LABDA at SemEval-2017 Task 10: Extracting Keyphrases from Scientific Publications by combining the BANNER tool and the UMLS Semantic Network

Isabel Segura-Bedmar, Cristóbal Colón-Ruiz, Paloma Martínez
Computer Science Department, Universidad Carlos III of Madrid
Avd. Universidad, 30, Leganés, 28911, Madrid, Spain
isegura,ccolon,pmf@inf.uc3m.es

Abstract

This paper describes the system presented by the LABDA group at SemEval 2017 Task 10 ScienceIE, specifically for the subtasks of identification and classification of keyphrases from scientific articles. For the task of identification, we use the BANNER tool, a named entity recognition system, which is based on conditional random fields (CRF) and has obtained successful results in the biomedical domain. To classify keyphrases, we study the UMLS semantic network and propose a possible linking between the keyphrase types and the UMLS semantic groups. Based on this semantic linking, we create a dictionary for each keyphrase type. Then, a feature indicating if a token is found in one of these dictionaries is incorporated to feature set used by the BANNER tool. The final results on the test dataset show that our system still needs to be improved, but the conditional random fields and, consequently, the BANNER system can be used as a first approximation to identify and classify keyphrases.

1 Introduction

In the era of big data, as it could not be otherwise, an enormous amount of scientific articles is available. Although during the last few years search engines have provided significant improvements in information access, researches still have to spend much time exploring the huge number of articles published in their research fields. This laborious task could be reduced if search engines were able to answer common questions such as: *which studies have dealt with a specific TASK?, which studies have explored a PROCESS? or which studies have employed such MATERIAL?*. The automatic detection and classification of keyphrases (which describe tasks, processes and materials) as well as the extraction of their relations between them from scientific articles can support to find the answers to the previous questions. This task is very important, but has hardly been explored at the present time (Augenstein and Sgaard, 2017).

The ScienceIE task at SemEval 2017 (Augenstein et al., 2017) aims the automatic extraction of keyphrases and their relations from scientific publications. The task consists of three subtasks: (1) the subtask A is focused on the identification of the keyphrases in a given article; (2) the subtask B is focused on the classification of keyphrases by one of the following types: MATERIAL, TASK, and PROCESS; and (3) the subtask C deals with the classification of the relationships between keyphrases by one of the following types: HYPONYM-OF, SYNONYM-OF, and NONE. For the evaluation of the task, the organizers have defined three different scenarios, which the participating teams can choose to submit their outputs. For example, in scenario 1, the test dataset consists of plain texts without any annotation and participants can submit their outputs for all subtasks; for the scenario 2, the texts in the test dataset also include the annotation of keyphrases with their offsets in texts, but without providing their types. In this case, the teams can only submit their outputs to the subtask B and C. Finally, in scenario 3, which is only valid for the subtask C, test documents contain the keyphrases annotated with their offsets and their types.

In this paper, we describe the participation of the group LABDA in the subtasks A and B. Our approach for identifying and classifying keyphrases from scientific articles combines the use of the BANNER tool (Leaman et al.,

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 947–950,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

2008) and the UMLS semantic network (McCray, 1989)[1]. The paper is organized as follows. Section 2 describes our approach. Experiments, results, and discussion are described in Section 3. Finally, the paper is concluded and future work is proposed in Section 4.

2 Combining the BANNER tool and UMLS to identify and classify keyphrases

This section describes the system proposed by the LABDA group for participation for subtask A and B. BANNER is a named entity recognition (NER) system, which is based on conditional random fields (CRF). CRF is a class of statistical modelling method for sequence labelling and makes use of a rich set of lexical and syntactic features. Based on successful results provided by this approach for NER in the biomedical domain (Krallinger et al., 2015; Wei et al., 2015; Segura-Bedmar et al., 2015), in this paper, we explore the recognition of keyphrases as a sequence labeling problem by using the BANNER tool. This tool is designed to maximize domain independence and allows to recognize named entities from different domains.

BANNER has a 3-stage pipeline, whose input is a sentence. The first process splits the sentence into tokens. Then, each token is represented by a set of features: lemma, prefixes and suffixes of up to 2, 3 and 4 characters, bigrams and trigrams, as well as a series of regular expressions to normalize numeric values. Moreover, the word-class feature also normalizes the possible forms of a token based on their letters by converting upper-case letters to 'A', lower-case ones to 'a' and numbers to '0'.

We also incorporate a new feature that indicates if the token is found in a given dictionary. In particular, for each type of keyphrase (TASK, MATERIAL, PROCESS), we define a dictionary based on the semantic groups of UMLS. To create these dictionaries, we studied in depth the UMLS semantic network and proposed the links between the keyphrase types and the UMLS semantic groups shown in Table 1. Then, we traverse the UMLS methatesaurus and their terms are stored in their corresponding dictionary based on the classification shown in Table 1. The UMLS semantic groups as well as their semantic types can be found at https://semanticnetwork.nlm.nih.gov/. Thus, if a token is found in one of the three dictionaries, the feature is set to the name of the dictionary.

To label tokens, we try with different IOB tagging schemas (O=outside, B=beginning of an entity, I=inside of an entity, E=end of an entity, W=a single entity). Finally, a CRF model is trained using the features for each token from the training data. We consider the three types of keyphrases as the three possible types of entities to be recognized by BANNER. Thus, our approach performs both subtasks, identification and classification, as one only process. We train a single model for the three types.

3 Evaluation

As said before, we have only participated in the subtasks A (identification) and B (classification). That is, our experiments are performed on scenarios 1 and 2. Our approach for identification is evaluated on the scenario 1, where the test documents do not contain any annotation. Our approach for classification is evaluated on the scenario 2, where texts include the offsets of the keyphrases, but not their types. Actually, as said above, we use the same system to identify and classify keyphrases.

For evaluating the classification task on the scenario 2, we take the list of keyphrase mentions (without their types) provided as input of this scenario and compare it with the output of the BANNER tool, which was trained to classify the three types of keyphrases: MATERIAL, TASK and PROCESS. If the mention was classified as a keyphrase by BANNER, we return the type provided by BANNER. If the mention was classified with the tag O by BANNER (that is, outside token), our system was not be able to classify it. However, if it is actually a keyphrase (because it is in the input of the scenario 2), we decide to classify it with the most frequent type (PROCESS). The keyphrases classified by BANNER, but not found in the input of the scenario 2, are ignored.

Table 2 shows the results on the development set for each keyphrase type: MATERIAL, TASK and PROCESS. We tried with different variations of the IOB schema and with different combinations of the dictionaries defined from the UMLS semantic network.

The best results are achieved for the type MATERIAL with an F1 of 35.33%, followed by PRO-

[1] https://semanticnetwork.nlm.nih.gov/

Type	**UMLS groups**
MATERIAL	ANAT:Anatomy, CHEM:Chemicals and Drugs, GENE:Genes and Molecular sequences, LIVB:Living beings, OBJC:Objects, CONC:Concepts and Ideas
PROCESS	ACTI:Activities and Behaviors, DISO(T050:Experimental Model of Disease), OCCU:Occupations, PROC:Procedures, CONC:Concepts and Ideas (T185:Classification, T089:Regulation or Law, T170:Intellectual product, T171:Language, T080:Qualitative Concept, T081:Quantitative Concept, T079:Temporal Concept)
TASK	DISO:disorders(all concepts except those classified with the semantic type T050), PHEN:phenomena, PHYS:physology

Table 1: Linking between keyphrase types and UMLS semantic groups.

Type	IOB schema	dictionaries	**Precision(%)**	**Recall(%)**	**F-Measure(%)**
MATERIAL	IO	NO	59.45	23.48	33.67
	IO	Material	**59.74**	**25.08**	**35.33**
	IOB	Material	61.29	23.66	34.14
	IOBEW	Material	62.73	23.66	34.36
TASK	IO	NO	**18.51**	**7.29**	**10.47**
	IO	Task	16.12	7.29	10.05
	IOB	Task	17.02	5.83	8.69
	IOBEW	Task	19.51	5.83	8.98
PROCESS	IO	NO	39.93	25.82	**31.36**
	IO	Process	40.00	25.60	31.22
	IOB	Process	41.76	24.06	30.53
	IOBEW	Process	40.87	22.73	29.21

Table 2: Results on the development set for each type of keyphrase (scenario 1).

CESS with a 31.36% of F1. The system achieves the worst results for TASK (F1=10.47%). We study the list of keyphrases in the training dataset in order to know how many words form each keyphrase type. We observe that 41% of MATERIALS are formed by a single word, 32% of them are formed by two words, and the rest of MATERIALS (27%) are phrases with more than two words. Therefore, we can claim that a high percent of MATERIALS could be named entities. For the type of PROCESS, more than half are formed by one or two words (that is, they can be named entities), while the rest (48%) are phrases with more than two words. However, many of TASKS (74%) have three or more words. Thus, while CRF models have succeeded in the task of NER from the biomedical texts, the sequence labelling approach may not be the most appropriate for identifying keyphrases when they are formed by three or more words. Another possible cause of low results for TASK could be that the semantic linking between TASK and the UMLS semantic groups, which we defined for this work, is not suitable for the task.

Regarding the different settings, the IO schema seems to achieve the best results for the three keyphrase types. Only the use of the dictionary for MATERIALS achieves a significant improvement, while the rest of dictionaries do not seem to improve the performance. We proposed the three submitted runs based on the results on the development set.

The final results of the task show that our system achieved an F1 of 0.33 for the subtask A and 0.23 for the subtask B, when the system is evaluated on the scenario 1 (without annotations). As expected, our results are better for the subtask B when it is classified on the scenario 2 (the offsets of the keyphrases are provided for the participants), achieving an F1 of 0.51.

4 Conclusion

In this paper, we study if a sequence labelling approach is appropriate for the tasks of identification and classification of keyphrases from scientific publications. In particular, we use the BANNER tool, based on a CRF model and a rich set of lexical features. To classify the keyphrases, we study the UMLS semantic network and propose a linking between the keyphrases types and the UMLS semantic groups. Then, we extend the BANNER tool by incorporating a new feature that indicates if the token is found in one of the three dictionaries built from UMLS. Results are modest yet suggest promise for MATERIAL and PROCESS. As a future work, we plan to explore other dictionaries for the areas of computer science, physics and material science. Moreover, we plan to study an approach based on deep learning methods. Because keyphrases are usually longer phrases than named entities, we would like to create a phrase embedding model capable of measuring the similarity between keyphrases. This approach could be a solution to deal with nested keyphrases.

Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (eGovernAbility-Access project TIN2014-52665-C2-2-R).

References

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.

Isabelle Augenstein and Anders Sgaard. 2017. Multi-Task Learning of Keyphrase Boundary Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. Chemdner: The drugs and chemical names extraction challenge. *Journal of cheminformatics* 7(1):S1.

Robert Leaman, Graciela Gonzalez, et al. 2008. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific symposium on biocomputing*. volume 13, pages 652–663.

Alexa T McCray. 1989. The umls semantic network. In *Proceedings of the Annual Symposium on Computer Application in Medical Care.*. American Medical Informatics Association, pages 503–507.

Isabel Segura-Bedmar, Vıctor Suárez-Paniagua, and Paloma Martınez. 2015. Combining conditional random fields and word embeddings for the chemdner-patents task. In *Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain*. pages 90–93.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015. Overview of the biocreative v chemical disease relation (cdr) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*. pages 154–166.