Combining rule-based and embedding-based approaches to normalize textual entities with an ontology

Arnaud Ferré[1,2], Louise Deléger[1], 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, louise.deleger@inra.fr, pierre.zweigenbaum@limsi.fr, claire.nedellec@inra.fr

Abstract

In this paper, we propose a two-step method to normalize multi-word terms with concepts from a domain-specific ontology. Normalization is a critical step of information extraction. The method uses vector representations of terms computed with word embedding information and hierarchical information among ontology concepts. A training dataset and a first result dataset with high precision and low recall are generated by using the ToMap unsupervised normalization method. It is based on the similarities between the form of the term to normalize and the form of concept labels. Then, a projection of the space of terms towards the space of concepts is learned by globally minimizing the distances between vectors of terms and vectors of concepts. It applies multivariate linear regression using the previously generated training dataset. Finally, a distance calculation is carried out between the projections of term vectors and the concept vectors, providing a prediction of normalization by a concept for each term. This method was evaluated through the categorization task of bacterial habitats of BioNLP Shared Task 2016. Our results largely outperform all existing systems on this task, opening up very encouraging prospects.

Keywords: ontology-based entity normalization, vector-based model, distributional semantics, multivariate linear model

1. Introduction

An important part of knowledge is expressed in textual form, such as in scientific articles. Specialized literature is characterized by the presence of terms of interest which are often complex nominal groups (e.g. "epithelial cells of the intestine") and display a high variability in their forms. At the same time, life sciences are a field with many structured, albeit incomplete, representations of knowledge: ontologies. These representations have the advantage of being machine interpretable, which can greatly improve the ability of programs to extract and reuse information from texts. The biomedical/biological field is therefore a good candidate for the development of more efficient generic methods that make use of these structured representations.

To extract entity information from texts, two steps are commonly applied: recognition of named entities and normalization of these entities (also called entity linking in the general domain). The recognition step detects terms of interest (e.g. bacterial habitat references) while normalization identifies them precisely by linking them to specific concepts or categories of an ontology (e.g. "T-cells" is a bacterial habitat mention that can be identified by the "lymphocyte" labeled concept). We are focusing on this particular task.

Today's best-performing normalization methods commonly rely on supervised learning from data that is manually annotated by experts of the field. However in specialized fields, these annotations are rare because they are difficult to obtain. Moreover, given the large number of target concepts in these fields (e.g. the biomedical metathesaurus UMLS contains more than 3 million concepts, the Ontobiotope ontology contains more than 2 thousand concepts, etc.), it seems unlikely to obtain sufficient data to cover all the possibilities of learning. An efficient and distant supervised method (i.e. with learning based on the results of an unsupervised method) would therefore be of great interest.

In this work, we propose several methods, notably a distant supervised normalization method, which we evaluated through the BioNLP 2016 Shared-Task (Deléger et al., 2016) and its task of bacterial habitat categorization. Habitat entities are often designated by complex nominal groups with variable forms, offering a relevant case study.

1.1 Related work

In the biomedical domain, normalization approaches relying on dictionaries and similarities of form between terms and labels from a knowledge source (Hanisch et al., 2005; Schuemie et al., 2007) are historically the oldest approaches and are still much used. They often provide good precision but a low recall because they have difficulty dealing with important variations in the form of terms/labels (e.g. synonymy such as "T-cells" / "lymphocyte", hyperonymy such as "Chondrus crispus" / "algae", etc.). These types of approach also often combine dictionaries with manually defined heuristic rules (Gerner et al., 2010; Kang et al., 2013). However rules are time-consuming to implement and highly dependent on the task and domain.

Statistical approaches aimed at learning directly the associations between terms and labels from corpora. Deep neural networks and word embeddings are those that have achieved the best performance at present (Mehryary et al., 2017), but the lack of training data and the large number of concepts seem to be a limitation on their potential for improvement.

Another recent method, CONTES (Ferré et al., 2017) is based on an approach that does not take into account the form of terms and concept labels, but only distributional information for terms, and hierarchical ontological information for concepts. It is based on the ability of the latest word embedding methods to generate relevant semantic spaces, as well as on building "concept embeddings" that preserves hierarchical information between concepts. This kind of method aims to overcome the problem of variability of forms of associated terms/labels, but does not otherwise take into account relevant morphological information.

In this work, we propose to improve the CONTES method by combining it with a rule-based approach, the ToMap method (Golik et al., 2011).

2. Material

Part of the data used is from the Bacteria Biotope categorization task of BioNLP Shared Task in 2016. The documents are MEDLINE references, consisting of titles and abstracts of scientific articles in the field of biology. The task is to assign concepts from the OntoBiotope ontology[1] to textual entities denoting bacterial habitats (entities are provided and do not have to be detected beforehand). The corpus is divided into two parts: a development corpus (combining the initial training and development corpora of the shared task) and a test corpus which we used to evaluate our method for the normalization task. The entities of each of these corpora have been annotated manually (see Table 1).

	BB		
	Dev.	**Test**	**Total**
Documents	107	54	161
Words	25,185	13,797	38,982
Entities	1,201	720	1,921
Distinct entities	743	478	1,125
Semantic categories	1,360	861	2,221
Distinct sem. categories	332	177	329

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

We used an expanded corpus to generate embeddings. It consists of all titles and abstracts of MEDLINE articles matching the MeshTerm "bacteria" from 2016 (see Table 2). The selection of this corpus was motivated by the need to use a corpus that is representative of the specific field of interest for this normalization task (i.e., habitats of bacteria). Considering that a high quantity of data for computing word embeddings does not guarantee a high quality in the biomedical field (Chiu et al., 2016), we chose to use this smaller targeted corpus as opposed to a larger and less relevant corpus.

sentences	7,714,841
raw words	154,749,541
raw words without stopwords	74,808,541
unique word (stopwords included)	1,565,740

Table 2: Descriptive statistics of extended corpus

3. Method

3.1 Rule-based approach: ToMap

As first approach to normalize entities, we used the ToMap method (Golik et al., 2011). ToMap relies on the internal morpho-syntactic structure of entities and maps them based on their syntactic heads, the underlying assumption being that the head is often the most informative component. The system first looks for a match between the syntactic head of an entity and the syntactic heads of the ontology concepts. Then, the entity is assigned the concept(s) with matching head(s). As there may be multiple candidates for a given entity, a Jaccard index is also computed between the entity and each of the concepts, and the concept with the highest score is selected.

The core algorithm is complemented by a set of heuristics designed to handle specific cases. For instance, a list of uninformative syntactic heads is provided so that if a head belongs to this list, then the algorithm tries to match the modifiers instead (e.g., in "water sample", "sample" is not very informative so "water" will be chosen to perform the mapping). Other heuristics include disambiguation rules targeted at particularly ambiguous terms (e.g., "plant" which can designate either a processing factory or living things such as trees, flowers, etc.). These heuristics are dependent on the type of entities (in our experiments, heuristics are designed for habitat entities). In the remainder of the article, we refer to the core algorithm alone (without the specific heuristics) as "simple ToMap".

Not all entities can be matched to an ontology concept with the ToMap method. When the syntactic head of an entity has no equivalent in the ontology, the entity cannot be normalized to a precise concept and is simply assigned the root concept of the ontology (i.e., "Bacteria habitat").

3.2 Embedding-based approach: CONTES (CONcept-TErm System)

3.2.1 Word embeddings with Word2Vec and Skip-Gram architecture

To train the word embedding method on the expanded corpus, the sentences were randomly shuffled and converted to lower case. Next, we applied the Word2Vec method with the Skip-Gram architecture (Mikolov et al., 2013). To evaluate our different methods, all based on embeddings, we chose to adopt the optimal parameters given in similar work on biomedical literature, which does not seem to drastically impact results (Chiu et al., 2016).

alpha	0.05
min-count	0
negative	5
sample	0.001
vector size	200
window size	2

Table 3: Main Word2Vec/Skip-Gram parameters used to calculate word embeddings for this work

We chose to keep all words, even those appearing once in the whole corpus, because we want to normalize entities that contain words with a low frequency and it does not seem to have a real impact on the global vector space. The most impacting parameters are the size of the vectors and the size of the contextual windows. We chose the vector size following the initial work on CONTES (Ferré et al.,

[1]http://2016.bionlp-st.org/tasks/bb2/OntoBiotope_BioNLP-ST-2016.obo?attredirects=0

2017), which obtained the best results with an output vector size of 200. However, we choose a smaller size of the context window, because of their possible potential to affect the nature of semantic proximity in the embedding space (Turney, 2012). Indeed, a common hypothesis is that larger window emphasizes the learning of domain similarity between words, while a narrow context window leads to hyponymic gathering. Thus, a smaller size of context window could be more preferable in our task.

All remaining parameters are those by default. Table 3 gives the main parameters.

3.2.2 Term and concept embeddings

The CONTES method computes embeddings for each word of the expanded corpus as described in the previous subsection. Then, for each entity of the training corpus and of the test corpus, a term vector is calculated by computing the barycenter of the vectors of the words that compose the entity. In parallel, concept vectors are calculated for all concepts of the ontology of interest, thus generating an ontological space. Each concept is associated with a vector of equivalent size to the number of concepts in the ontology. Each dimension is associated with a fixed concept. The vector is initialized as a one-hot (i.e. all weights are set to zero except the weight associated with the current concept, which is set to one), then each of the weights of the dimensions associated with the ancestors of the current concept are also set to one. The method encodes the hierarchical information of the ontology: if we estimate the cosine similarity between ontology concepts, the parent/child concepts are always the closest. This hierarchical information has the advantage of being the most frequent semantic relation (*is_a* relation) in ontologies.

After the generation of the term vectors (that represent the textual entities) and of concept vectors (that represent the concepts from the ontology), the method performs a projection of the term vectors into the ontological space. We use the training corpus to learn the optimal projection. That is, the projection that globally minimizes the distance between the projection of terms in the ontological space and the vectors of the associated concepts is determined. The learning method we used is a multivariate linear regression in order to limit the overfitting risk, particularly with regard to a relatively small training corpus. This projection is then applied to the term vectors of the test corpus, allowing to obtain new term vectors in the ontological space, and to calculate a cosine similarity with the concept vectors. The closest concept is chosen to normalize a term. The method has been designed to address the problem of term variability, because it does not rely on the similarity of form between terms and concept labels.

3.3 Combining the two approaches: HONOR (Hierarchical Ontological NORmalization)

The aim of the HONOR method is to take advantage of the precision of ToMap and to complement it with the CONTES method which has the potential to address the problem of form variability. More specifically, cases that are not handled by ToMap will be normalized by CONTES. Our hypothesis is that CONTES, which is not based on form similarity, should have the potential to propose relevant normalization predictions in cases where ToMap cannot. The overall scheme of the method is shown in Figure 1.

3.4 Distant supervision version

Additionally, as ToMap is not supervised and potentially yields good precision results, we can also use it to generate a first prediction on a larger corpus, which will be used to train CONTES rather than using the gold standard. In this setting, the method becomes a distant supervised method, as it would not rely on manual annotation anymore to train its learning model. We chose to use the same corpus as the one used for the training of Word2Vec. We tested for many random subsamples of the full predictions of ToMap on this corpus and analyzed the impact of the number of selected examples on the global performance of these methods.

We tested four different versions of the distant supervised method: two versions are based on the unsupervised predictions of ToMap without using its supplementary heuristics (simple ToMap), and the two others are based on the predictions of the complete version of ToMap. For each configuration (simple ToMap vs. complete ToMap), there is one version which only uses the CONTES method, and another version which uses the combined approach of HONOR. For both versions of the HONOR method, we studied the impact of the number of examples (i.e. the predictions of ToMap on the expanded corpus). The smallest batches have the same order of magnitude as the gold standard and the biggest batch is a hundred times bigger.

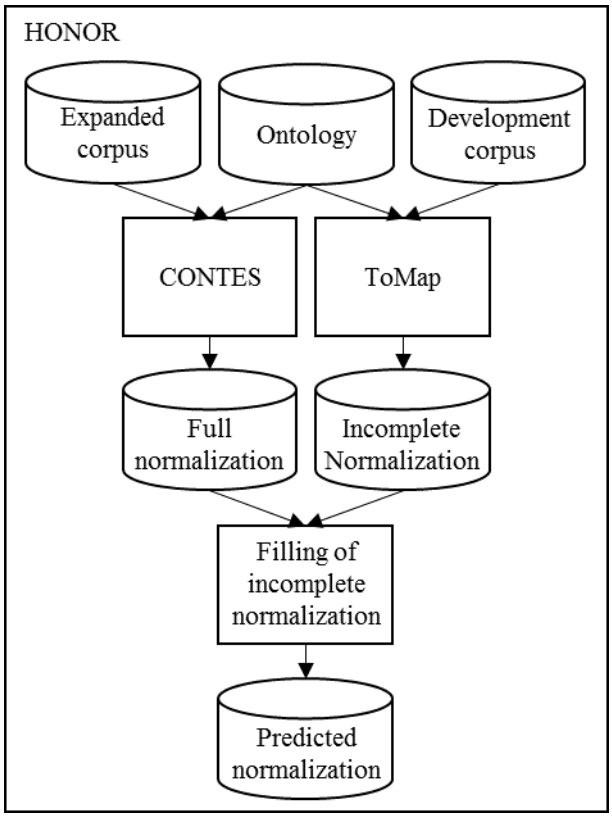

Figure 1: Global schema of the method HONOR

An implementation of the ToMap and CONTES methods is available via the AlvisNLP/ML[2] engine (Ba et al., 2016).

[2] https://bibliome.github.io/alvisnlp/

You need to download the module for the CONTES method[3]. AlvisNLP/ML allows to combine the two methods and to emulate HONOR.

4. Results

4.1 Evaluation

We evaluated the performance of ToMap and CONTES individually and of our combined approach HONOR (both in its supervised and distant supervised settings), on the Bacteria Biotope normalization task of the BioNLP Shared Task 2016. The predicted concept 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 of the BioNLP-ST 2016 challenge site. We computed baseline results by assigning all terms to the concept "bacteria habitat", which is the root of the OntoBiotope ontology hierarchy. The baseline obtained a score of 32% (Table 6). For comparison, we also included results from other systems that proposed methods to normalize bacterial habitats and evaluated them on the shared task corpus (Grouin, 2016; Tiftikci et al., 2016; Mehryary et al., 2017).

4.2 ToMap and CONTES individually

ToMap can provide a normalization prediction for 54% of the terms. Even if there are still problems of ambiguity, it enables us to estimate the proportion of terms that have a form similar to an ontology label at least at the level of their syntactic head. Despite its limitation, ToMap obtained good performances on the test corpus and ranked above existing methods (Table 6). This shows that the method has a really good precision when it can provide a prediction.

As long as terms are composed with tokens from the initial vocabulary, which allows to calculate an embedding for the terms, CONTES can provide a normalization prediction for all terms to normalize. The method obtained a score of 61% (Table 6). Compared to previous work, this version of CONTES includes a few improvements (refined embedding parameters and extended training corpus). We also tested the impact of the hierarchical information encoded in concept vectors to evaluate the gain when using this information compared to a simple one-hot representation (Table 4). Indeed, a one-hot encoding does not use hierarchical information because all vectors are equidistant from each other. The improvement of 7% validates the usefulness of this information to improve the matching between the two vector spaces.

Taking into account hierarchical information	0.61
One-hot encoding	0.54

Table 4: Comparison between the CONTES method encoding of concepts and an approach with one-hot encoding of concepts

4.3 HONOR

We compared the results of HONOR to those of existing systems. We report all results in Table 5 and Table 6. In 2016, two teams participated in this task. The best method was the BOUN method (Tiftikci et al., 2016) which combined a form similarity approach and an information-retrieval based approach (based on tf-idf) and obtained a 0.61 score. The method of LIMSI (Grouin, 2016), which is based on a form similarity approach, had the lowest performance (0.43). Since the 2016 shared task, the University of Turku has proposed an end-to-end neural network method (Mehryary et al., 2017) which has outperformed these methods with a 0.63 score.

Compared to these systems, our method performs well above by a 10 points increase compared to the Turku system. It also brings a significant gain to the ToMap method (+7 points in the supervised version and +6 points in the distant supervised version). There is only a one-point difference between the supervised and distant supervised HONOR methods, which shows that our method could enable to perform without manual annotated data and without a significant loss of performance.

Method	Score
Unsupervised	
ToMap	0.66
Simple ToMap	0.62
Distant Supervised	
HONOR (ToMap)	0.72
HONOR (simple ToMap)	0.66
CONTES (ToMap)	0.61
CONTES (simple ToMap)	0.59
Supervised	
HONOR (ToMap)	**0.73**
CONTES (improved)	0.61

Table 5: Results of all the methods described in this article on the normalization task of BioNLP-ST 2016

System	Score
Supervised HONOR	**0.73**
Distant supervised HONOR	**0.72**
Turku	0.63
BOUN	0.62
CONTES (2017)	0.60
LIMSI	0.43
Baseline	0.32

Table 6: Results on the normalization task of BioNLP-ST 2016

4.4 Impact of the number of examples in distant supervision

For the two distant supervised versions of HONOR, we evaluated performances obtained with three different sizes of data batches. These batches have been constituted by randomly choosing a variable number of examples in the predictions of ToMap on the expanded corpus. For each batch size, variations on the score have been estimated over many executions. The results remain relatively stable across different batches of the same size: less than 0.2% variation for the biggest batch and less than 1.5% for the smallest. It seems that there is a small gain for both versions

[3] https://github.com/ArnaudFerre/CONTES

to use 10^5 rather 10^4 examples. The means of these results are reported in Table 7:

Number of predictions from ToMap:	**10^3**	**10^4**	**10^5**
HONOR with simple ToMap:	0.64	0.66	0.66
HONOR with ToMap:	0.70	0.72	0.72

Table 7: Results for the two distant supervised version of HONOR.

5. Discussion and future work

The key hypothesis behind the efficiency of the CONTES and HONOR methods is that the semantic space of terms (based on distributional semantics) and the semantic space of the ontology (based on the specific hierarchical information of the current ontology) are homologous. Even if this work seems to indicate that there are at least some similarities between these two vector spaces, this hypothesis is most likely too strong. It would certainly be interesting to alter the space of embeddings to increase this similarity and then use a non-linear algorithm to find a better projection from the embedding space to the ontological space.

With this kind of word embedding-based approach, there is a problem with computing words that have not been met before (i.e. out-of-vocabulary words). Consequently, any new word encountered should require a complete recalculation of word embeddings to be taken into account. Recently, a new method to calculate embeddings seems to overcome this difficulty (Bojanowski et al., 2016) and we plan to estimate its performance on our method in further work.

Like the CONTES method, a problem of high dimensional representation of the concept vectors persists. We plan to address this issue in further work.

The bacterial habitat normalization task includes the normalization of an entity by multiple concepts. Currently, no system has successfully addressed this issue. We would like to investigate this problem in the future.

Finally, beyond the benefits of using a multivariate linear regression, work is underway to explore the possibility that a non-linear learning method can provide a better projection between vector space of terms and vector space of concepts.

6. Conclusion

Our method seems to open up interesting perspectives for joint use of word embeddings, ontologies and form similarity based methods, particularly for the domain specific literature. This kind of synergy could also address challenges in fields where sufficient manual annotations are difficult to obtain.

7. Acknowledgment

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

8. Bibliographical References

BA MOUHAMADOU, BOSSY ROBERT (2016), Interoperability of Corpus Processing Workflow Engines: The Case of. AlvisNLP/ML in OpenMinTeD, np, in: *Meeting of working Group Medicago sativa*.

BOJANOWSKI PIOTR, GRAVE EDOUARD, JOULIN ARMAND, MIKOLOV TOMAS (2016), Enriching Word Vectors with Subword Information. *ACL 2017*.

CHIU BILLY, CRICHTON GAMAL, KORHONEN ANNA, PYYSALO SAMPO (2016), How to Train Good Word Embeddings for Biomedical NLP. *Proceedings of BioNLP16*: 166.

DELÉGER LOUISE, CHAIX ESTELLE, BA MOUHAMADOU, FERRÉ ARNAUD, BESSIÈRES PHILIPPE, NÉDELLEC CLAIRE (2016), Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016.

FERRÉ ARNAUD, ZWEIGENBAUM PIERRE, NÉDELLEC CLAIRE (2017), Representation of Complex Terms in a Vector Space Structured by an Ontology for a Normalization Task. *BioNLP 2017*: 99–106.

GERNER MARTIN, NENADIC GORAN, BERGMAN CASEY M. (2010), LINNAEUS: A Species Name Identification System for Biomedical Literature. *BMC bioinformatics*, 11(1): 85.

GOLIK WIKTORIA, WARNIER PIERRE, NÉDELLEC CLAIRE (2011), Corpus-Based Extension of Termino-Ontology by Linguistic Analysis: A Use Case in Biomedical Event Extraction, 37–39, in: *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence*.

GROUIN CYRIL (2016), Identification of Mentions and Relations between Bacteria and Biotope from PubMed Abstracts. *ACL 2016*: 64.

HANISCH DANIEL, FUNDEL KATRIN, MEVISSEN HEINZ-THEODOR, ZIMMER RALF, FLUCK JULIANE (2005), ProMiner: Rule-Based Protein and Gene Entity Recognition. *BMC Bioinformatics*, 6(Suppl 1): S14.

KANG NING, SINGH BHARAT, AFZAL ZUBAIR, VAN MULLIGEN ERIK M, KORS JAN A (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.

MEHRYARY FARROKH, HAKALA KAI, KAEWPHAN SUWISA, BJÖRNE JARI, SALAKOSKI TAPIO, GINTER FILIP (2017), End-to-End System for Bacteria Habitat Extraction. *BioNLP 2017*: 80.

MIKOLOV TOMAS, CHEN KAI, CORRADO GREG, DEAN JEFFREY (2013), Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

SCHUEMIE MARTIJN J., JELIER ROB, KORS JAN A. (2007), Peregrine: Lightweight Gene Name Normalization by Dictionary Lookup, 131–133, in: *Proc of the Second BioCreative Challenge Evaluation Workshop*.

TIFTIKCI MERT, SAHIN HAKAN, BÜYÜKÖZ BERFU, YAYIKÇI ALPER, OZGÜR ARZUCAN (2016), Ontology-Based Categorization of Bacteria and Habitat Entities Using Information Retrieval Techniques. *ACL 2016*: 56.

TURNEY PETER D. (2012), Domain and Function: A Dual-Space Model of Semantic Relations and Compositions. *Journal of Artificial Intelligence Research*.

WANG JAMES Z., DU ZHIDIAN, PAYATTAKOOL RAPEEPORN, YU PHILIP S., CHEN CHIN-FU (2007), A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics*, 23(10): 1274–1281.