Representation of complex terms in a vector space structured by an ontology for a normalization task

Arnaud Ferré [1,2], Pierre Zweigenbaum[2], Claire Nédellec[1]
[1]MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2]LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
arnaud.ferre@universite-paris-saclay.fr

Abstract

We propose in this paper a semi-supervised method for labeling terms of texts with concepts of a domain ontology. The method generates continuous vector representations of complex terms in a semantic space structured by the ontology. The proposed method relies on a distributional semantics approach, which generates initial vectors for each of the extracted terms. Then these vectors are embedded in the vector space constructed from the structure of the ontology. This embedding is carried out by training a linear model. Finally, we apply a cosine similarity to determine the proximity between vectors of terms and vectors of concepts and thus to assign ontology labels to terms. We have evaluated the quality of these representations for a normalization task by using the concepts of an ontology as semantic labels. Normalization of terms is an important step to extract a part of the information contained in texts, but the vector space generated might find other applications. The performance of this method is comparable to that of the state of the art for this task of standardization, opening up encouraging prospects.

1 Introduction

A lot of biomedical or biological knowledge is in a non-structured form, such as that expressed in scientific articles (Kang et al., 2013). For experts from these fields, the substantial increase in the specialized literature has created a significant need for automatic methods of information extraction (Ananiadou and McNaught, 2006). The task of normalization is one of the main tasks to respond to this need.

Normalization consists in standardizing terms (single- or multi-word) extracted from texts by linking them to non-ambiguous references, such as entries from existing knowledge bases. Concepts from an ontology can be used to represent these references in a formal and structured way. Term and their relationships carry a lot of the knowledge contained in texts, thus successful term identification is a key to getting access to the information (Krauthammer and Nenadic, 2004).

Standardization encounters several difficulties, such as the significant variability of the form of the terms, whether they are represented by one word (e.g. “child” / “kid” or “accommodation” / ”home”, etc.) or by several (e.g. “child” / ”little boy” or “accommodation” / ”dwelling place”, etc.) (Nazarenko et al., 2006). Multiword terms, which have varied morphosyntactic structures and complex imbrications (mainly complex noun phrases), are particularly difficult to normalize (e.g. only with a different syntactic organization: “breast cancer” / “cancer of the breast”). In the literature, such as scientific articles in life sciences, complex noun groups are abundant (Maniez, 2007). An approach based on the similarity of form between term and semantic label appears limited to perform this task (Golik et al., 2011), because the form of the labels of the concepts is not necessarily close to the form of the terms to be annotated. Another difficulty arises from the large number of ontology concepts, making a supervised classification approach costly in manual annotation (e.g. over 2,000 categories for example in the ontology of bacterial habitats OntoBiotope (Bossy et al., 2015)).

An alternative approach is to calculate the semantic proximity between terms by distributional semantics. It is an approach based on the correlation between the similarity of meaning and the distribution similarity of semantic units (word, combination of words, sentence, documents, ...) (Firth, 1957; Harris, 1954). A semantic unit can then be represented by a vector: it is constructed from the context information in which the semantic unit is found. The proximity of vectors in

Proceedings of the BioNLP 2017 workshop, pages 99–106,
Vancouver, Canada, August 4, 2017.

this space can be transposed to a semantic proximity (Fabre and Lenci, 2015). Today, there are many methods for generating such vector spaces, such as Word2Vec (Mikolov et al., 2013), but they usually focus on massive data sets (Fabre et al., 2014) in which information is often repeated.

The question is: how to use distributional semantics to normalize terms by an ontology? In other words how to relate distributional information to the categories of ontology? In the context of specialized literature, we often deal with relatively small corpora and a large number of semantic categories.

We propose an original method in which we represent complex terms based on word embedding, embed the ontology in a vector space, and learn a transformation from term vectors to concept vectors. Then, this transformation is used to determine the most suitable concept for an input term.

2 Material

The data used are those of the Bacteria Biotope categorization task (Task 3) of the 2016 BioNLP Shared Task (Deléger et al., 2016). The documents are references from MEDLINE, composed of titles and abstracts of scientific articles in the field of biology. The task consists in assigning a category from the OntoBiotope ontology to given corpus terms related to bacterial habitats. The corpus is divided into three subparts: the training corpus, the development corpus and the test corpus. In the training and development corpus, the categories of terms are given: they have been used to train our method. The terms from the test corpus are those which categories have to be predicted: it is the corpus used to evaluate our method for the task of normalization. The entities of each of these corpora have been manually annotated. Table 1 provides a summary of their characteristics:

	Train	Dev.	Test	**Total**
Documents	71	36	54	**161**
Words	16,295	8,890	13,797	**38,982**
Entities	747	454	720	**1,921**
Distinct entities	476	267	478	**1,125**
Semantic cat.	825	535	861	**2,221**
Distinct cat.	210	122	177	**329**

Table 1: Descriptive statistics for the Bacteria Biotope corpus ("cat." = categories, "Dev." = development corpus)

In addition to this corpus, an extended corpus of the same domain is used to generate vector representations of each word. It is composed of approximately 100,000 sentences (4,800,000 words) from titles and abstracts of scientific articles in the field of biology available on PubMed. This represents a relatively small size corpus, which contains a majority of words with a low frequency of occurrence (cf. Table 2). Other corpus, larger and/or more general could be used, also direct words embedding as the one released by BioASQ (Pavlopoulos et al., 2014). Nevertheless, the very accurate domain of the used extended corpus and its desired small size seemed to be more adapted.

Repeated >2	72,412	35%
Repeated 2 times	31,569	15%
Not repeated	105,364	50%
Words (without stopwords)	209,345	100%

Table 2: Descriptive statistics of extended corpus

3 Method

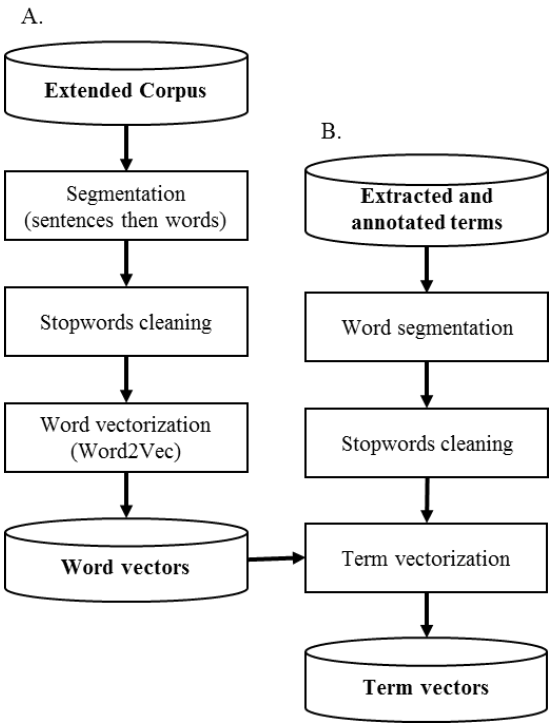

Figure 1: A. Process to create word vectors. B. Process to create term vectors.

3.1 Word vectors

The vector space of the terms (VST) is obtained by generating a vector for each word of the extended corpus and the Bacteria Biotope corpus. For this, we used the Word2Vec tool

(Mikolov et al., 2013), taking as context of a word, a list containing all the words of their sentence. To have enough training data for the generation of meaningful word vectors, and also to avoid taking into account typos or errors, it is usually advisable to use Word2Vec without the infrequent words appearing only once or twice throughout the corpus. But our corpus contains many words of interest with a low frequency, so we choose not to apply this frequency threshold. After some performance tests, the dimension 200 was selected for the output vectors (cf. Figure 1A), which is of the same order of magnitude as what is usually advised (Mikolov et al., 2013).

3.2 Term vectors

To compute the vector representations of the multiword terms (cf. Figure 1B), segmenting them into words is the first step. For each word, which is not a stopword, the vector calculated by Word2Vec is used. Then the vector of the multiword term is obtained by averaging the vectors of the words which compose it:

$$v_{t_k} = \sum_{i=1}^{n_k} v_{m_i^k}/n_k \qquad (1)$$

where v_{t_k} is the associated vector of the term t_k, n_k is the number of words (without stopwords) of the term t_k, $v_{m_i^k}$ is the vector of the word m_i^k from our Word2Vec computation, and the term t_k is such that :

$$\forall i \in [1, n_k], m_i^k \in t_k \qquad (2)$$

Even if it is not the aim of this paper, future works could test other methods.

3.3 Concept vectors

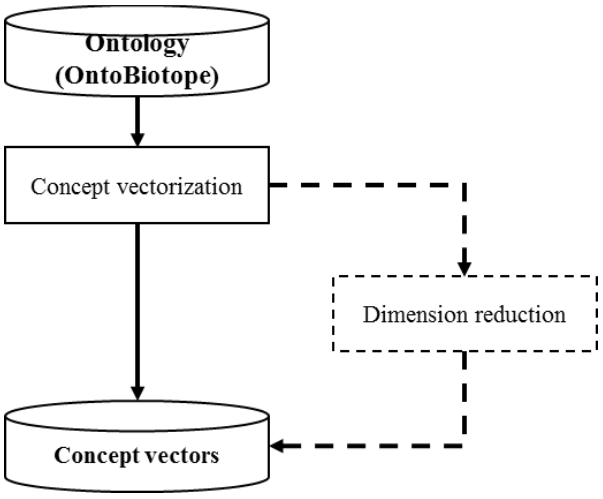

Figure 2: Process to create concept vector

To construct the concept vectors and thus a vector space of an ontology (VSO), null vectors with as many dimensions as the number of concepts in the ontology are initialized. Each value of the vector is thus related to one of the concepts of the ontology, which is set to 1 for the considered concept. The value is also 1 if the current axis is related to a concept which is an ancestor of the considered concept, and 0 otherwise:

$$v_{c_k} = \left(w_{c_k}^0, \ldots, w_{c_k}^i, \ldots, w_{c_k}^n \right) \qquad (3)$$

where v_{c_k} is the vector related to the concept c_k, n is the number of concepts in the ontology and $w_{c_k}^i$ is the value of vector v_{c_k} for the axis i, such as:

$$w_{c_k}^i = \begin{cases} 1, if\ i = k \\ 1, if\ c_i\ parent\ of\ c_k \\ 0, \text{otherwise} \end{cases} \qquad (4)$$

This representation has the advantage of preserving the similarity arrangement (with cosine distance) expected between the concepts (cf. Figure 3 and Table 3): a concept is more similar to his children and his parents.

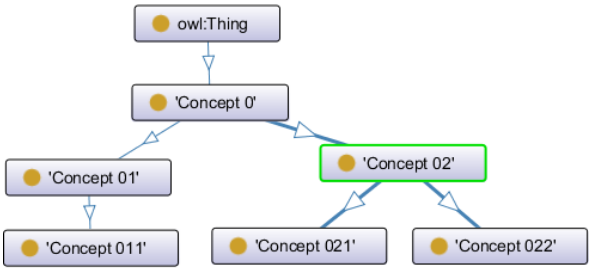

Figure 3: Abstract ontology representation (displayed by Protégé)

Concept 02	Similarity
Concept 02	1,0000
Concept 021	0,8165
Concept 022	0,8165
Concept 0	0,7071
Concept 01	0,5000
Concept 011	0,4082

Table 3: Cosine distances between concepts of an abstract ontology (cf. Figure 3)

We can notice that the dimension of the generated VSO is the number of concepts of the ontology (e.g. more than 2,000 for the OntoBiotope ontology). It is a high dimension in comparison to the VST but concept vectors are very sparse (with a maximum of 13 non-zero values in a vector) and they only contain binary values. Therefore, to make them more comparable to term vectors, we experimented with reducing the VSO to denser

representations in a lower-dimension space (cf. Figure 2). Two methods have been tested: Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS).

3.4 Training with general linear model

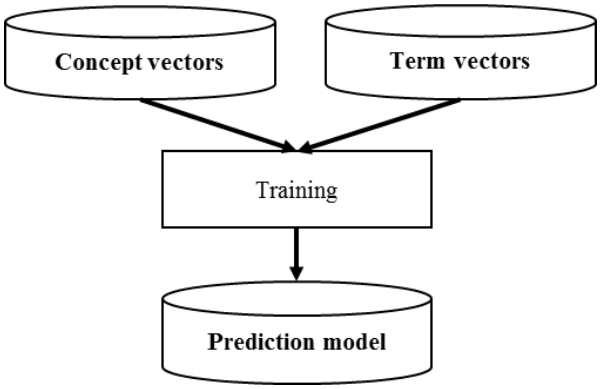

Figure 4: Training process to determine a transformation VST to VSO

The objective of the training step is to determine a transformation from VST to VSO, which minimizes all the distances between the vectors of terms resulting from this transformation and the vectors of the associated concepts. In this paper, a linear transformation is studied with the aim of keeping a strong similarity between the distribution of term vectors in the VST and the distribution of the projections in the VSO. Indeed, a non-linear transformation could strongly distort the resulting distribution to fit better to training data.

This training aims to obtain the best parameters to approximate this matrix equation:

$$Y = X.B + U \qquad (5)$$

where Y is a matrix resulting in a series of concept vectors, X is a matrix resulting in a series of term vectors (where the ith line of X is the vector of a term which has for category a concept which has for vector the ith line of Y), B is a matrix containing parameters that are usually to be estimated and U is a matrix containing noise following a multivariate Gaussian distribution. This training is performed on the training and development corpora (cf. Figure 4).

The obtained matrix enables us to design a linear transformation function then make it possible to predict new vectors associated with the terms of the test corpus expressed in the VSO:

$$f: \begin{pmatrix} VST \rightarrow VSO \\ v_{\text{term}} \rightarrow v'_{\text{term}} = f(v_{\text{term}}) \end{pmatrix} \qquad (6)$$

where v_{term} is a vector of term in the VST and v'_{term} is the resulting vector of the same term projected in the VSO. To satisfy the requirements of the evaluation task, the concept vector nearest to v'_{term} (as determined by cosine distance) is chosen as category for the annotated term (cf. Figure 5).

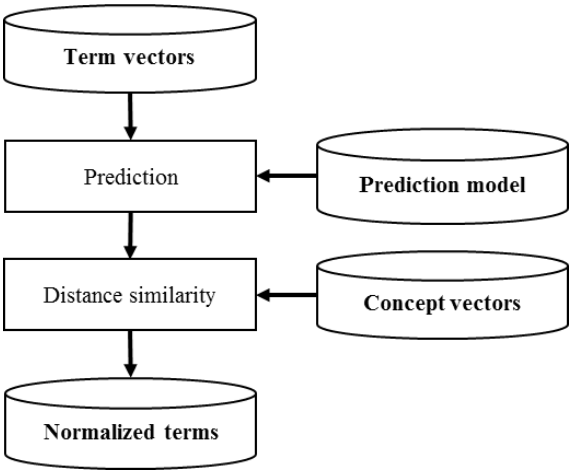

Figure 5: Process of predicting semantic categories associated with extracted terms

3.5 Evaluation

We evaluate the performance of our normalization method on the Bacteria Biotope normalization task of the BioNLP Shared Task 2016. The dataset was presented in Section 2. The predicted concepts identifiers are compared to the gold standard concepts according to the similarity measure of (Wang et al., 2007), with the weight parameter set to 0.65. The evaluation was performed by submitting our results to the evaluation server run at the BioNLP-ST 2016 challenge site.

4 Results

4.1 Normalization

Team	Similarity score
BOUN	0.6200
CONTES	**0.5968**
LIMSI	0.4380
Baseline	0.3217

Table 4: Results on the normalization task of BioNLP-ST 2016

We applied our concept normalization method to the test dataset of the Bacteria Biotope 2017 Task 3. We computed baseline results by assigning all terms to the concept "bacteria habitat", which is the root of the OntoBiotope ontology hierarchy. We also compared these results to those of the two teams who participated in this task of BioNLP-ST 2016. We report all results in Table 4. The baseline obtains a score of 0.3217. Our method

(CONTES - CONcept-TErm System) obtained a score of 59.68%, much higher than the baseline, and close to that of the top team (Tiftikci et al., 2016). This score is also significantly above the method of LIMSI (Grouin, 2016), which is based on a morphological approach.

4.2 Term vectors

In spite of the low frequency of occurrence of the words of the extended corpus (cf. Table 2), the resulting word vectors seem to have relatively satisfactory proximities, from the point of view of the semantic similarity of the associated terms. Moreover, the method used to compute vectors for complex terms also seems satisfactory, as illustrated Table 5.

cell	**Similarity**
HCE cell	0.9999
13C-labeled cell	0.9998
parietal cell	0.9989
Schwann cell	0.9965
CD8+ T cell	0.9770
PMN cell	0.9669
macrophage cell	0.9473

Table 5: Terms nearest to the term "cell"

It also appears that lexical variation can be overcome (cf. Table 6 and Table 7), which was one of the desired properties. Although more generally, it seems that terms with similar lexical forms are closer (Table 5).

Nevertheless, the co-occurrence of some words seems to cluster certain terms from different categories: two words appearing frequently in common contexts are then found close. This similarity persists when calculating multiword term vectors. This applies, for example, to the terms relating to fish and those relating to fish farms (cf. Table 8). These cases are less satisfactory because they do not differentiate between terms which should be annotated with different semantic categories (e.g. "fish" and "healthy fish" should be annotated by <OBT:001902: fish>, "fish farm" and "disease-free fish farm" by <OBT:000295: fish farm> and "fish farm sediments" by <OBT:000704: sediment>).

younger ones	**Similarity**
children less than five years of age	0.8087
children less than 2 years of age	0.8060
children less than two years of age	0.7995

Table 6: Terms nearest to the term 'younger ones'

seawater	**Similarity**
sediments	0.7696
sediment sample from a disease-free fish farm	0.7499
fish farm sediments	0.7342
subterranean brine	0.7320
lagoon on the outskirts of the city of Cagliari	0.7128
petroleum reservoir	0.7095
marine environments	0.7077
marine bivalves	0.6896
sediment samples from diseased farms	0.6870
urine sediments	0.6819
petroleum	0.6576
subterranean environment	0.6497
fresh water	0.6494
fresh water supply	0.6395
Seafood	0.6390
marine	0.6366

Table 7: Terms nearest to the term 'seawater'

fish	**Similarity**
fish farming	0.9875
fish farm	0.9170
disease-free fish farm	0.9124
fish farm sediments	0.8683
healthy fish	0.8145

Table 8: Terms nearest to the term 'fish'

4.3 Concept vectors

<OBT:001922: algae> sans ACP	**Similarity**
<OBT:001777: aquatic plant>	0.9258
<OBT:001895: submersed aquatic plant>	0.8571
<OBT:001967: seaweed>	0.8018

Table 9: Concepts nearest to the concept <OBT:001922: algae>

We can estimate the quality of the created concept vectors by observing the consistency between the proximity of two vectors and the similarity of their meanings. Table 9 and Figure 6 show the example of the 'algae' concept: the nearest neighbors of its vector are its father in the ontology, its sibling and the immediate descendant of its sibling.

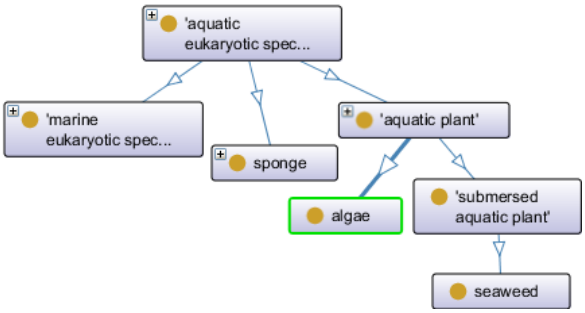

Figure 6: Taxonomy of concepts around concept "algae" (displayed by Protégé)

By comparing several examples, it seems that PCA does not modify the order of proximity of the concepts, but an increase in vector density can be observed (cf. comparison between Table 9 and Table 10).

<OBT:001922: algae> avec ACP	**Similarity**
<OBT:001777: aquatic plant>	0.9990
<OBT:001895: submersed aquatic plant>	0.9982
<OBT:001967: seaweed>	0.9943
<OBT:000372: sponge>	0.9303
<OBT:000269: marine eukaryotic species>	0.9303

Table 10: Concepts nearest to the concept <OBT:001922: algae> after a PCA with a final dimension of 100

4.4 Impact of the size of the VST

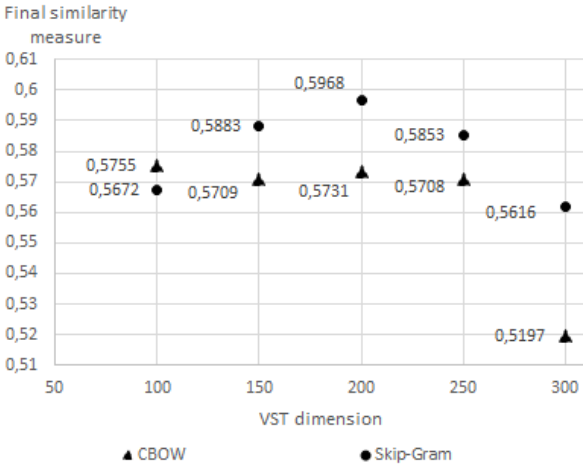

Figure 7: Comparison between CBOW and Skip-Gram architectures for the VST

Word2Vec allows the use of 2 different architectures to generate word vectors from a corpus: Continuous Bag Of Words (CBOW) and Skip-Gram. We tested the 2 architectures on different output vector sizes (cf. Figure 7). For vector spaces generated with a dimension between 100 and 250, the final scores appear to be relatively stable, especially with CBOW. Similarly, the score difference between the two architectures remains below 3%. Above a dimension of 250, there is a decrease in the score for the 2 architectures, with a greater slope for CBOW.

4.5 Impact of a dimension reduction on the VSO

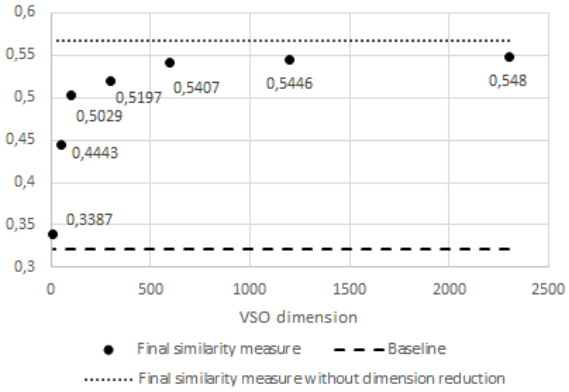

Figure 8: Evolution of performance depending on the final size of the VSO after reduction (here with a VST with 100 dimensions)

The VSO has a large dimension compared to the specific information that it contains (i.e. the ontology structure). This may present combinatorial but also theoretical difficulties: a linear projection of the VST on the VSO (with a higher dimension than the VST) should then only be performed on a subspace of the VSO. Thus, it theoretically limits the results. It was therefore interesting to study the impact of a reduction of the VSO size on the final score. We can then observe that a reduction PCA (with similar results with MDS) systematically decreases the score obtained when using a non-reduced VSO (cf. Figure 8).

Nevertheless, there is a level with relatively high performance (less than 3% below the score without reduction) which collapses below a certain dimension. This threshold might have a link with the number of concepts that have at least 2 distinct parents.

5 Discussion

To extend the interpretations derived from examples, it would be interesting to evaluate the overall quality of the generated vector spaces: vector spaces of words, terms, concepts as well as the final space containing the transformations of the vectors of the terms. We plan to perform this in further work.

One of the difficulties of the task is that in this normalization task, a term can be annotated by several distinct concepts of the ontology (e.g. "school age children with wheezing illness" should be annotated by the concept <OBT: 002307: pediatric patient> as well as the concept <OBT: 002187: patient with disease>). This difficulty is linked to the ontology of interest. In 2016, all participating systems of the task skip this difficulty, which is not anecdotal among the extracted terms.

6 Future work

For future work, it would be relevant to apply methods of global evaluation of the quality of the generated vector spaces. In particular, this would make it possible to evaluate the intermediate processes more thoroughly and to observe the impact of the modifications on their internal parameters more precisely. New methods could then be considered to improve outcomes. For example, it would certainly be positive to use a method of vector representation of an ontology that would generate a space with a smaller dimension while retaining the possibility of discerning the initial structure of the ontology. Similarly, the method used here to generate the VST vectors could be improved to take into account the syntactic context of the terms. This could solve the semantic similarity problems between "fish" and "fish farm" (cf. Table 8).

In the Bacteria Biotope normalization task, terms often have to be annotated with several concepts of the target ontology (for example, "children greater than 9 years of age who had lower respiratory illness" should be annotated by the concept <OBT: 002307: pediatric patient> and by the concept <OBT: 002187: patient with disease>). Having a completely defined ontology (i.e. containing all the concepts sufficient to annotate uniquely each possible extracted term - for example, a concept 'pediatric patient with disease' which is a subset of <OBT: 002307: pediatric patient> and of <OBT: 002187: patient with disease>) should improve the results. If such ontologies seem to be relatively rare in the biological domain, it might be interesting to start by automatically generating all the concepts equivalent to the intersection of the non-disjoint concepts to answer this problem. Nevertheless, if the concepts share many intersections between them or the disjoint property has not been formalized, the size of the generated ontology may pose combinatorial difficulties.

We addressed a task in which entities have already been detected in text. Since entity detection and terminology extraction methods have relatively acceptable performance, it would be useful to use them to extend the current task to an end-to-end concept detection and normalization system.

Finally, despite the inherent limitation of normalization methods based on word form similarity, these could nevertheless be used to carry out a pre-normalization of the corpus. As a result, one might consider using these annotations to drive the training part of the method (cf. 3.4 Training with general linear model) instead of using a manual annotation (i.e. a test corpora). Thus, this would transform this method into a fully unsupervised method.

7 Conclusion

The aim of this article was to propose an approach for the creation of vector representations for (complex or non-complex) terms in a semantic space. In addition, it aimed to propose a method capable of adapting to a small specialized corpus where the interest terms appear with a relatively low frequency. The most widely used methods currently generate vector spaces which meaning is difficult to interpret other than in terms of spatial proximity / semantic similarity. Our method seems to show that by combining relatively classical approaches, it is possible to use an ontology to generate vectors in a more interpretable vector space. The results are comparable to those of the state of the art, which seems to open up encouraging prospects. Beyond the standardization task, new efficient methods of generating interpretable vector spaces could apply to a number of further tasks.

Acknowledgments

This work is supported by the "IDI 2015" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

References

Sophia Ananiadou and John McNaught, editors. 2006. *Text mining for biology and biomedicine*. Artech House, Boston.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessières, and Claire Nédellec. 2015. Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task. *BMC bioinformatics*, 16(10):S1.

Louise Deléger, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016.

Cécile Fabre, Nabil Hathout, Franck Sajous, and Ludovic Tanguy. 2014. Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, pages 266–279.

Cécile Fabre and Alessandro Lenci. 2015. Distributional Semantics Today Introduction to the special issue. *Traitement Automatique des Langues*, 56(2):7–20.

J. R. Firth. 1957. The technique of semantics.

Wiktoria Golik, Pierre Warnier, and Claire Nédellec. 2011. Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence*, pages 37–39.

Cyril Grouin. 2016. Identification of Mentions and Relations between Bacteria and Biotope from PubMed Abstracts. *ACL 2016*:64.

Zellig S. Harris. 1954. Distributional Structure. WORD, 10(2–3):146–162, August.

Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. 2013. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 20(5):876–881, September.

Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526, December.

François Maniez. 2007. Prémodification et coordination : quelques problèmes de traduction des groupes nominaux complexes en anglais médical. *ASp*(51–52):71–94, December.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Adeline Nazarenko, Claire Nédellec, Erick Alphonse, Sophie Aubin, Thierry Hamon, and Alain-Pierre Manine. 2006. Semantic annotation in the alvis project. In *International Workshop on Intelligent Information Access (IIIA)*, page 5–pages.

Ioannis Pavlopoulos, Aris Kosmopoulos, and Ion Androutsopoulos. 2014. Continuous Space Word Vectors Obtained by Applying Word2Vec to Abstracts of Biomedical Articles. March.

Mert Tiftikci, Hakan Sahin, Berfu Büyüköz, Alper Yayıkçı, and Arzucan Ozgür. 2016. Ontology-based Categorization of Bacteria and Habitat Entities using Information Retrieval Techniques. *ACL 2016*:56.

J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen. 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, May.