domain. Other setups check stability. We can see that our method clearly outperforms information gain and most of the time it reaches flat F_1-measure values (falls into $\pm 1\%$ range) faster than the method based on weights learned from a mixture of domains.

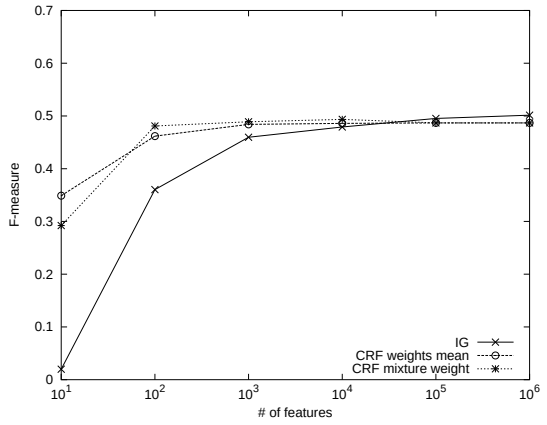
To test statistical significance of the obtained results, we used approximate randomization test described in (Yeh, 2000). It samples a mixture two algorithms being compared and tests if it is better than the baseline. With 100000 iterations we observed that performance advantage of our method against information gain is statistically significant up to more than 1000 features in all experiments. The same holds almost all the time for another weights-based method.

Apart from bias-reduction, feature selection also improves performance of the system. In our experiments the throughput grew from 31 to 45 and 66 tokens per millisecond with reducing 10^5 features to 10^3 and 10^2 respectively.

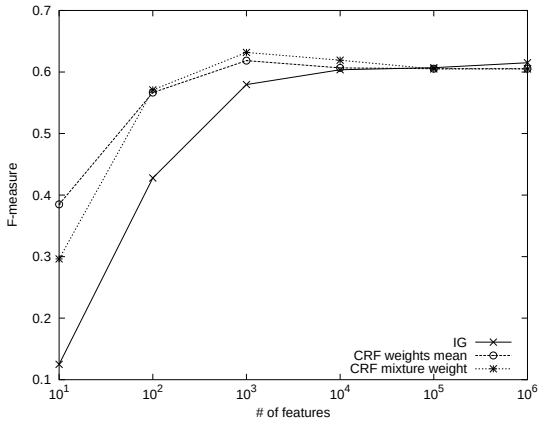
5 Conclusions

In this paper we presented evidence that, in terms of NER F_1-measure, ranking by the mean of feature weights learned on the known domains is a better method of fast feature selection than regular ones (e.g. information gain). It is also competitive against ranking by a weight learned on the mixture of known domains. The experiments on OntoNotes benchmark show that our method obtains higher F_1 measure on a small number of features as compared to other fast feature selection methods. Consequently, our method is less prone to over-fitting.

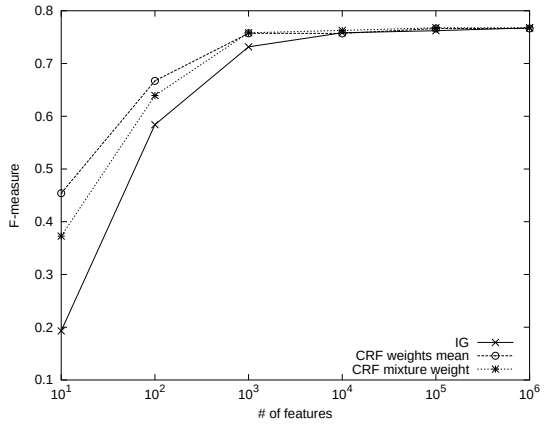
We have explored the two extremes: using a mixture of domains to learn feature weights and taking the mean of feature weights learned on each domain. While we show that the approaches are competitive, our future work is to explore possible combination of the two approaches, e.g., to automatically learn coefficients with which each domain should be taken into account.



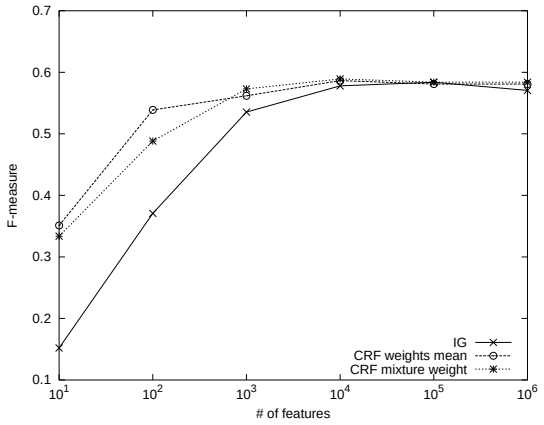
(a) bc-msnbc



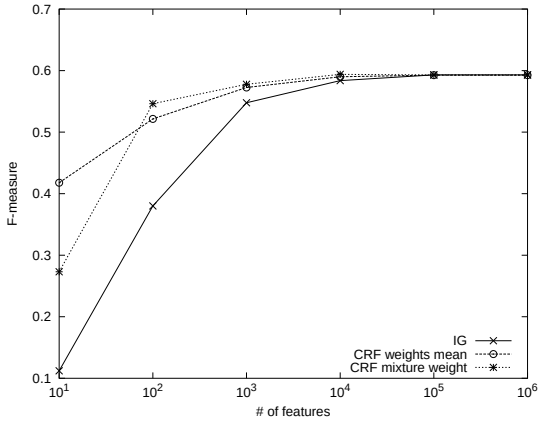
(b) bn-cnn



(c) mz-sinorama



(d) nw-xinhua



(e) wb-eng

Figure 1: Feature selection on different training data. The test data is mz-sinorama. The Y-axis on all charts is F_1-measure. The X-axis is the number of features. Lines with crosses stand for feature selection based on information gain (IG). Lines with stars stand for feature selection based on the feature weight in a mixture of domains (CRF mixture weight). Lines with circles stand for feature selection based on the feature weights mean (CRF weights mean). One can see that the latter lines start at higher values of F_1 and most of the time reach flat part of the chart faster than the lines corresponding to other methods.

References

Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *ACL*. The Association for Computer Linguistics.

Shai Ben-david, John Blitzer, Koby Crammer, and Presented Marina Sokolova. 2007. Analysis of representations for domain adaptation. In *In NIPS*. MIT Press.

Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Richard Evans. 2003. A framework for named entity recognition in the open domain. In *In Proceedings of the Recent Advances in Natural Language Processing (RANLP*, pages 137–144.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90 In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics.

Jing Jiang and Chengxiang Zhai. 2006. Exploiting domain structure for named entity recognition. In *In Human Language Technology Conference*, pages 74–81.

Jing Jiang and ChengXiang Zhai. 2007. A two-stage approach to domain adaptation for statistical classifiers. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 401–410, New York, NY, USA. ACM.

Jing Jiang. 2008. A Literature Survey on Domain Adaptation of Statistical Classifiers, March.

Roman Klinger and Christoph M. Friedrich. 2009. Feature subset selection in conditional random fields for named entity recognition. In *Proceedings of the International Conference RANLP-2009*, pages 185–191, Borovets, Bulgaria, September. Association for Computational Linguistics.

Vijay Krishnan and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1121–1128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New Tools for Web-Scale N-grams.

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 the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, January.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 6.

Sandeepkumar Satpal and Sunita Sarawagi. 2007. Domain adaptation of conditional probability models via feature subsetting. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *PKDD*, volume 4702 of *Lecture Notes in Computer Science*, pages 224–235. Springer.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147, Morristown, NJ, USA. Association for Computational Linguistics.

Maksim Tkatchenko, Alexander Ulanov, and Andrey Simanovsky. 2011. Classifying wikipedia enti-

ties into fine-grained classes. In *ICDE Workshops*, pages 212–217.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences.