Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1959–1966
Marseille, 11–16 May 2020

Handling Entity Normalization with no Annotated Corpus: Weakly Supervised Methods Based on Distributional Representation and Ontological Information

Arnaud Ferré[1,2], Robert Bossy[1], Mouhamadou Ba[1], Louise Deléger[1], Thomas Lavergne[2], Pierre Zweigenbaum[2], Claire Nédellec[1]
[1]Université Paris-Saclay, INRAE, MaIAGE, allée de Vilvert, 78350 Jouy-en-Josas, France
[2]Université Paris-Saclay, CNRS, LIMSI, rue John Von Neumann, 91400 Orsay, France
{arnaud.ferre, robert.bossy, mouhamadou.ba, louise.deleger, claire.nedellec}@inrae.fr
{thomas.lavergne, pierre.zweigenbaum}@limsi.fr

Abstract

Entity normalization (or entity linking) is an important subtask of information extraction that links entity mentions in text to categories or concepts in a reference vocabulary. Machine learning based normalization methods have good adaptability as long as they have enough training data per reference with a sufficient quality. Distributional representations are commonly used because of their capacity to handle different expressions with similar meanings. However, in specific technical and scientific domains, the small amount of training data and the relatively small size of specialized corpora remain major challenges. Recently, the machine learning-based CONTES method has addressed these challenges for reference vocabularies that are ontologies, as is often the case in life sciences and biomedical domains. Its performance is dependent on manually annotated corpus. Furthermore, like other machine learning based methods, parametrization remains tricky. We propose a new approach to address the scarcity of training data that extends the CONTES method by corpus selection, pre-processing and weak supervision strategies, which can yield high-performance results without any manually annotated examples. We also study which hyperparameters are most influential, with sometimes different patterns compared to previous work. The results show that our approach significantly improves accuracy and outperforms previous state-of-the-art algorithms.

Keywords: Information Extraction, entity normalization, entity linking, supervised learning

1. Introduction

Entity Normalization, also known as Entity Linking or Entity Grounding, is an Information Extraction subtask that consists in the assignment of categories or concepts from a reference vocabulary to entity mentions in the text (i.e. words or sequences of words). Ontologies and taxonomies have been frequently used as reference vocabularies since the late 1990s (Faure and Nédellec, 1998; Hwang, 1999). For instance, entity normalization with concepts of an ontology could consist in linking the textual entity mention "*T-cell"* to a concept referenced by a unique identifier and a label such as: <OBT:001342: lymphocyte>.

Entity Normalization contributes to the reuse of the extracted information and to its integration with data in reference databases (Nédellec et al., 2009). Entity Normalization has recently gained traction in technical and scientific domains, especially in Life Sciences and Health Sciences. Several benchmark datasets have been proposed in these domains (Wei et al., 2015; Roberts et al., 2017; Deléger et al., 2016). Entity Normalization methods handle the problem as a classification problem. They are based either on pattern-matching rules (Aronson, 2001) or on Machine Learning (ML) algorithms (Leaman et al., 2013). Both are able to operate on the surface form of entities (i.e. their sequence of characters), on NLP analyses (lemmatization, POS-tagging, syntactic parsing) (Aronson, 2001), or on distributional semantic representations such as word embeddings (Limsopatham and Collier, 2016).

The main limitation of rule-based and entity-form-based methods is that they require entity mentions to present some similarity with concept labels to be efficient. For instance, rule-based methods cannot assign the mention "*T-cell*" to a concept labelled "*lymphocyte*", unless a similar form of the term "*T-cell*" is added to the concept (Pratt and Yetisgen-Yildiz, 2003). For a real task with thousands or millions of concepts, this represents a tremendous work, and can quickly introduce some ambiguities between concepts ("*plant*", vegetal or factory entity?), which require additional and special processing (Hanisch et al., 2005; Morgan et al., 2008; Aronson and Lang, 2010).

ML and embedding-based methods aim to address this limitation. They have also shown a better adaptability to various specific normalization tasks. These methods yield good results if provided with sufficient training data (Sil et al., 2018). However training data in the form of manually annotated corpora are costly to produce and are generally limited in size (Uschold and King, 1995), in particular in specific domains where the level of expertise required to annotate training data is high and the number of target concepts is large (Lipscomb, 2000; McCray, 1989; Nédellec et al., 2018). Learning without examples is a problem called zero-data learning (or few-shot learning) (Larochelle et al., 2008; Ravi and Larochelle, 2016), and is a well-known challenge for ML.

CONTES (Ferré et al., 2017) and HONOR (Ferré et al., 2018) are two recent methods that address training data paucity by exploiting ontological subsumption information (*is_a* relation between concepts or categories). CONTES uses the subsumption graph of the ontology together with word embeddings. HONOR combines CONTES and the rule-based method ToMap (Golik et al., 2011). HONOR achieves state of the art performance on the Bacteria Biotope normalization task of BioNLP Shared Task 2016 (BB3) (Deléger et al., 2016). Nevertheless, both of these methods still need an annotated corpus which provides ground truth training examples. In this paper we present novel approaches to entity normalization, using CONTES and HONOR with no manually annotated corpus, based on a weak supervision strategy.

Moreover, embedding based methods vary depending on the representation of examples and algorithm hyperparameters (Chiu et al., 2016). We propose an experimental setting with the aim of highlighting the influence of a wide range of different factors (corpus selection, pre-processing, word embedding hyper-parameters) and finding the optimal configuration for adapting a generic normalization method to a specific task. Pre-processing still receives too little focus when evaluating embedding-based systems (Camacho-Collados and Pilehvar, 2018), in favour of hyperparameter study or the use of general precomputed embeddings.

We evaluated our approach on the Bacteria Biotope normalization task of BioNLP Shared Task 2016 (BB3). This task illustrates the challenge that we aim to address: it is a domain-specific normalization task, well-recognized in the BioNLP community, and with a small amount of training data available compared to the number of concepts of the task.

2. Related Work

2.1 A Brief History of Entity Normalization Methods in Scientific Domains

Most normalization methods in technical and scientific domains rely on the similarity between entity forms and concept labels. Due to frequent linguistic variations (*e.g.* noun-phrase inversion, typographic variations, synonymy), these methods are dependent on comprehensive lexicons. Several strategies are used to ensure comprehensiveness: third-party resources (Gerner et al., 2010; Lee et al., 2015) inflection generation (Hanisch et al., 2005; Tsuruoka et al., 2007; Ghiasvand and Kate, 2014), pre-processing (lemmatization, stemming or stopword filtering) (Schuemie et al., 2007), giving more weight to syntactic heads of mentions and labels (Aronson, 2001; Golik et al., 2011).

To handle domain-specific ambiguities, which notably increase with strategies such as third-party resources and inflection generation, and simultaneously to specialize for the task domain, some methods use a hand-crafted blacklist: for instance, the ToMap method (Golik et al., 2011) has a version adapted to the BB3 task, and the Peregrine method (Schuemie et al., 2007) has a version adapted to BioCreative II. Another method by Claveau (2013) weights tokens with an Information Retrieval measure, which aims to automatically mitigate the weight of ambiguous words.

Other methods use vector representations to compute a similarity measure between text mentions and concept labels. Notably, Tiftikci et al. (2016) and Mehryary et al. (2017) estimate the semantic similarity between two expressions by computing a cosine similarity between TF-IDF bag-of-words representations (Manning et al., 2009). These representations are based on word forms, and fail to link mentions that do not share any common token with the correct concept label. To address this limitation, the ML-based DNorm method (Leaman et al., 2013) learns a function that estimates high similarities between distant TF-IDF bag-of-words representations of mentions and associated concept labels.

Some advances in NLP have been made through word embeddings as built by Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2016) or more recently ELMo (Peters et al., 2018) or BERT (Devlin et al., 2018), as a way to compute and represent the meaning of words from the contexts in which they are observed. Word embeddings are vectors with the advantage of a smaller number of dimensions. However, their acquisition requires large amounts of untagged corpora.

To favour their mapping, text mentions and labels can be represented by embeddings in the same space. So, a first approach to normalize a mention embedding is to find the nearest concept label embedding. For instance, the BOUNEL method (Karadeniz and Özgür, 2019) computes a cosine similarity between these embeddings to find the best concept candidates for a mention. Limsopatham and Collier (Limsopatham and Collier, 2016) also use word embeddings for the representation of mentions, and a convolutional neural network architecture to learn to classify each mention representation by the correct concept.

The results of word embedding-based methods significantly depend on the choice of a large unannotated corpus, on the chosen hyper-parameters (Chiu et al., 2016), and on parameter initialization. Moreover, through specialized corpora, domain specialized embeddings can increase the performance of methods (Roberts, 2016). This specialization can be emphasized by exploiting external knowledge (Faruqui et al., 2014; De Vine et al., 2014; Celikyilmaz, 2015), such as that contained in ontologies (Ferré et al., 2017; Yen et al., 2018).

2.2 The CONTES and HONOR Methods

The word embeddings- and ML-based CONTES method (Ferré et al., 2017) differs from these methods by its use of concept vectors instead of concept label embeddings or concept one-hot vectors. Each concept has a unique vector in a vector space that is different from the space of mention embeddings. CONTES then performs a multivariate linear regression to find a linear projection from the vector space of mentions to the vector space of concepts. The learning optimization goal is to minimize globally the Euclidean distance between each projected mention vector and its concept vector(s) (see Figure 1).

CONTES then uses the learned parameters to project any mention vector onto the ontological space. It computes a cosine similarity between the projected vector and every concept vector. CONTES finally selects the concept with the most similar vector as the prediction for normalization (see Figure 2).

The vector space of concepts includes vectors computed with ontological information rather than one-hot vectors (where all weights are set to zero except the weight associated with the current concept, which is set to one), as in Limsopatham and Collier (Limsopatham and Collier, 2016). Each concept is associated with a vector whose size is the number of concepts in the ontology. Each dimension is associated with a fixed concept.

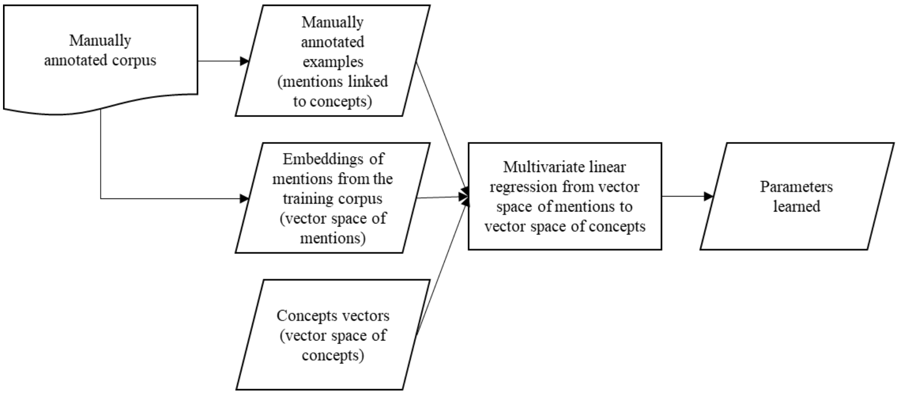
Manually annotated corpus
Manually annotated examples (mentions linked to concepts)
Embeddings of mentions from the training corpus (vector space of mentions)
Concepts vectors (vector space of concepts)
Multivariate linear regression from vector space of mentions to vector space of concepts
Parameters learned

Figure 1: Training step of the CONTES method.

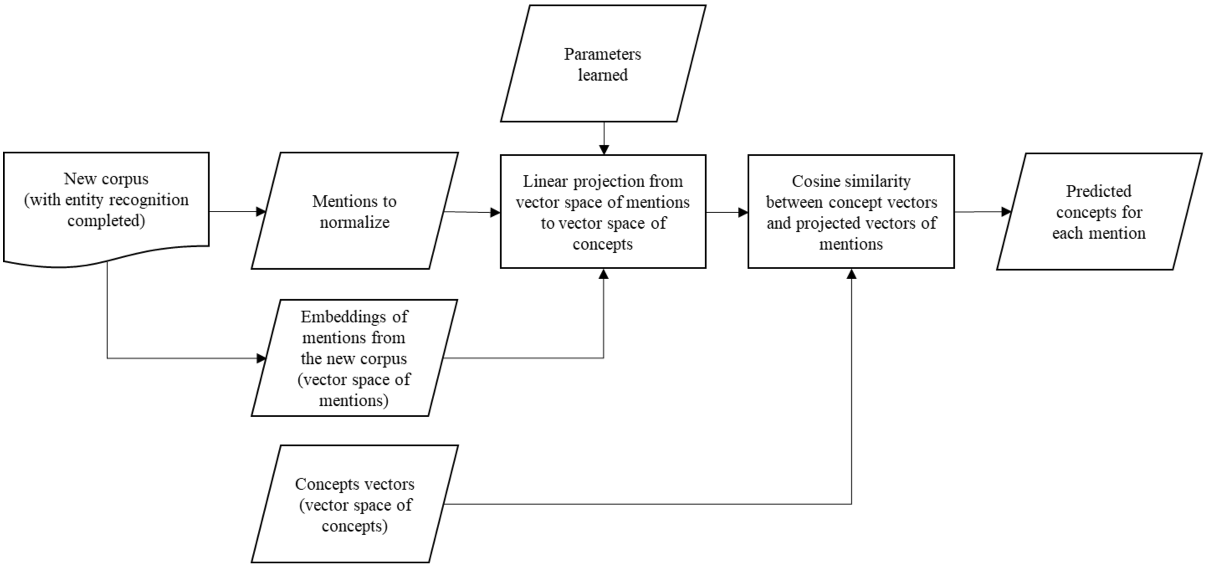
Parameters learned
New corpus (with entity recognition completed)
Mentions to normalize
Linear projection from vector space of mentions to vector space of concepts
Cosine similarity between concept vectors and projected vectors of mentions
Predicted concepts for each mention
Embeddings of mentions from the new corpus (vector space of mentions)
Concepts vectors (vector space of concepts)

Figure 2: Prediction step of the CONTES method.

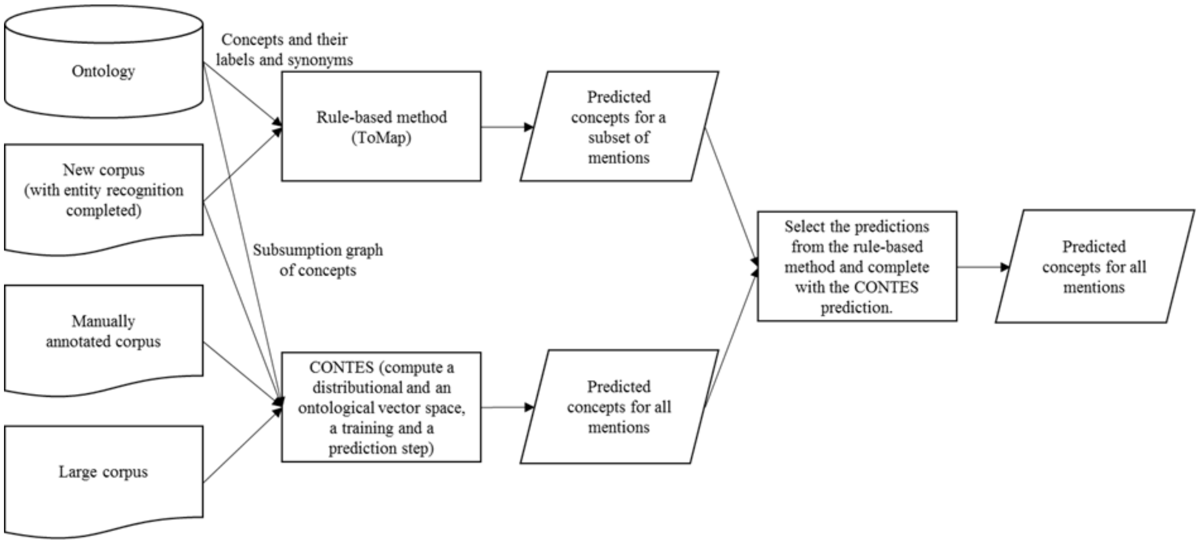
Ontology
Concepts and their labels and synonyms
Rule-based method (ToMap)
Predicted concepts for a subset of mentions
New corpus (with entity recognition completed)
Subsumption graph of concepts
Select the predictions from the rule-based method and complete with the CONTES prediction.
Predicted concepts for all mentions
Manually annotated corpus
CONTES (compute a distributional and an ontological vector space, a training and a prediction step)
Predicted concepts for all mentions
Large corpus

Figure 3: Diagram of the HONOR method, combining the CONTES and the rule-based ToMap methods.

The vector is initialized as a one-hot, then each of the weights of the dimensions associated with all ancestors of the current concept are also set to one. This way, the representation encodes the hierarchical information of the ontology: when computing cosine similarity between ontology concepts, parent/child concepts are always nearest to each other. This hierarchical relation is the most frequent semantic relation (is_a relation) in existing ontologies.

To compute mention embeddings, CONTES uses word embeddings from a large corpus. For each recognized mention of a corpus, a mention embedding is computed by calculating the barycenter of the vectors of the words that compose the entity. Roberts (2016) has shown that using word embeddings trained on a specific biomedical corpus for a biomedical task obtains better results than with a larger out-of-domain corpus. Thus, in order to improve the similarity between the two spaces, CONTES used specific word embeddings trained on a biomedical corpus rather than on large general corpora.

CONTES authors hypothesize that it is possible to learn a structural similarity between a distributional space and an ontological space from the same domain.

Rule-based methods predict correct concepts with a better precision whereas ML and embedding-based methods usually achieve better recall. To benefit from the best of both approaches, the HONOR method combines the ML and embedding-based CONTES with the rule-based method ToMap (Golik et al., 2011) in two steps. First, ToMap predicts concepts for all mentions it can, then CONTES predicts a concept for the mentions left out by ToMap (see Figure 3).

3. Task Data Sets

In this section, we present the BB3 normalization task dataset that we used to evaluate the methods. The task consists in linking mentions of bacteria species names to a taxonomy, and bacterial habitats to one or more of the 2,320 concepts of the dedicated ontology OntoBiotope[1]. Each concept can also contain some synonyms of the label (0.2 synonyms on average), which gives a total of 2,739 terms associated to concepts of the ontology.

The normalization of taxonomic mentions of bacteria is not much of a challenge for the BioNLP community because the nomenclature is complete, variations are relatively standardized, and synonymy is rare (except in some special cases such as strain names). Thus string matching with basic variations yields decent results (Grouin, 2016). Habitat mentions are subject to much more variations, and, due to the microscopic nature of bacteria, any object or place can be construed as a habitat. The normalization of habitat mentions is thus a challenging task that has generated many studies.

The BB3 corpus is made of titles and abstracts from PubMed entries annotated with entities. The annotated corpus is split into training, development, and test sets. The BB3 online evaluation service measures the performance of the predictions on the test set. The total volume of the annotated corpus is rather small: the whole training and development corpus contains around 25 thousand tokens, around 1,200 mentions of bacterial habitats and only 12% of the concepts of OntoBiotope occur in this corpus (266 distinct concepts).

The evaluation metrics is based on a semantic similarity between the reference concept and the predicted concept. The similarity equals 1 if they are equal, and tends to zero if the concepts are farther in the hierarchy. The overall score is the mean of the similarity for each mention in the test set (Wang et al., 2007).

4. Methods

4.1 Weakly Supervised Strategy

CONTES achieves good results on BB3 with a small set of training data but it still needs manually annotated examples (Ferré et al., 2017). To address this limitation, we developed a weak supervision strategy that exploits the lexicalization of the ontology: ontology concepts are usually labeled by terms and synonyms. We use these labels and synonyms as training examples, instead of, or combined with, annotated mentions in the training corpus. This gives about twice as many training examples as in the training corpus, and at least one training example per concept.

4.2 Preprocessing

We also extend CONTES by studying the impact of linguistic preprocessing as complementary to the study of hyperparameters for embeddings calculation that embedding-based methods usually focuse on. Preprocessing steps such as lemmatization or stemming, stop-words filtering and masking (i.e. replacing some expressions with a single mnemonic form) have the positive effect of decreasing the size of the vocabulary and increasing the distributional signal for each token. At the same time, lemmatization and stemming can have undesirable effects such as merging tokens that have no common meaning, and stopword filtering and masking can delete useful information. Thus, we study the impact of these preprocessing strategies on the performance of the CONTES and HONOR methods.

For stemming, we tested an implementation of the Snowball algorithm (Porter, 1980), and for lemmatization we used GeniaTagger[2], a state-of-the-art lemmatizer and POS-tagger for the biomedical domain. For stopword filtering, we removed grammatical words (determiners, prepositions, conjunctions, “to”), and punctuations. For masking, we replaced each numerals and species names with a unique token.

4.3 Word2Vec Parametrization and Corpus Selection

We have extended the study of Word2Vec hyperparameters beyond those in CONTES and HONOR (Ferré et al., 2017; Ferré et al., 2018). We used the Skip-Gram architecture of Word2Vec with negative sampling (Mikolov et al., 2013) in a similar way as CONTES. We

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

[2] http://www.nactem.ac.uk/GENIA/tagger/

have screened the hyperparameters, notably the size of the contextual window and the size of the word embedding vectors. We set the minimum occurrences of words to zero in order to consider rare domain words.

We trained the embeddings on a corpus dedicated to the domain of BB3, i.e. microbiology. From the whole PubMed collection, we selected entries indexed by keywords of the MeSH controlled vocabulary[3] indicating that the entry's topic is microbiology ("*Bacteria*", "*Microbiology*", etc.). The resulting corpus contains 2,333,943 entries (412,240,083 tokens). We refer to it as the Microbiology Corpus in the following. This corpus is smaller than the usual corpora used to calculate embeddings such as Wikipedia or Google News, which are still commonly used for biomedical tasks. However, it constitutes a relatively large biomedical corpus (Roberts, 2016). It allows to generate embeddings in less than an hour on a standard server computer. It is manageable enough to enable the screening of the hyperparameters and of the pre-processing settings in a reasonable time. We also compared the results obtained with this corpus to publicly available domain-specific and general domain word embeddings[4].

5. Results

All the presented scores are averaged over ten runs on the development set, except for the final evaluation (Table 4) which contains unique scores on the test set. In all our experiments, we observed a maximal standard deviation of the score of 0.011, due to the random initialization of Word2Vec.

5.1 Impact of Preprocessing and Word2Vec Parametrization

We observed no impact of the window size (see Table 1). Thus, we set a short symmetrical window of two tokens in all subsequent experiments.

We observe however a significant impact of the vector size on the performance (see Table 2). Previous work with the CONTES method had reported an optimal value of 200. We find the same optimal value for the supervised version but, surprisingly, we find a different optimal value of 1,000 for the weak supervision strategy.

Pre-processing yields rather mixed results. Lemmatization does not significantly improve the results, and stemming even degrades them (see Table 3). We hypothesize that stemming shortens the token forms too aggressively and entails a loss of information in the word embeddings. Masking numerals and named entities also did not significantly affect the results. In contrast, we consistently observe a significant improvement of the score when stop words are removed (see Table 3).

The embeddings calculated with the Microbiology Corpus and the pre-processed corpus give consistently better scores (0.59) than those obtained with off-the-shelf embeddings trained on Wikipedia (0.57) or the whole PubMed (0.56).

Window size	**Stopword filtering**	**No filtering**
1	0.62	0.59
2	0.60	0.61
3	0.61	0.61
5	0.60	0.59
8	0.61	0.59
12	**0.64**	0.60
20	0.62	0.60

Table 1: Screening of the window size, with and without stopword filtering. CONTES results on the BioNLP-ST 2016 Bacteria Biotope normalization task, trained on the training set and evaluated on the development set. Vector size is set to 200.

Vector size	**CONTES (training)**	**CONTES (labels)**	**CONTES (training and labels)**
25	0.52	0.46	0.49
50	0.55	0.49	0.53
100	0.57	0.49	0.58
200	**0.59**	0.55	0.62
250	0.58	0.57	0.65
300	0.53	0.57	0.65
500	0.36	0.61	0.67
1000	0.31	**0.63**	**0.72**
2000	0.37	0.38	0.47
3000	0.39	0.39	0.46

Table 2: Screening of the vector size. CONTES results on the BioNLP-ST 2016 Bacteria Biotope normalization task, trained either on the training set ("training"), the labels and synonyms ("labels"), or both ("training and labels"), and evaluated on the development set. The window size is set to 2. Stopwords are filtered.

[3] https://meshb.nlm.nih.gov/

[4] https://vecto.readthedocs.io/en/docs/tutorial/getting_vectors.html

Token normalisation	No filtering	Stopword filtering
None	0.578±0.013	**0.598**±0.011
Lemmatization	0.582±0.006	0.588±0.013
Stemming	0.567±0.012	0.569±0.016

Table 3: Screening of token preprocessing. The standard deviation is calculated on all masking preprocessing. CONTES results on the BioNLP-ST 2016 Bacteria Biotope normalization task, trained on the training set and evaluated on the development set. The vector size is set to 200, the window size is set to 2.

5.2 Results for the Weak Supervision and Mixed Approaches

To evaluate the weak supervision strategy, we used word embeddings obtained with the following settings:

- the Microbiology Corpus with stopword filtering and no other pre-processing;
- a window size of 2;
- a vector size of 1,000.

The results are shown in Table 4. For comparison purposes we also report the score obtained by a baseline method. It consists in a strict string matching of concept labels and synonyms against lemmatized mentions. We also report published results obtained by four other systems on the same benchmark: BOUN (Tiftikci et al., 2016), Turku (Mehryary et al., 2017), BOUNEL (Karadeniz and Özgür, 2019), and ToMap (Golik et al., 2011). We trained, ran, and evaluated CONTES and HONOR using the generated embeddings. We trained them with three different training sets:

- BB3 training and development sets (*training*);
- concept labels and synonyms (*labels*) for the weak supervision approach;
- the union of the above (*training and labels*).

We observed that both CONTES and HONOR perform similarly when trained either on the training and development sets, or on concept labels and synonyms. These two training sets are yet different:

- the annotated corpus contains in the order of one third of training data compared to the ontology,
- the ontology has at least one training example per concept, while the annotated corpus contains mentions only from 12% of the concepts of the ontology.

It can mean that the concepts mentioned in the annotated corpus (whole training and development set) are largely those mentioned in the annotated test corpus. Thus, both training sets have the same order of training examples for concepts mentioned in the test corpus, which could explain the similar performance. We note that we reproduced the results previously published for CONTES and HONOR. When both training sets are combined, the results considerably outperform the state-of-the-art systems on the BB3 task by 10 points and previous HONOR results by 4 points. This can be explained by complementary information contained in these training sets, or simply by an increase in the number of training examples.

Method	Similarity score
Baseline	0.54
ToMap	0.66
BOUN	0.62
Turku	0.63
BOUNEL	0.66
CONTES (training) – 2017	0.60
HONOR (training) – 2018	0.73
CONTES (training)	0.61
CONTES (labels)	0.62
CONTES (training and labels)	0.70
HONOR (training)	0.72
HONOR (labels)	0.72
HONOR (training and labels)	**0.76**

Table 4: Results on the BioNLP-ST 2016 Bacteria Biotope normalization task (test set).

6. Conclusion and Discussion

We address the Entity Normalization task by concepts of an ontology under the conditions of scarce training data. This situation is common for technical and scientific domains such as Life Sciences. Using the CONTES and HONOR methods on the BB3 task, we experimented with several strategies and highlighted the relative merits of each.

Both CONTES and HONOR use word embeddings. We have shown that training word embeddings on a reduced but targeted and adapted corpus yields better results than off-the-shelf word embeddings trained on a huge general corpus. Moreover, the reduced size of the training corpus allows us to screen and optimize hyperparameters. The word embeddings vector size has the largest influence on the prediction performance.

We have also shown that token normalization has little, and sometimes a negative influence. Indeed lemmatization, stemming, numeral masking and entity masking did not improve results. On the other hand, stop word filtering produces better word embeddings for this task.

The main result of our experiments is that CONTES achieves similar results when trained on annotated entity mentions or on concept labels and synonyms. Moreover when trained on the union of labels and the annotated training corpus, the results improve even more. This demonstrates that the use of the concept labels and synonyms can overcome the lack of annotated training data, or positively complement small training datasets.

Our hypothesis is that both training sets are complementary because they present different characteristics. On one hand labels and synonyms cover the whole ontology because each concept has at least one label. On the other hand annotated corpora represent more faithfully mentions in texts. Further study should include the performance of the methods specifically on mentions of concepts that have not been mentioned in the training dataset. This could allow us to see the specific contribution of these methods to the problem of zero-data learning.

By exploring parameters and training data, we were able to substantially outperform the state-of-the-art. In order to draw more general conclusions, we plan to apply the approach to other benchmarks, such as BioCreative V Chemical-Disease-Relations (Wei et al., 2015) and TAC Adverse Drug Reaction Extraction from Drug Labels (Roberts et al., 2017), as future work.

The presented strategies all rely on publicly available tools[5] and on adequate data preparation and thus they are within the reach of non-ML specialists.

7. Acknowledgements

We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources.

This work is partially supported by the French National Research Agency under grant ANR-15-CE23-0025-01 (ContentCheck project) and by the CoSO (French Committee for Open Science) under the Visa TM project.

This work was partially supported by the French National Research Agency under grant ANR-15-CE23-0025-01 (ContentCheck project).

8. Bibliographical References

Aronson, A.R. (2001), Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program., *Proceedings of the AMIA Symposium*. American Medical Informatics Association.

Aronson, A.R. and Lang, F.M. (2010), An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association*, 17(3): 229–236.

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2016), Enriching Word Vectors with Subword Information. *ACL 2017*.

Camacho-Collados, J. and Pilehvar, M.T. (2018), On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis, 40–46, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics.

Celikyilmaz, A. (2015), Enriching Word Embeddings Using Knowledge Graph for Semantic Tagging in Conversational Dialog Systems. *AAAI Spring Symposium Series.*

Chiu, B., Crichton, G., Korhonen, A. and Pyysalo, S. (2016), How to Train Good Word Embeddings for Biomedical NLP. *Proceedings of BioNLP16*: 166.

Claveau, V. (2013), IRISA Participation to BioNLP-ST13: Lazy-Learning and Information Retrieval for Information Extraction Tasks, 188–196, *Proceedings of the BioNLP Shared Task 2013 Workshop*. Sofia, Bulgaria: Association for Computational Linguistics.

De Vine, L., Zuccon, G., Koopman, B., Sitbon, L. and Bruza, P. (2014), Medical Semantic Similarity with a Neural Language Model, 1819–1822, *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM'14*. Shanghai, China: ACM Press.

Deléger, L., Bossy, R., Chaix, E., Ba, M., Ferré, A., Bessières, P. and Nėdellec, C. (2016), Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016, 12–22, *Proceedings of the 4th BioNLP Shared Task Workshop*.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018), BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1.

Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E. and Smith, N.A. (2014), Retrofitting Word Vectors to Semantic Lexicons. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Faure, D. and Nédellec, C. (1998), A Corpus-Based Conceptual Clustering Method for Verb Frames and Ontology Acquisition, 5–12, *LREC workshop on adapting lexical and corpus resources to sublanguages and applications.*

Ferré, A., Deléger, L., Zweigenbaum, P. and Nédellec, C. (2018), Combining Rule-Based and Embedding-Based Approaches to Normalize Textual Entities with an Ontology. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Ferré, A., Zweigenbaum, P. and Nédellec, C. (2017), Representation of Complex Terms in a Vector Space Structured by an Ontology for a Normalization Task. *BioNLP 2017*: 99–106.

Gerner, M., Nenadic, G. and Bergman, C.M. (2010), LINNAEUS: A Species Name Identification System for Biomedical Literature. *BMC bioinformatics*, 11(1): 85.

Ghiasvand, O. and Kate, R. (2014), UWM: Disorder Mention Extraction from Clinical Text Using CRFs

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

and Normalization Using Learned Edit Distance Patterns, 828–832, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

Golik, W., Warnier, P. and Nédellec, C. (2011), Corpus-Based Extension of Termino-Ontology by Linguistic Analysis: A Use Case in Biomedical Event Extraction, 37–39, *WS 2 Workshop Extended Abstracts, TIA 2011*.

Grouin, C. (2016), Identification of Mentions and Relations between Bacteria and Biotope from PubMed Abstracts. *Proceedings of the 4th BioNLP Shared Task Workshop*.

Hanisch, D., Fundel, K., Mevissen, H.T., Zimmer, R. and Fluck, J. (2005), ProMiner: Rule-Based Protein and Gene Entity Recognition. *BMC Bioinformatics*, 6(Suppl 1): S14.

Hwang, C.H. (1999), Incompletely and Imprecisely Speaking : Using Dynamic Ontologies for Representing and Retrieving Information. *KRDB*: 13.

Karadeniz, İ. and Özgür, A. (2019), Linking Entities through an Ontology Using Word Embeddings and Syntactic Re-Ranking. *BMC Bioinformatics*, 20(1): 156.

Larochelle, H., Erhan, D. and Bengio, Y. (2008), Zero-Data Learning of New Tasks., 3, *AAAI*.

Leaman, R., Islamaj, D.R. and Lu, Z. (2013), DNorm: Disease Name Normalization with Pairwise Learning to Rank. *Bioinformatics*, 29(22): 2909–2917.

Lee, H.C., Hsu, Y.Y. and Kao, H.Y. (2015), An Enhanced CRF-Based System for Disease Name Entity Recognition and Normalization on BioCreative V DNER Task. *Proceedings of the fifth biocreative challenge evaluation workshop*.

Limsopatham, N. and Collier, N. (2016), Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation, 1014–1023, *ACL 2016*. Berlin, Germany: Association for Computational Linguistics.

Lipscomb, C. E. (2000), Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3): 265–266.

Manning, C., Raghavan, P. and Schuetze, H. (2009), Introduction to Information Retrieval. *Natural Language Engineering*: 581.

McCray, A.T. (1989), The UMLS Semantic Network., 503–507, *Proceedings. Symposium on Computer Applications in Medical Care* (pp. 503-507). American Medical Informatics Association.

Mehryary, F., Hakala, K., Kaewphan, S., Björne, J., Salakoski, T. and Ginter, F. (2017), End-to-End System for Bacteria Habitat Extraction. *BioNLP 2017*: 80.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.H., Torres, R., Krauthammer, M., Lau, W.W., Liu, H., Hsu, C.N., Schuemie, M., Cohen, K.B. and Hirschman, L. (2008), Overview of BioCreative II Gene Normalization. *Genome Biology*, 9(Suppl 2): S3.

Nédellec, C., Bossy, R., Chaix, E. and Deléger, L. (2018), Text-Mining and Ontologies: New Approaches to Knowledge Discovery of Microbial Diversity. arXiv preprint arXiv:1805.04107.

Nédellec, C., Nazarenko, A. and Bossy, R. (2009), Information Extraction, 663–685, *Handbook on ontologies*. Springer.

Pennington, J., Socher, R. and Manning, C. (2014), GloVe: Global Vectors for Word Representation. *EMNLP*.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018), Deep Contextualized Word Representations. *Proceedings of NAACL-HLT*.

Porter, M.F. (1980), An Algorithm for Suffix Stripping. *Program*, 14(3): 130–137.

Pratt, W. and Yetisgen-Yildiz, M. (2003), A Study of Biomedical Concept Identification: MetaMap vs. People. *AMIA Annual Symposium Proceedings*, 2003: 529–533.

Ravi, S. and Larochelle, H. (2017), Optimization as a Model for Few-Shot Learning. *Eighth International Conference on Learning Representations*.

Roberts, K. (2016), Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP, 54–63, *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. Osaka, Japan: The COLING 2016 Organizing Committee.

Roberts, K., Demner-Fushman, D. and Tonning, J.M. (2017), Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track, *TAC*.

Schuemie, M.J., Jelier, R. and Kors, J.A. (2007), Peregrine: Lightweight Gene Name Normalization by Dictionary Lookup, 131–133, *Processing of the Second BioCreative Challenge Evaluation Workshop*.

Sil, A., Kundu, G., Florian R. and Hamza W. (2018), Neural Cross-Lingual Entity Linking, *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tiftikci, M., Sahin, H., Büyüköz, B., Yayikçi, A. and Ozgür, A. (2016), Ontology-Based Categorization of Bacteria and Habitat Entities Using Information Retrieval Techniques. *Proceedings of the 4th BioNLP Shared Task Workshop*: 56.

Tsuruoka, Y., McNaught, J., Tsujii, J. and Ananiadou, S. (2007), Learning String Similarity Measures for Gene/Protein Name Dictionary Look-up Using Logistic Regression. *Bioinformatics*, 23(20): 2768–2774.

Uschold, M. and King, M. (1995), Towards a Methodology for Building Ontologies. *Citeseer*: 15.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. and Chen, C.F. (2007), A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics*, 23(10): 1274-1281.

Wei, C.H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J., Li, J., Wiegers, T.C. and Lu, Z. (2015), Overview of the BioCreative V Chemical Disease Relation (CDR) Task. *Proceedings of the fifth BioCreative challenge evaluation workshop*: 14.

Yen, T.Y., Lee, Y.Y., Huang, H.H. and Chen, H.H. (2018), That Makes Sense: Joint Sense Retrofitting from Contextual and Ontological Information, 15–16, *ACM Press*.