MayoNLP at SemEval 2017 Task 10: Word Embedding Distance Pattern for Keyphrase Classification in Scientific Publications

Sijia Liu[1,2], **Feichen Shen**[1], **Vipin Chaudhary**[2], and **Hongfang Liu**[1]

[1]Department of Health Science Research, Mayo Clinic, USA
{liu.sijia,shen.feichen,liu.hongfang}@mayo.edu
[2]Department of Computer Science and Engineering, SUNY at Buffalo, USA
vipin@buffalo.edu

Abstract

In this paper, we present MayoNLP's results from the participation in the ScienceIE share task at SemEval 2017. We focused on the keyphrase classification task (Subtask B). We explored semantic similarities and patterns of keyphrases in scientific publications using pre-trained word embedding models. Word Embedding Distance Pattern, which uses the head noun word embedding to generate distance patterns based on labeled keyphrases, is proposed as an incremental feature set to enhance the conventional Named Entity Recognition feature sets. Support vector machine is used as the supervised classifier for keyphrase classification. Our system achieved an overall F1 score of 0.67 for keyphrase classification and 0.64 for keyphrase classification and relation detection.

1 Introduction

In this paper, we present details of our participation in the SemEval 2017 Task 10, ScienceIE (Augenstein et al., 2017). Named Entity Recognition (NER) is one of the major challenges in Natural Language Processing (NLP) and text mining. The interesting entity types in NER tasks vary from communities and corpora. In general, NLP community mainly focused on the identification of proper nouns or noun phrases, e.g., locations, names and organizations in news corpora (Nadeau and Sekine, 2007). In contrast, biomedical community is more interested in finding biomedical or clinical terminologies (Leaman and Gonzalez, 2008; Tsuruoka and Tsujii, 2005) in biomedical texts and scientific literatures. There are several machine learning based methods used in biomedical NER, which include Support Vector Machine (SVM) (Lee et al., 2004), Hidden Markov Model (HMM) (Zhou and Su, 2004) and Conditional Random Field (CRF) (Tsai et al., 2006).

Semantic word embedding (Mikolov et al., 2013) is designed to capture different degrees of similarity between words using a vectorized representation, which preserves semantic and syntactic relationships. Word embeddings and word embedding based features have drawn increasing attention for classification tasks (Ma et al., 2015) and similarity prediction tasks (Afzal et al., 2016).

We leveraged pre-trained word embeddings to obtain head noun pattern features, and combined several other NER feature sets to improve the keyphrase classification performance. Although our team participated in Scenario 2 (keyphrase classification and relation detection), our efforts were focused on keyphrase classification task (Subtask B). For the relation detection problem (Subtask C), we implemented a straightforward rule-based system to detect synonyms and hyponyms given annotated keyphrases.

The rest of the paper is organized as follows. Section 2 briefly introduces the corpus used in this task. Section 3 discusses the methods proposed in our NER system. Section 4 addresses the experimental results in the development set, our submitted runs and official evaluation results. Finally, Section 5 concludes the paper with possible extensions for future work.

2 Materials

The corpus provided by the ScienceIE organizers consisted of 500 introductory paragraphs from ScienceDirect journal articles in Computer Science, Material Sciences and Physics. The corpus

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 956–960,
Vancouver, Canada, August 3 - 4, 2017.

was divided into training, development and test sets, which contained 350, 50 and 100 documents, respectively. It is the first publicly available corpus with annotations focused on the research topics and goals of general domain scientific literature. The annotated keyphrases were relatively longer than other annotated corpora, which makes the boundary detection and classification task very challenging. More details of the corpus can be found in (Augenstein et al., 2017).

3 Methods

3.1 Preprocessing

To facilitate feature extraction for supervised classification, all plain text sentences and annotations were pre-processed by NLTK[1] for tokenizing, Part-of-Speech (POS) tagging and sentence detection.

3.2 Head Noun Extraction

Intuitively, the head noun of a keyphrase provides important information of its semantic category (Li, 2010). For example, in the phrase "*homonuclear chains of tangent Mie spherical CG segments*" from the ScienceIE 2017 corpus, the noun "*chains*" determines the phrase is from the category "Material". In another example, the category of the phrase "*applications of methodology of research*" is determined by the head noun "*application*", which is an instance of "Task". Extracting the head noun can help eliminate ambiguous contexts while preserving the semantic information for the classification step. Therefore, we used the extracted head noun features, rather than the features from whole phrase, to determine the semantic category.

A shallow parsing approach is applied to extract the head nouns from given phrases. We removed the part at and after the preposition token "of", "with", "for" and "on", and kept only the features from the head noun for the feature extraction step. In the above examples, we extracted the head noun "*chains*" and "*applications*".

3.3 Feature Set

Given a sentence and a head noun token w_i, we adopted several commonly used feature sets as the input of supervised classifiers for the baseline system.

[1] http://www.nltk.org/

Lexical features The lower case of tokens in ± 2 window.

Orthographic features The set of case, character and symbolic features of given token. Orthographic features are all binary features: if the token contains only upper case letters, if only the first letter is in upper case, if the token contains only alphabetic characters, if the token contains numbers, and if the token starts with alphabetic characters and ends with numbers.

Part-Of-Speech features The Part-Of-Speech (POS) tags for the tokens in ± 2 window.

Lemma features Lemmatized word of w_i and its verb form from WordNet. For example, for the token "*derivations*", the lemmatized word is "*derivation*" and the verb form is "*derive*".

3.4 Word Embedding Distance Pattern

The extraction of head nouns in keyphrases enables utilizing word embedding information as features in the keyphrase classification task.

To improve the performance using baseline NER features described above, we proposed Word Embedding Distance Pattern (WEDP). It is based on the assumption that the differences among the head nouns in each semantic category should follow similar patterns in semantic word vector space. We would like to validate and obtain the patterns in this keyphrase classification task.

We selected 10 most frequent head nouns from each category in the training corpus. After excluding the duplications, we obtained the following list of keywords M={*model, particle, data, system, film, problem, algorithm, function, effect, equation, reaction, method, surface, alloy, layer, structure*}. We also added the category names (*task, material, process*) into M.

Given a token w, the word embedding distance to each of the k-th word-embedding above is calculated by

$$d_k(w) = dist(w2v(w), w2v(M_k)), \qquad (1)$$

where $k = 1, \ldots, |K|$, the distance function $dist$ is the cosine distance, and $w2v$ is the dictionary lookup method, which returns the embedding of the input token from a pre-trained word to vector (word2vec) model. If the token w cannot be found in the word embedding dictionary, we set $d_k(w) = 1$ for all k.

In this study, we used the word embedding

“ ()”, “, where .”, “, i.e.)”,
“ (i.e., in terms of)”, “, or equivalently, .”,
“, which is the ,”, “, the so-called ”,
“, which are called [,.]”, “, which is called [,.]”,
“ (the)”

Table 1: Matching contexts for synonym detection, separated by comma (“,”)

model GloVe [2]. We tested the overall classification performance against different dimensions, ranging from 50 to 300, but found the differences are negligible. For better algorithm efficiency, we selected the 50-dimension model as the final solution.

3.5 Classification

In this study, we modeled the keyphrase classification task as a supervised multi-class classification problem. All features described in Section 3.3 were encoded into a sparse vector, and then combined with the WEDP as the input of supervised classifiers.

3.6 Relation Extraction

For the relation extraction subtask, we implemented a simple rule-based system. For each sentence, we considered all possible pairs of the entities as relation candidates. For each candidate, the context texts between two entities were extracted, including one character after the entity appeared later of the pair. The matching patterns we used are shown in Table 1. If any of those patterns matched with the context, we identified the pair as a detected relation. Relations sharing at least one entity were grouped together as one relation, according to the requirement of output format.

We used hearstPattern[3] which implements Hearst patterns (Hearst, 1992) for hyponym detection.

4 Results

We tested several supervised classification methods, The results on development set are shown in Table 2. The L2-loss linear kernel SVM was selected as the classifier and used the scikit-learn[4] implementation. The result also validated that SVM can outperform other classification methods in high dimensional data (Chang and Lin, 2011). The hyperparameter C was tested in the range from 0.01 to 10. The F1 scores[5] range from 0.70 to 0.78 and yields the highest F1-score on the development set when C is set to 0.5.

Classifier	Material	Process	Task	Avg
ExtraTrees	0.77	0.69	0.45	0.68
SGD	0.76	0.67	0.35	0.65
5-NN	0.66	0.59	0.24	0.56
RBF-SVM	0.76	0.71	0.29	0.65
Linear SVM	0.88	0.75	0.45	**0.78**

Table 2: F1 scores for different classification methods on development set for Subtask B. (5-NN: 5 Nearest Neighbor; SGD: Stochastic Gradient Descent; RBF: Radial Basis Function)

Ablation experiments were conducted on the development set to find the importance of individual feature sets. The ablation results in F1 scores are shown in Table 3. From Table 3, we see that both the baseline feature sets and the WEDP contributed to the overall performance, since the combination of these two sets outperform the other feature settings.

Feature sets	F1 score
Lexical features, ± 1 window	0.68
+ ± 2 window	0.69
+ Orthographic features	0.71
+ POS features	0.72
+ Lemma feastures	0.72
baseline features only	0.72
WEDP features only	0.67
All	0.78

Table 3: Ablation F1 scores of keyphrase classification on the development set

The official evaluation uses the standard precision (P), recall (R) and F1 score as the metrics. We submitted two runs for official evaluation. Run 1 uses the feature set described in Section 3.3 with synonym detection result. Run 2 is derived by extending Run 1 by predicted hyponyms. Both runs achieved F1 score of 0.64 for Subtasks B and C. This was due to the insignificance from the positive cases of “Hyponym-of” relations on Run 2. The results of Run 2 are shown in

[2] http://nlp.stanford.edu/projects/glove/
[3] https://github.com/mmichelsonIF/hearst_patterns_python
[4] http://scikit-learn.org/
[5] Unless specified, the F1 scores mentioned in this section are micro-average F1 scores in keyphrase classification.

Category	P	R	F1 score	Support
Material	0.74	0.78	0.76	904
Process	0.69	0.64	0.66	954
Task	0.28	0.29	0.28	193
Synonym-of	0.42	0.27	0.33	112
Hyponym-of	0.16	0.03	0.05	95
Entity	0.67	0.67	0.67	2051
Relation	0.37	0.16	0.23	207
Overall	0.66	0.62	0.64	2258

Table 4: Official evaluation results of the best submitted run on the test set using annotated keyphrase boundaries (Scenario 2).

Table 4. From the results, "Task" is the most difficult category for our proposed method, but its relatively low proportion reduces its impact on the overall F1 score. Compared to the development set Table 2, the F1 scores of all three categories drop by at least 0.09, which indicates the selected classifier suffers from overfitting.

5 Conclusion

In this paper, we presented details of MayoNLP's participation in the ScienceIE share task at SemEval 2017. We used a supervised classifier for the keyphrase classification task using word embedding distance patterns, which improves the performance of conventional feature sets. Our system achieved an overall F1 score of 0.67 for keyphrase classification subtask and 0.64 for keyphrase classification and relation detection subtasks. It outperformed other participating systems in Scenario 2.

A future extension of this work is to test the patterns on different pre-trained word embeddings. We will also develop methods for more accurate key noun extraction such as dependency parsing, to improve the overall classification performance.

Acknowledgments

We would like to thank Yanshan Wang, Ravikumar Komandur Elayavilli, and Majid Rastegar-Mojarad for their valuable suggestions. This work is supported by NIH grants R01GM102282-01A1 and R01EB19403-01 and NSF IPA grant.

References

Naveed Afzal, Yanshan Wang, and Hongfang Liu. 2016. MayoNLP at SemEval-2016 Task 1: Semantic textual similarity based on lexical semantic net and deep learning semantic model. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. San Diego, CA, USA.

Isabelle Augenstein, Mrinal Kanti Das, Sebastian Riedel, Lakshmi Nair Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Vancouver, Canada.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2(3):27:1–27:27. https://doi.org/10.1145/1961189.1961199.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING 1992, pages 539–545. https://doi.org/10.3115/992133.992154.

Robert Leaman and Graciela Gonzalez. 2008. Banner: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing (PSB)*. pages 652–663.

Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, and Hae-Chang Rim. 2004. Biomedical named entity recognition using two-phase model based on SVMs. *Journal of Biomedical Informatics* 37(6):436 – 447. Named Entity Recognition in Biomedicine. https://doi.org/10.1016/j.jbi.2004.08.012.

Xiao Li. 2010. Understanding the semantic structure of noun phrase queries. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '10, pages 1337–1345. http://dl.acm.org/citation.cfm?id=1858681.1858817.

Mingbo Ma, Liang Huang, Bing Xiang, and Bowen Zhou. 2015. Dependency-based convolutional neural networks for sentence embedding. In *Proceedings of the 53st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Beijing, China, pages 174–179. http://aclweb.org/anthology/P/P15/P15-2029.pdf.

Tomas Mikolov, Scott Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvistic Investigationes* 30(1):3–26. https://doi.org/10.1075/li.30.1.03nad.

Tzong-han Tsai, Wen-Chi Chou, Shih-Hung Wu, Ting-Yi Sung, Jieh Hsiang, and Wen-Lian Hsu. 2006. Integrating linguistic knowledge into a conditional random fieldframework to identify biomedical named entities. *Expert Syst. Appl.* 30(1):117–128. https://doi.org/10.1016/j.eswa.2005.09.072.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '05, pages 467–474. https://doi.org/10.3115/1220575.1220634.

Guodong Zhou and Jiang Su. 2004. Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*. Association for Computational Linguistics, Stroudsburg, PA, USA, JNLPBA '04, pages 96–99. http://dl.acm.org/citation.cfm?id=1567594.1567616.