Unsupervised relation extraction from scientific texts using self-organizing maps

Elena Manishina, Mouna Kamel, Cassia Trojahn dos Santos, Nathalie Aussenac-Gilles

Unsupervised relation extraction from scientific texts using self-organizing maps

Elena Manishina, Mouna Kamel, Cassia Trojahn and Nathalie Aussenac-Gilles

IRIT (UT3)
18 ROUTE DE NARBONNE
F-31062 TOULOUSE
firstname.lastname@irit.fr

Résumé : Scientific texts represent a rich source of unstructured knowledge. Extracting this knowledge in a supervised manner can become highly expensive in time and human resources. Moreover supervised models are domain- and language-dependent which make them hard to maintain and extend. Hence unsupervised methods have received a lot of attention from researchers in the fields of information extraction and data mining. In this paper, we present our experiments with self-organizing maps (SOMs) for the task of open relation extraction. We combine contextual features of different level (lemmas and parts-of-speech) to help the algorithm to automatically discover lexical and morphological patterns in the corpus. The evaluation results show that our model yields a better performance than the widely used K-means clustering algorithm with the same feature set.

Mots-clés : Open relation extraction, clustering, self-organizing maps

1 Introduction

With the ever growing quantity of available unstructured texts in various domains containing a large amount of information about the world, the need for efficient knowledge extraction techniques becomes apparent. This is especially true for scientific fields : new entities and the corresponding scientific terms as well as relations between them are discovered and updated regularly. Oftentimes this information is stored in the domain ontologies which need to be maintained and populated with newly discovered terms and relations. Updating ontologies is a laborious process involving domain experts and requiring extensive manual work.

Recent advances in automatic term and relation extraction procedures are designed to solve this problem for domains with significant amounts of annotated data. Unfortunately not all domains have annotated training resources ; most scientific texts available in sufficient quantities do not contain any kind of annotation necessary to train supervised relation extraction models. The manual annotation being costly in time and human resources, more and more researchers turn to unsupervised learning techniques, with various clustering algorithms being among the most widely used ones. A number of unsupervised learning techniques have been successfully implemented and tested on the general domain corpora. However their performance on the domain-specific texts is yet to be evaluated.

In this paper we present our relation extraction model which uses self-organizing maps - a neural-network-based clustering algorithm - which allows for flexible clustering unconstrained by random factors like the initial number of clusters and the the initial centroid coordinates.

A SOM organizes the data, mapping the space of training instances to a two-dimensional neural grid. This grid format allows for data examination as well as interactive assignment and labeling of data clusters. In addition, the SOM grid allows to explore the topology of the data (feature space) and adjust the feature set respectively.

This paper is structured as follows : in section 2 we give a brief outline of the work carried out previously in the domain of unsupervised relation extraction and the application of SOM in similar tasks ; in section 3 we outline the theory behind our implementation of the SOM for the task of relation extraction as well as its place among similar clustering frameworks (like K-means) ; in section 4 we present the experiments with the SOM and the results ; finally we conclude the paper with a discussion and some future directions for our work.

2 Background

Various supervised learning algorithms have been successfully applied to the task of relation extraction (RE) : tree kernels within a Support Vector Machine (Culotta & Sorensen (2004)), Maximum Entropy models Kambhatla (2004), etc.

Neural networks (NN) are among the most recent learning techniques applied to RE. Specifically convolutional deep neural network (DNN) have been used to learn relation representations (Xu *et al.* (2015)) as well as to extract lexical and sentence level features (Zeng *et al.* (2014) and Nguyen & Grishman (2015)).

The lack of annotated training data triggered the interest in semi-supervised and unsupervised learning for RE. In the domain of semi-supervised learning : Chen *et al.* (2006) implements a label propagation (LP) algorithm as applied to RE, Krause *et al.* (2012) learns grammar-based RE rules from the Web by utilizing large numbers of relation instances as seed, etc.

Among the examples of a successful application of unsupervised learning to relation extraction in general domains are : Gonzalez & Turmo (2009) with the adaptation of K-means and Expectation Maximization algorithms ; custom semantic clustering heuristic based on WordNet path distance Eichler *et al.* (2008), etc.

Unsupervised NN algorithms, like Self-organizing map (SOM) which has traditionally been used for data visualization applications has also been successfully applied to NLP domains such as document clustering (Chifu & Cenan (2004)), co-reference resolution (Burkovski *et al.* (2011)) and relation extraction (Bloehdorn & Blohm (2006)).

Unsupervised learning techniques and SOM specifically also showed a good performance comparable to those of supervised algorithms on large datasets where supervised methods often exhaust their computational capacities and end up over-fitting the training data.

3 Methodology

Relation extraction consists in identifying the entities (terms) and semantic relations between them in a corpus of text (a scientific article in our case). For example in a sentence : *Proteinogenic amino acids, such as glutamate (standard glutamic acid) and gamma-amino-butyric acid also play critical non-protein roles within the body.* - the terms **glutamate** and **gamma-amino-butyric acid** are hyponyms of **proteinogenic amino acid**, i.e. they are both types of **proteinogenic amino acid**.

Relation extraction, especially using domain-specific corpus, is an important preliminary step in building domain ontologies and knowledge bases. In this context scientific articles present an invaluable source of knowledge. Apart from the typical unstructured text (paragraphs) scientific articles contain other data representation structures that may contain a lot

of additional information and thus require a separate study : tables, figures, enumerative structures, etc. In this section we present the theoretical ground for our relation extraction model applied to paragraphs of text.

3.1 Relation extraction as a classification problem

We transform the relation extraction into a classification problem : given a pair of terms occurring in the same sentence we want to classify these terms as related or not related and consequently cluster the related terms by the relation type. Each instance pair is represented by the concatenated context vectors of its component terms on the lexical and POS level. The proximity of those vectors in the search space should suggest the similarity between the instances (term couples) and supposedly similar relations between the terms.

3.2 Self-Organizing Maps

The SOM is a type of artificial neural network for unsupervised learning and data visualization which builds the map from input examples using vector quantization and a topological layout of the prototype vectors. SOMs allow for a mapping of high-dimensional input vectors onto a low dimensional output space. A map itself represents a grid of nodes or neurons. Each node is represented by a weight vector of the same length as the input data vectors and a given position on the map. Placing a vector from data space onto the map consists in finding the node with the closest weight vector to the data space vector. The proximity between the two instances on the SOM allows to suggest that the term couples share the same semantic relation. Based on the inherent properties of SOM, we derive the following hypotheses :

— The instances in a cluster point to sets of features that are often shared across contexts, and hence may indicate relatedness of entity pairs.

— The proximity of two term couples on the SOM suggests that these couples have semantically and syntactically similar contexts and thus are related in similar way, i.e. by the same relation.

To evaluate the potential of SOM we compare it to a widely used K-means algorithm which uses the same feature set.

3.3 Word embeddings as training instance format

Concerning data format, one of the important innovations in recent years was the revival of word embeddings (Mikolov *et al.* (2013)) which became the most current training instance format for unsupervised RE (Gupta *et al.* (2016), Hashimoto *et al.* (2015)). The intuition behind this is the following : word embeddings represent a word as a numerical vector, based on the context in which it appears ; this vector representation allows to perform vector operations and more importantly calculate the distance between vectors which can be seen as finding similar patterns in an unsupervised manner based on the distance between the context vectors. These vector properties make them an ideal data representation for unsupervised learning (specifically clustering algorithms).

4 Experiments and Results

In this section we present our experiments with SOM and K-means. Both algorithms use the same corpus and the same feature set.

4.1 Corpus

For our experiments we collected a corpus of articles from the Nature journal of the ISTEX digital library dating from 2000 till 2012. For this first experiment we removed structured data presentations (tables, figures) as well as vertical enumerative structures (lists) from the text. For each type of the removed data structure we constituted a separate corpus which is to be processed separately with a specific term annotation procedure and a different feature set. For the present experiment we kept only the text from the paragraphs. The corpus statistics is presented in Table 1. The first version of the corpus (Nature A) is the one annotated with the terms extracted using the weighed combination of terminological extractors. The second version (Nature B) is annotated with the terms from the NCIT ontology. Both annotation procedures are described in Section 4.2 below. POS-tagging is performed using TreeTagger (Schmid (1995)) in order to build POS context vectors at the feature extraction stage.

TABLE 1 – Corpus statistics

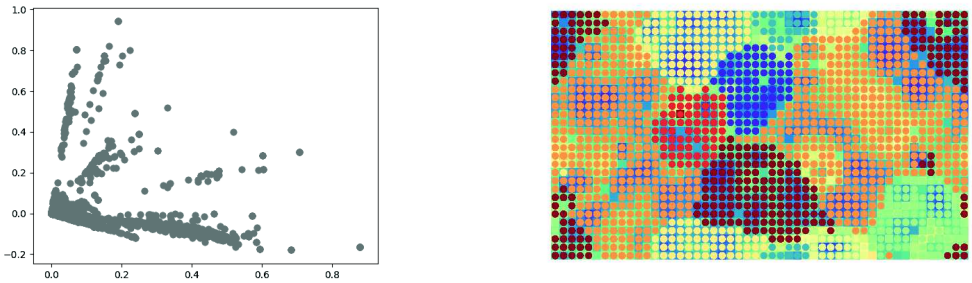
Corpus	**Nsents**	**Nterms**	**Ncouples**
Nature A	81706	2528	42611
Nature B	81706	2412	39611

4.2 Term identification

The original term extraction paradigm included running the weighed combination of the two term extractors on our corpus : Termsuite (Cram & Daille (2016)) and Yatea (Aubin & Hamon (2006)). Combining these tools yields a sufficiently accurate and comprehensive terminological annotation of our corpus. This procedure however does not allow us to test our model as in this case the manual evaluation of the output would be necessary and would require human expertise in the field of medical science. Thus in order to objectively test the viability of our method we resort to a domain ontology. We opted for the National Cancer Institute Thesaurus (NCIT) [1] which combines the vocabulary for clinical care, translational and basic research, as well as the vocabulary for public information and administrative activities. NCIT ontology has a total of 118941 classes and 173 properties. This ontology was selected for having the best coverage on our corpus (76% terminology coverage according to the medical ontology recommendation portal [2])].

Thus to test our implementation of the SOM algorithm we project the terms from the NCIT ontology on the corpus. We obtain 2412 unique terms in total (Table 1). After term projection multi-unit terms are replaced with a single token.

1. https ://ncit.nci.nih.gov/
2. http ://bioportal.bioontology.org/recommender



(a) Feature space

(b) SOM projection

FIGURE 1 – Training instances distribution

4.3 Features

To obtain the training instances we build the term couples according to the procedure described in Section 3. We apply the following constraints : both terms must be in the same sentence and the number of terms in between the two terms under consideration does not exceed 1 (i.e. maximum one term between the two target terms). Thus if a phrase contains 4 terms, 5 term couples will be extracted. For word2vec models we use two different formats of the training corpus : the first format represents the lemmas and the second one - POS tags. We train two models using skip-gram algorithm with the vector size of 100 and the window of 10 words on each corpus. Then for each couple of terms we generate an instance in the form of a feature vector. This feature vector represents a concatenation of the context word embedding vectors of the word tokens and the POS. In order to tackle the inherent computational complexity of the SOM training we opted for a smaller size of the word embeddings than the recommended defaults which resulted in a reduced dimensionality of the feature vectors. Another solution is to take the average of contextual vectors. We exploit both setups in our current experiments ; the results are presented in Section 4.5.

4.4 SOM implementation

For the implementation of the SOM algorithm we used the somoclu [3] python library (Wittek *et al.* (2013)) along with scikit-learn [4] for a K-means implementation. Somoclu is a parallel implementation of self-organizing maps which includes a sparse kernel. For this first experiment we picked a 30x50 grid. Given the dimension of the feature vectors and a relatively high number of instances in the corpus this grid size seemed like a satisfactory compromise between the computation time and the expected accuracy of the data projection. Figure 1b shows the resulting SOM.

3. https ://github.com/peterwittek/somoclu
4. http ://scikit-learn.org/stable/

4.5 Results

We compared our SOM method with a classical K-means algorithm on the same feature set. Here we focus on 8 relations from the NCIT ontology which emerge as the cluster-defining ones :

— Gene_Prod_Plays_Role_In_Bio_Process (GRB)
— Conceptual_Part_Of (CPO)
— EO_Disease_Has_Property_Or_Attribute (DPA)
— Gene_Plays_Role_In_Process (GPRP)
— Gene_Product_Has_Organism_Source (GPOS)
— Procedure_Has_Excised_Anatomy (PEA)
— Procedure_Uses_Manufactured_Object (PMO)
— Procedure_Has_Target_Disease (PTD)

TABLE 2 – F-score : SOM average vectors (F1), SOM concatenated vectors (F2) and K-means concatenated vectors (F3) (30x50 grid)

Relation	F1	F2	F3
EO_Disease_Has_Property_Or_Attribute	0.3	0.25	0.11
Gene_Product_Has_Organism_Source	0.34	0.31	0.24
Gene_Plays_Role_In_Process	**0.59**	**0.22**	0.11
Conceptual_Part_Of	0.44	0.43	0.18
Procedure_Has_Excised_Anatomy	**0.43**	**0.52**	0.25
Gene_Prod_Plays_Role_In_Bio_Process	0.5	0.33	0.08
Procedure_Has_Target_Disease	0.25	0.21	0.14
Procedure_Uses_Manufactured_Object	0.43	0.46	0.1

As we can see from the table above the performance of the classifiers for different relations varies with no one specific trend : in general SOM shows better results than K-means. But the difference in concatenated versus average vector values with SOM does not seem to have a well-defined general pattern and is indeed relation-specific. The difference in F-score for different relations may be explained by the specificities and the variety of the lexical realizations of each relation type, but this phenomenon requires a thorough analysis which is currently under way.

5 Conclusions and Future work

In this article we presented our unsupervised approach to relation extraction based on SOM neural network as applied to paragraph text in the corpus of scientific articles. The approach is language- and domain- independent and does not require external resources (apart from the evaluation stage).

As we mentioned above, one of the objectives of the study was, among other things, developing a paradigm for automatic populating of domain ontologies. Though our model does not reach 100% accuracy and thus the extracted terms and relations cannot be integrated into the

ontology directly, the results of the first raw extraction can be validated and refined by domain experts before integrating the new relations into the ontology.

As for our future work, we extend this approach to cover other information presentation formats : lists, tables, and images.

5.0.1 Acknowledgments

This work is funded by ISTEX project : ANR-10-IDEX-0004-02.

Références

AUBIN S. & HAMON T. (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, p. 380–387. Springer.

BLOEHDORN S. & BLOHM S. (2006). A self organizing map for relation extraction from wikipedia using structured data representations. In *Proc. Int. Workshop on Intelligent Information Access (IIIA-2006)*.

BURKOVSKI A., KESSLER W., HEIDEMANN G., KOBDANI H. & SCHÜTZE H. (2011). Self organizing maps in nlp : Exploration of coreference feature space. In *International Workshop on Self-Organizing Maps*, p. 228–237 : Springer.

CHEN J., JI D., TAN C. L. & NIU Z. (2006). Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 129–136 : Association for Computational Linguistics.

CHIFU E. S. & CENAN C. (2004). Discovering web document clusters with self-organizing maps. *Sci. Ann. Cuza Univ.*, **15**, 38–47.

CRAM D. & DAILLE B. (2016). Termsuite : Terminology extraction with term variant detection. *ACL 2016*, p.1̃3.

CULOTTA A. & SORENSEN J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, p. 423 : Association for Computational Linguistics.

EICHLER K., HEMSEN H. & NEUMANN G. (2008). Unsupervised relation extraction from web documents. In *LREC*, volume 8, p. 1674–1679.

GONZALEZ E. & TURMO J. (2009). Unsupervised relation extraction by massive clustering. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, p. 782–787 : IEEE.

GUPTA N., PODDER S., ANNERVAZ K. & SENGUPTA S. (2016). Domain ontology induction using word embeddings. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, p. 115–119 : IEEE.

HASHIMOTO K., STENETORP P., MIWA M. & TSURUOKA Y. (2015). Task-oriented learning of word embeddings for semantic relation classification. *arXiv preprint arXiv :1503.00095*.

KAMBHATLA N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, p.2̃2 : Association for Computational Linguistics.

KRAUSE S., LI H., USZKOREIT H. & XU F. (2012). Large-scale learning of relation-extraction rules with distant supervision from the web. *The Semantic Web–ISWC 2012*, p. 263–278.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

NGUYEN T. H. & GRISHMAN R. (2015). Relation extraction : Perspective from convolutional neural networks. In *Proceedings of NAACL-HLT*, p. 39–48.

SCHMID H. (1995). Improvements in part-of-speech tagging with an application to german. In *In proceedings of the acl sigdat-workshop* : Citeseer.

WITTEK P., GAO S. C., LIM I. S. & ZHAO L. (2013). Somoclu : An efficient parallel library for self-organizing maps.

XU K., FENG Y., HUANG S. & ZHAO D. (2015). Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv :1506.07650*.

ZENG D., LIU K., LAI S., ZHOU G., ZHAO J. *et al.* (2014). Relation classification via convolutional