Classification of Keyphrases from Scientific Publications using WordNet and Word Embeddings

Davide Buscaldi [1], **Simon David Hernandez** [1], **Thierry Charnois** [1]

Laboratoire d'Informatique de Paris Nord, CNRS (UMR 7030)
Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France
{davide.buscaldi,hernandez-perez,thierry.charnois}@lipn.univ-paris13.fr

ABSTRACT. *The ScienceIE task at SemEval-2017 introduced an epistemological classification of keyphrases in scientific publications, suggesting that research activities revolve around the key concepts of* process *(methods and systems),* material *(data and physical resources) and* task. *In this paper we present a method for the classification of keyphrases according to the ScienceIE classification, using WordNet and word embeddings derived features. The method outperforms the best system at SemEval-2017, although our experiments highlighted some issues with the collection.*

RÉSUMÉ. *Dans le contexte du challenge ScienceIE à SemEval-2017, ses organisateurs ont introduit une classification des phrases clés dans les publications scientifiques. Selon leur hypothèse, les activités de recherche tournent autour des concepts clés de "*process*" (methodes, systèmes), "*material*" (ressources matériellles, données, produits) et "*task*" (problèmes, activités à poursuivre). Dans cet article, nous présentons une méthode pour la classification des phrases clés selon la classification donné par ScienceIE, en utilisant des caractéristiques dérivées à partir de WordNet et de "word embeddings". La méthode proposée dépasse le meilleur système au SemEval-2017; toutefois, nos expériences ont mis en évidence certains problèmes d'annotation avec la collection.*

KEYWORDS: *Information Extraction, Text Mining on Scientific Literature, Keyphrase extraction.*

MOTS-CLÉS : *Extraction de mots clés, extraction d'information, fouille de textes sciéntifiques.*

1. Introduction

Nowadays, the number of scientific publications is continuously growing, in all disciplines. According to (Bjork *et al.*, 2009), 1.35 million articles were published in indexed journals in the single year 2006, and the growth rate in the number of scientific publications has been estimated by (Larsen, Von Ins, 2010) to be between 2.2% and 9% for journals and between 1.6% and 14% for conferences (depending on the disciplines) in the decade 1997-2007. It is becoming more and more difficult to search some informations required to write scientific papers, review the work of other researchers, or looking for expert. Usually this kind of search involves checking the originality of an idea or a method. Current search engines dedicated to the exploration of scientific literature, such as Google scholar[1] and Scopus[2], are based on text-based search, author and citation graphs. Recent works from the semantic web, scientometry and natural language processing communities have been aimed to improve the access to scientific literature (Osborne, Motta, 2015; Wolfram, 2016), and some initiatives have been started to gather researchers around this problem, like the SAVE-SD[3] workshops and the ScienceIE task (Augenstein *et al.*, 2017) at SemEval2017[4].

In particular, the ScienceIE task was focused on extracting keyphrases and relations between them, relying on the hypothesis that the ability of correctly recognising these semantic items in text will help in tasks related to the process of scientific publishing, such as to recommend articles to readers, highlight missing citations to authors, identify potential reviewers for submissions, and analyse research trends over time. The hypothesis made by the organizers is that some concepts, notably PROCESS, TASK and MATERIAL, are cardinal in scientific works, since they allow to answer questions like: “which papers addressed a *Task* using variants of some *Process* ?". In their vision, *Processes* correspond to methods and equipments and *Materials* to corpora and physical items. An example of text labelled with these concept is shown in Figure 1 (Augenstein *et al.*, 2017).

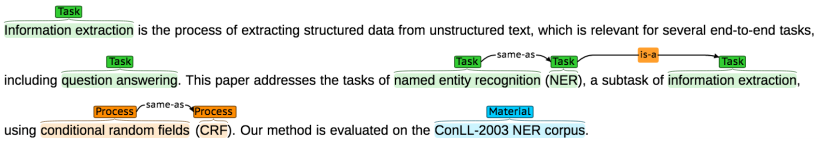


Figure 1. Example of annotation of a scientific document with ScienceIE concepts and relations.

In this paper, we propose a method to classify candidate terms into the three categories defined in the ScienceIE challenge, using surface features combined with WordNet-based features and word embeddings. This method outperforms the best

1. http://scholar.google.com
2. http://www.scopus.com
3. http://cs.unibo.it/save-sd/2017/index.html
4. https://scienceie.github.io

result obtained at ScienceIE. In the remainder of this paper, we describe the method and the features used in Section 2, then we show the obtained results in Section 3, and finally we draw some conclusions in Section 4

2. Proposed Method

The method we propose in this paper is based on Support Vector Machines (SVM), in particular the nu-SVM implementation by (Chang, Lin, 2011). SVMs are well known maximum margin classifiers; we chose them because of their robustness with regard to problems with a large number of features. Please note that the method we are describing in this paper only shares part of the WordNet-based features with the one we used to participate to the task (Hernandez *et al.*, 2017).

2.1. Base Features

The base features are constituted by all the {3,4,5}-prefixes and suffixes of keyphrases that appeared in the training set with frequency greater than 10. For instance, from the keyphrase “information extraction" we can identify the following features: *inf, info, infor* as prefixes and *ction, tion, ion* as suffixes. Together with the prefixes and suffixes, we have considered the following features:

– capitalization of the keyphrase (binary);

– uppercase ratio, calculated as number of uppercase characters divided by number of characters in the keyphrase;

– number of digits in the keyphrase;

– number of dashes;

– number of words.

2.2. WordNet-based Features

WordNet (Miller, 1995) is a well known lexical database for the English language. In WordNet, word senses are represented as *synsets*, or “set of synonyms", which may be connected to other synsets by some relationship. Some of the most common relationships are meronymy (part-of) and hyperonymy (is-a). We define a *synpath* as the list of synsets connecting a sense of a target word to the root of the hierarchy in WordNet, following the hyperonymy relation. In Figure 2 we show the synpaths corresponding to the three senses of the word *extraction* in WordNet 3.0. The definitions of the senses are as follows:

1. extraction#1:the process of obtaining something from a mixture or compound by chemical or physical or mechanical means;

2. extraction#2: properties attributable to your ancestry;

3. extraction#3: the action of taking out something (especially using effort or force).

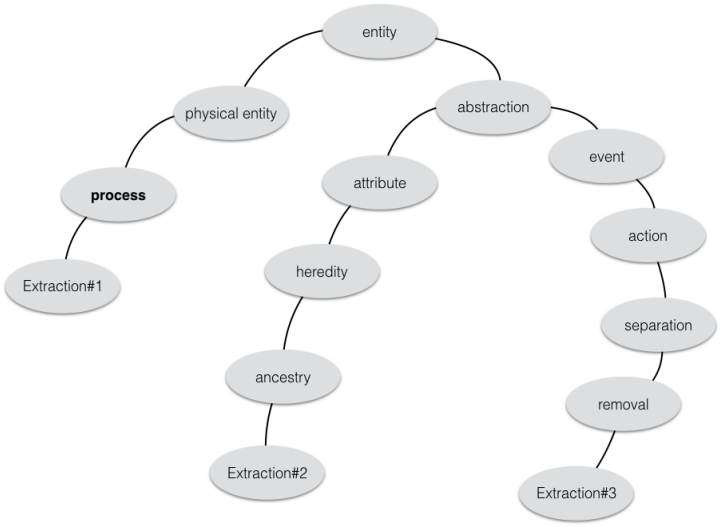


Figure 2. Example of synpaths for the word "extraction" in WordNet 3.0 (simplified by removing some synsets).

From Figure 2 it can be observed that the synset *process* is in the synpath (process, physical_entity) of `extraction#1`, which seems an important clue to classify this keyword as a PROCESS, according to the ScienceIE classification. Therefore, we supposed that synpaths can be effectively used as features to predict the category of a keyword. Given the number of synsets in WordNet (more than $117,000$), we opted to select only a subset of those synset, in particular by limiting the scope to the synsets that are particularly distinctive for each of the three classes. We calculated, on the training corpus of ScienceIE, the probability $p(s|C)$ for each synset with respect to class C. Subsequently, we ordered in decreasing order, for each class, the synsets according to the difference $p(s|C_i) - \frac{p(s|C_j)+p(s|C_k)}{2}$. We show in Table 1 the most distinctive synsets for each category. The semantic correlation between the MATERIAL category and its distinctive synsets is particularly evident.

Table 1. Top 5 distinctive synsets for each category.

$PROCESS$	$MATERIAL$	$TASK$
$psychological\_feature.n.01$	$physical\_entity.n.01$	$science.n.01$
$event.n.01$	$object.n.01$	$possession.n.02$
$abstraction.n.06$	$whole.n.02$	$natural\_science.n.01$
$act.n.02$	$artifact.n.01$	$question.n.02$
$cognition.n.01$	$matter.n.03$	$subject.n.01$

We arbitrarily selected the top 20 distinctive synsets for each category and we used them to extract some binary features[5]. These features are true for a token if they are present in any of the hypernym paths connecting the noun synsets to the root synset. Note that these features were added only for the nouns, since there is no hierarchy for the other lexical categories (if we exclude verbs, whose hierarchy is in any case very shallow, if compared to nouns). If the keyphrase was composed by more terms, then we searched the synpaths for the rightmost noun in the keyphrase.

2.3. Word Embeddings Features

Word embeddings, as introduced by (Bengio *et al.*, 2006), are vector representations of words that capture a certain number of syntactic and semantic relationships, generated with neural networks. In this work, we used the pre-trained vectors trained on 100 billion words from a Google News dataset (Mikolov *et al.*, 2013). The vocabulary size is 3 million words and the vector length is 300. One of the problem we had to solve to include embeddings was to deal with keyphrases composed by more than one term: vectors are linked to single words (or, in some cases to compound words or terms). (De Boom *et al.*, 2016) showed that it's possible to exploit the properties of embeddings to represent sentences with the average, the max, or the min of the vectors of the composing words. We chose to use the max.

3. Experiments and Results

We carried out our experiments on the ScienceIE dataset[6], consisting in a set of 450 articles collected from ScienceDirect, distributed among the domains Computer Science, Material Sciences and Physics. The training set consists of 350 documents, while the test set consists of 100 documents. The organizers also distributed 50 documents as development set, but we didn't use these data. The task consisted in three sub-tasks:

– A) Mention-level keyphrase identification;

– B) Mention-level keyphrase classification;

– C) Mention-level semantic relation extraction between keyphrases with the same keyphrase types. Relation types are HYPONYM-OF and SYNONYM-OF.

We consider in this paper only sub-task B), while for the evaluation, we refer to the evaluation scenario in which the text is manually annotated and keyphrase boundaries are given (Augenstein *et al.*, 2017).

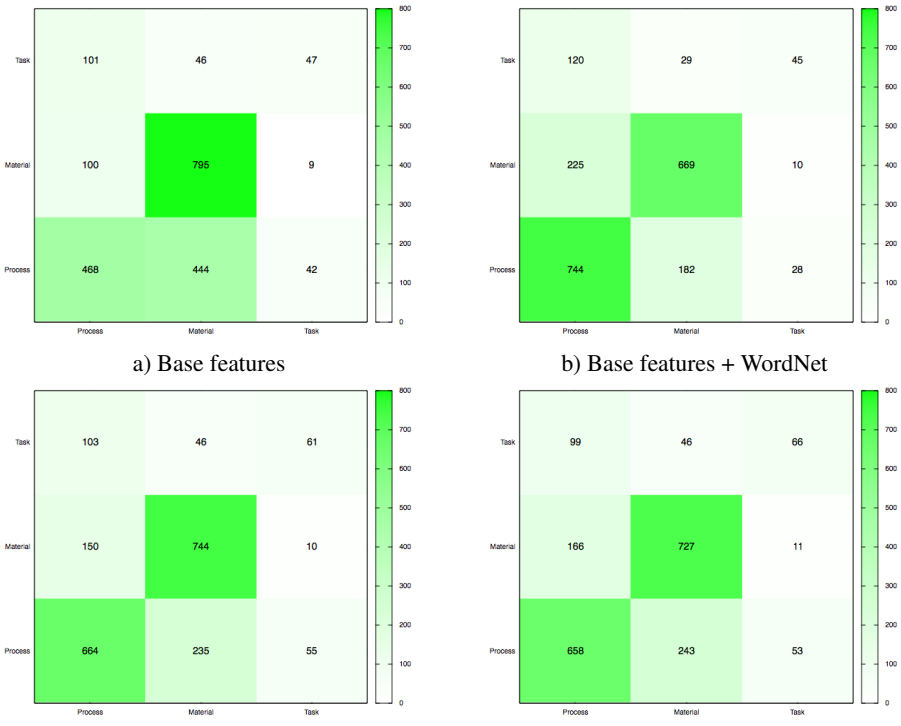
5. full list: https://github.com/snovd/corpus-data/blob/master/SemEval2017Task10/SynsetsRelatedToTrainingData.txt
6. https://scienceie.github.io/resources.html

In Table 2 we show the results obtained with different combination of features, compared to the best system at the SemEval 2017 ScienceIE (B subtask, with the evaluation scenario where the keyphrase boundaries are given).

Table 2. F-measure obtained for each test configuration, compared with the best system at SemEval 2017 ScienceIE.

	$PROCESS$	$MATERIAL$	$TASK$	all
$Base$	.577	.726	.322	.619
$Base + WN$	**.728**	.750	.325	.700
$All features$	.710	**.778**	.381	**.716**
$Base + Embeddings$	.701	.764	**.407**	.701
$best@SemEval2017$	.660	.760	.280	.670

From these results and the confusion matrices in Figure 3 it can be seen that WordNet features are very helpful in discriminating the MATERIAL from the PROCESS class, while the word embeddings features had a positive impact on the TASK class, which was the most difficult one.



a) Base features

b) Base features + WordNet

c) Base features + WordNet +embeddings

d) Base features + embeddings

Figure 3. Confusion matrices for the 4 configurations tested.

The confusion matrices show also that TASK is often confused with PROCESS, which in turn seem to be too predominant, indicating a bias in the collection towards this class. An analysis of the annotated collection showed certain inconsistencies in annotations that may be at the origin of the errors: for instance, in file `2212667814000732.ann`, we found a conflicting annotation for “synthetic assessment method": alone is annotated as PROCESS, but the keyphrase “synthetic assessment method based on cloud theory" is annotated as TASK, which seems odd. In file `S2212671612002351.ann`, we found that “position estimation method" is labelled as TASK, when it should instead be a process.

4. Conclusions

We developed a method to classify keyphrases into a predefined set of categories provided by the ScienceIE task at SemEval-2017. This method integrates external knowledge, acquired either from an existing resource like WordNet or learned from a large corpus of text and encoded using word embeddings, as features for a SVM classifier. The obtained results outperform those obtained by the best system presented at SemEval-2017. Our method presents margins of improvement, since some parameters were chosen arbitrarily and further investigation is needed to discover the optimal ones. We plan to exploit the domain of the document as an additional feature, supposing that keyphrase styles may vary depending on the domain. The experiments also highlighted some problems with the ScienceIE collection: on one side one of the classes seems underrepresented and our analysis exposed a certain number of annotation errors which may require a manual re-annotation.

Acknowledgements

This work has been partly supported by the program “Investissements d’Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL).

References

Augenstein I., Das M. K., Riedel S., Vikraman L. N., McCallum A. (2017, August). SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the international workshop on semantic evaluation.* Vancouver, Canada, Association for Computational Linguistics.

Bengio Y., Schwenk H., Senécal J.-S., Morin F., Gauvain J.-L. (2006). Neural probabilistic language models. In D. E. Holmes, L. C. Jain (Eds.), *Innovations in machine learning: Theory and applications*, pp. 137–186. Berlin, Heidelberg, Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/3-540-33486-6_6

Bjork B.-C., Roos A., Lauri M. (2009). Scientific journal publishing: yearly volume and open access availability. *Information Research: An International Electronic Journal*, Vol. 14, No. 1.

Chang C.-C., Lin C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1–27:27. (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm)

De Boom C., Van Canneyt S., Demeester T., Dhoedt B. (2016, September). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recogn. Lett.*, Vol. 80, No. C, pp. 150–156. Retrieved from https://doi.org/10.1016/j.patrec.2016.06.012

Hernandez S. D., Buscaldi D., Charnois T. (2017, August). LIPN at SemEval-2017 Task 10: Filtering Candidate Keyphrases from Scientific Publications with Part-of-Speech Tag Sequences to Train a Sequence Labeling Model. In *Proceedings of the international workshop on semantic evaluation.* Vancouver, Canada, Association for Computational Linguistics.

Larsen P. O., Von Ins M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, Vol. 84, No. 3, pp. 575–603.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Miller G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41.

Osborne F., Motta E. (2015). Klink-2: Integrating multiple web sources to generate semantic topic networks. In *Proceedings of the 14th international conference on the semantic web - iswc 2015 - volume 9366*, pp. 408–424. New York, NY, USA, Springer-Verlag New York, Inc. Retrieved from http://dx.doi.org/10.1007/978-3-319-25007-6_24

Wolfram D. (2016). Bibliometrics, information retrieval and natural language processing: Natural synergies to support digital library research. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)*, pp. 6–13. Retrieved from http://ceur-ws.org/Vol-1610/paper1.pdf