Focused training sets to reduce noise in NER feature models

Amber McKenzie
Computer Science and Engineering Department
University of South Carolina
mckenzie.amber@gmail.com

Abstract

Feature and context aggregation play a large role in current NER systems, allowing significant opportunities for research into optimizing these features to cater to different domains. This work strives to reduce the noise introduced into aggregated features from disparate and generic training data in order to allow for contextual features that more closely model the entities in the target data. The proposed approach trains models based on only a part of the training set that is more similar to the target domain. To this end, models are trained for an existing NER system using the top documents from the training set that are similar to the target document in order to demonstrate that this technique can be applied to improve any pre-built NER system. Initial results show an improvement over the University of Illinois NE tagger with a weighted average F1 score of 91.67 compared to the Illinois tagger's score of 91.32. This research serves as a proof-of-concept for future planned work to cluster the training documents to produce a number of more focused models from a given training set, thereby reducing noise and extracting a more representative feature set.

1 Introduction

Though research in the area of named entity recognition (NER) is fairly extensive, current state-of-the-art solutions are generic, succeeding only for domains similar to their training data, and still fail to adequately provide functionality that is adaptable to a broad range of domains (Tkachenko and Simanovsky, 2012). This leaves room for improvement in designing a system that can more easily adapt to previously unseen data. In particular, the increasingly popular feature set produced by feature and context aggregation provides many opportunities for different types of optimization given the strong correlation between the training input and the feature values that are produced. This is due to the fact that aggregation looks at features at a document or corpus level, rather than at the token level, and therefore will be sensitive to changes in the training set. This research looks to exploit this aspect of feature and context aggregation by identifying portions of a training set that are more similar to the target data and will thus provide feature values that are likely more representative of the entities within that data.

Rather than train a model with a full training set, this approach extracts portions of the training data that are most similar to the target data and trains a model using only those documents. This initial work tailors a model to a given target document to demonstrate that less, but more appropriate, training data is preferable to a full generic training set.

Similar to that of Dalton et al. (2011), in which they use passage retrieval to expand their feature set, cosine similarity is used to retrieve documents containing similar entity instances in an effort to achieve a more relevant feature set that will result in more likely output label predictions. However, the proposed approach conducts document similarity above the tagger level, without modifying the underlying tagging system. This allows for domain adaptation improvements using any available NER tagger. This approach is able to be implemented with any pre-existing NER tagger in order to improve the performance of the tagger for out-of-domain data. Initial results show an improvement over the standard NE tagger from the University of Illinois at Urbana-Champaign using a smaller training set and no additional external data sources.

2 Related work

Feature aggregation refers to collecting feature information from across a document or document set, rather than simply taking the information from a particular word instance. With feature aggrega-

Proceedings of NAACL-HLT 2013, pages 411–415,
Atlanta, Georgia, 9–14 June 2013.

tion, researchers strive to expand the context used to predict the classification of a given token. Much of the recent work on features for NER has been related to aggregation of some sort in an effort to widen model coverage, decrease human interaction in the feature generation process, and increase detection and classification accuracy. Many systems incorporating feature aggregation have seen performance improvements over other nearly state-of-the-art systems.

The global features discussed by Chieu and Ng (2003) represent context aggregation in that they extract features about the word in multiple instances within a document. Krishnan and Manning (2006) introduce a two-stage approach to feature aggregation layering two CRFs in which the second uses the output of the first as features, aggregated over both documents and the entire corpus.

Ratinov and Roth (2009) use a similar implementation for their work, substituting relative frequencies of tags within a 1000 token window for the majority tags used by Krishnan and Manning. They refer to the information gathered from aggregation as non-local features and categorize the different approaches as context aggregation, two-stage prediction aggregation and extended prediction history. In an effort not to treat all tokens in a text similarly, which they assert is the case with context aggregation and two-stage prediction, Ratinov and Roth developed an approach for non-local feature generation based on extended prediction history. Their approach is based on the idea that named entities are easier to spot at the beginning of texts where they are first introduced. They keep track of all label assignments for the token in the last 1000 words and use that probability information as a prediction history feature for the token.

Huang and Yates (2009) present their feature aggregation approaches in the form of smoothing of the dataset. Their goal for smoothing is the same as for aggregation in that they strive to extend the usefulness of the model by sharing information about multiple contexts for a token in order to provide more information about words that are rarely, or never, seen in training. In experimentation, the authors found that their smoothing approach improved performance on rare words, out-of-domain text, and smaller training sets.

Dalton et al. (2011) take an external knowledge approach to context aggregation. Using an information retrieval method called Pseudo-Relevance Feedback (PRF), they query for relevant passages in an external data set using the context for the target token. Given that they searched for the context that the entity occurs in, it is assumed that the top returned passages all contain instances of the entity with the same label. They then aggregate the features for this token across a number of the top retrieved documents and induce features based on this information. Their approach is compared with the Stanford and Illinois NER systems and found that their aggregated features improved performance over those systems.

Apart from the body of work attempting to incorporate external data sources, such as Wikipedia, to augment training data, approaches for domain adaptation for NER focus on either adapting features to fit the domain or searching for more abstract features that can span multiple domains (Zhang and Johnson, 2003; Huang and Yates, 2009; Lin and Wu, 2009). This is largely due to the assumption that a domain-specific, tagged training set will not be available for most target domains.

This research expands on previous work by providing a more informative training set that is a closer representation of the features contained in the target documents. Further, the proposed system does not require external knowledge sources or additional tagged data to augment the utilized training set. The modifications that are made are implemented above the tagger level allowing for any existing tagger to be used without need to alter the underlying source code.

3 NER approach

Feature aggregation has become an integral part of building an NER prediction model. Because aggregating the context of every named entity across an entire training set can be fairly computationally expensive and introduces significant noise into the features due to the many contexts in which an entity may occur, many researchers have chosen instead to conduct local aggregation, such as across a document, or with a certain window of tokens that may span several documents. The NER tagger produced by the University of Illinois at Urbana-Champaign, one of the best performing systems on the CoNLL 2003 data set, uses a 1000 token window across which to take their global context aggregation (Ratinov and Roth, 2009). By choosing 1000 tokens, the researchers hope to be able to

capture a large enough example set to provide a robust feature value while maintaining a reasonable computation time. However, this method leaves the choice of context to chance: determined by how the documents are organized within the training set. A better option would be to choose the context that best represents the entities to be tagged. To that end, this work serves to provide a more useful and informative training set from which to pull context information.

The hypothesis explored in this work is that the context aggregation feature would prove more useful if the training data were more specific to the target entities. For this research, documents from the training set were compiled based on their similarity to the target document. These documents were then used to train a model for the Illinois NE tagger. In this way we strive to reduce the noise present in the context aggregation feature as a result of the generic contexts found in a large, often heterogeneous, training set and produce feature values that are more representative of the target entities, thus producing more reliable output labels.

3.1 Methodology

For an initial proof-of-concept test, vectors were created for all test (not the development set) and training documents in the CoNLL-2003 shared task data. This corpus was chosen due to the previous NER research using this corpus and the results available using the LBJ tagger. Also, it has been noted that the test and training sets within the corpus are not as similar in nature as are the development and training sets (Ratinov and Roth, 2009). The training set contains 946 documents, while the test set contains 231. For each test document, a specified number of the top documents from the training set most similar to that test document was collected. For this initial work, a simple cosine similarity measure was used. These top similar documents were used as a training set for the LBJ tagger, and the test document was then tagged using the resultant model. The system was tested by pulling the top 20, 50, 100, and 300 similar training documents to train the models. The performance of this customized model is compared to that of the standard, two-phase LBJ tagger trained on the full CoNLL '03 training set.

3.2 Results

For this research, because each test document is tagged using a different model, we chose to measure our performance on a per-document basis, rather than the standard overall measure for the entire test set.[1] This performance is compared to that achieved by the standard LBJ tagger on the same document. Figure 1 shows how many documents were tagged more accurately using the proposed system compared to the LBJ tagger.

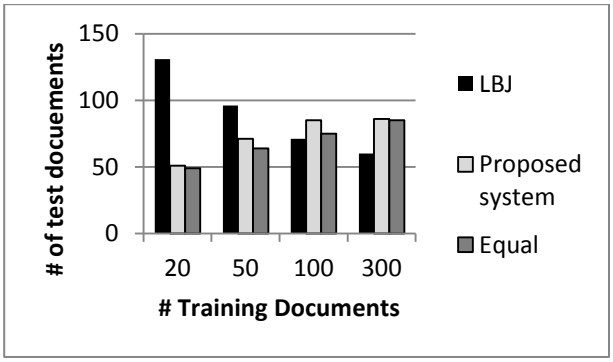

Figure 1 – Results showing the number of documents for which each system performed better or for which they had equal F1 scores.

Further, Figure 2 displays the average percentage better and worse in terms of F1 score for each training document size. In contrast to Figure 1, Figure 2 demonstrates the average difference in F1 scores between the LBJ tagger trained on the entire training set and the proposed system trained on varying numbers of training documents. These numbers indicate that there exists an optimal balance that can achieve the dual advantages of having a smaller, more relevant training set while also maintaining enough data to ensure enough features to accurately predict NER labels.

The overall aggregated difference is also provided as a more global view of performance achievements. This measurement is calculated by multiplying the F1 score of a given document by the number of entity tokens contained in that document, summing these calculations, and then dividing by the total number of entity tokens across the test dataset. These results reveal an improvement over the Illinois tagger for the 300 document train-

[1] The Illinois NE tagger only provides performance information in the form of percentages and does not give enough information to calculate an overall F1 score for the test set using the CoNLL eval script.

ing set with a weighted average F1 score of 91.67 compared to the Illinois score of 91.32.

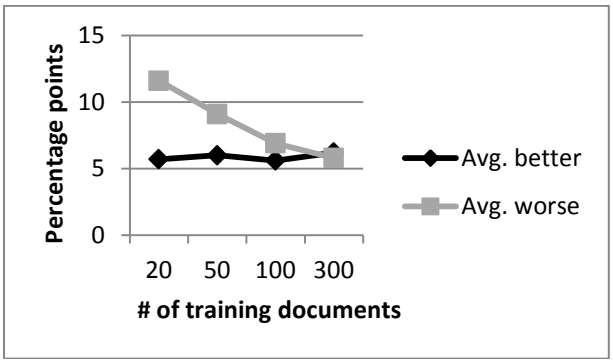

Figure 2 – Average percentage points better and worse in the F1 score that the proposed system achieved compared to the standard LBJ tagger for models trained with the top 20, 50, 100, and 300 similar documents.

These initial results demonstrate that an available training set can be easily tailored to better serve the needs of a target data set that differs from the training set and showed improvements on an existing competitive NER system by modifying the training data set used to build the prediction model. By identifying a smaller, relevant training set, the sequence tagging model is better equipped to accurately predict output labels for target data that does not closely align with the training documents.

4 Future work

Given the computational expense of training a model for each individual document to be tagged, improvements must be made to the approach to transform it into a viable long-term NER solution. The next logical step in this research will be to cluster the training documents and train models based on those clusters. Subsequently, the test documents can be clustered to the training set clusters and be tagged using the appropriate model for that cluster set. Alternatively, the test set could be initially clustered, with the training set then fit to those clusters. Tests must be conducted to determine which option produces the best prediction accuracy levels. Once a viable clustering methodology has been developed, further testing will be conducted to compare it with some of the best current techniques (e.g. the work of Dalton et. al 2011) to provide a more comprehensive evaluation.

The results presented here were achieved using baseline document representation and document similarity techniques. Significant work remains for experimentation to determine which alternative methodologies will result in the optimal NER performance. Not only could different clustering algorithms be employed, but an investigation into which type of clustering, in particular linear or hierarchical, is better suited for NER would be prudent. Also, further work will test the validity of this approach for successful domain adaptation by demonstrating that it is extensible to other data sets.

5 Summary

This research has implications in the NER domain adaptation space as it demonstrates that fewer training documents are required as long as they are sufficiently similar to the targeted test set. This methodology could potentially allow for better utilization of existing, freely-available (possibly generic) training sets by extracting portions of the training set that are more similar to the target data. It also allows for existing NER systems to be better adapted to domain-specific data without modification for feature augmentation or the inclusion of additional external data sources. The opportunities for continuing this tread of research are numerous, and initial results illustrate significant promise given the relative simplicity of the execution compared with its achievement.

6 Acknowledgements

This document was prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285; managed by UT-Battelle, LLC, for the US Department of Energy under contract number DE-AC05-00OR22725.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

Dan Wu, Wee Sun Lee, and Hai Leong Chieu. 2009. Domain adaptive bootstrapping for named entity recognition. In *EMNLP*, pp. 142 - 147.

Fei Hueng, and Alexander Yates. 2009. Distributional representations for handling sparsity in supervised sequence-labeling. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 1*, pp. 495-503. Suntec, Singapore: ACL.

Hai Leong Chieu, and Hwee Tou Ng. 2003. Named entity recognition with a maximum entropy approach. *Proceedings CoNLL 2003*, (pp. 160-163). Edmonton, Canada.

Jeffrey Dalton, James Allan, and David A. Smith. 2011. Passage retrieval for incorporating global evidence in sequence labeling. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 355-364). Glasgow, UK: ACM.

Lev Ratinov, and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. *CoNLL '09 Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 147-155). Boulder, CO: Association for Computational Linguistics.

Maksim Tkachenko and Andrey Simanovsky. 2012. Named entity recognition: exploiting features. *Proceedings of KONVENS 2012*. Vienna, Austria.

Terry Koo, Xavier Carreras, and Michael John Collins. 2008. Simple semi-supervised dependency parsing. *Proceedings of ACL*, (pp. 595-603).

Vijay Krishnan, and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, (pp. 1121-1128). Sydney, Australia.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 1*, pp. 359-367. Portland, OR: Association for Computational Linguistics.