RDF triples extraction from company web pages: comparison of state-of-the-art Deep Models[1]

Wouter BAES [a,b], François PORTET [a], Hamid MIRISAEE [b], and Cyril LABBÉ [a]
[a] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France*
[b] *Skopai, 38400 Saint-Martin-d'Hères, France*

Abstract. Relation extraction (RE) is a promising way to extend the semantic web from web pages. However, it is unclear how RE can deal with the several challenges of web pages such as noise, data sparsity and conflicting information. In this paper, we benchmark state-of-the-art RE approaches on the particular case of company web pages, since company web pages are important source of information for Fintech and BusinnessTech. To this end, we present a method to build a corpus mimicking web pages characteristics. This corpus was used to evaluate several deep learning RE models and compared to another benchmark corpus.

Keywords. relation extraction, NLP, linked data, Deep Learning

1. Introduction

Relation Extraction (RE) refers to the process of identifying semantic links between entities in a sentence [1]. As an example, *Bill Gates founded Microsoft*, has *Bill Gates* and *Microsoft* as entities and the *founder* relation as a semantic link between those two. RE has been successfully applied to a wide range of domains such as knowledge base enrichment [2] and Question-Answering [3]. With the extremely fast growth of the internet, web pages are now considered as a very rich source for populating knowledge bases. Those pages, however, contain information in plain text, or in a poorly structured form. Extracting this information is not easy as they suffer from noise and data sparsity. Accordingly, the extracted information can be incomplete, or in conflict with other information. Although RE is a mature technology which has been evaluated on some benchmarks, it is still difficult to predict how it will behave on new datasets different from those of the benchmarks. For instance, in Skopai[2], a company that uses deep learning techniques to analyze and classify startups, one of the objectives is to extract information from company web pages in a form that is exploitable for reasoning. Such company needs an efficient semantics extraction from web-pages to reduce the amount of corrections to be performed by human experts. Furthermore, storing the extracted relations as

[2] `https://www.skopai.com/`

RDF-triples in an ontology allows for reasoning, deduction of implicit information and automatic updating of the information, all of which help with the company's objective.

In this paper, we present the results of a study aiming at comparing current state-of-the-art deep learning RE models to the specific domain of company web pages. To do this, we built a corpus which mimics the characteristics of the desired data.

The contributions of this paper are (1) the construction of a dataset for the task of relation extraction (with a focus on RE from company webpages) and (2) the comparison of several state-of-the-art relation extraction models on different benchmarks.

2. State of the art

Relation Extraction (RE) task is to detect and classify semantic relationship mentions from plain free-text. Several techniques for RE from patterns matching to statistical models have been proposed. However, recent advance in deep learning has made it the current state-of-the-art [1]. In the literature, the task of RE has been studied both in the supervised and unsupervised paradigm. For instance, [4] proposes a CNN-based technique to extract lexical features for relation classification. The biggest drawback of this approach is the need for large amount of high-quality, manually labeled training and test data, which is costly and time-consuming to make [5]. Unsupervised techniques do not require human labor for labeling, but they usually lead to inferior results [6].

To assess the progress of the domain, several benchmarks and challenges have emerged this last decade. In the domain of supervised RE, a widely used benchmark is the SemEval 2010 Task 8 [7] challenge. The dataset used during the challenge, was composed 10k sentences each of which annotated with one of 19 possible relations (9 bi-directional relations, and one *Other*).

In supervised RE, popular Deep Neural Network (DNN) architectures are convolutional neural networks (CNN) and recurrent neural networks (RNN). For instance, [4] proposed a 'CNN + Softmax' model which reached 78.9 % of F-measure on the SemEval 2010 Task 8 challenge. Since then, BERT-based models have shown a definite improvement. For instance, [8] proposed a model called 'BERT-Entity + MTB' which used the representation of entity markers (more specifically, the end markers of those entities) as output of the final hidden transformer layer. MTB signifies that the transformer model used is not regular BERT, but one that has been pre-trained to 'Match the Blanks' (MTB), meaning it got fed sentences with words blanked out, where the goal was to predict what these blanked out words were. This model reached 89.5 % of F-measure on the SemEval 2010 Task 8 challenge far above the CNN model.

These DNN models generally perform well but heavily depend on the availability of a large amount of high-quality, manually labeled training and test data. This is costly and time-consuming in human labor [5]. We partially address this problem by creating semi-automatically a corpus dedicated to company web pages.

3. Method

A general overview of the approach to acquire a new dataset and train RE models from it is given, with the different steps laid out in a schema. The ontology definition and the alignment process are then detailed.

3.1. Overview of the approach

Figure 1 shows the different steps undertaken in this study. To extract semantic information, the list of the concepts and their relations is first defined within an **Ontology**. This process is explained in Section 3.2. At the beginning of the process, we consider a free-text corpus and a set of semantic relations none of which being aligned. For instance a fact such as `founder(Bill Gates,Microsoft)` is given but it is not known which sentence in the corpus describes this fact. These facts, together with the free text sentences, compose the **Unaligned dataset**.

Using the ontology terminology, the dataset facts are processed to **populate** the ontology. This step can be seen as the transformation of an arbitrary semantic information into RDF-triples.

Once the set of RDF-triples are processed, the alignment step seeks the sentences that are the most probably associated to each triple. The aim is to produce an **aligned dataset** where sentences are annotated with the triples. This process is detailed in Section 3.2. The aligned output can then be used to train and evaluate some of the deep learning **models** described in Section 2.

3.2. Ontology definition

To build the reference ontology for description of companies we used the DBpedia [3] OWL structure, in particular the relations linked to the `Organisation` and `Company` entities since it already contains most of the needed relations. It also plays the role of a *top-ontology* where the links between the classes are established and could be used to infer further information. Moreover, using DBpedia makes it much easier to ensure the interoperability.

The ontology was then confronted to the professional Skopai database, by looking at the possible attributes that can be present in the collections. Not all of these attributes can be modeled using the relations already extracted from DBpedia. Hence, some extra predicates such as those related to patents, funding and awards were added to the final ontology.

3.3. Data to Triple alignment

The alignment consists of matching two sources of information, each with a list of sentences and a list of RDF-triples. Each of those sentences may or may not actually describe one or more of the triples. Hence, the objective is to align those sentences with

[3] `https://wiki.dbpedia.org/`

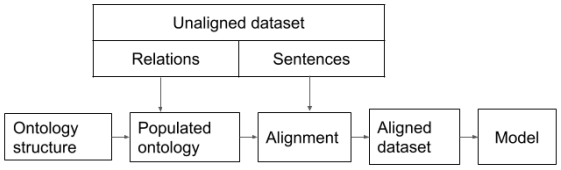

Figure 1. Overview of the different steps of this study.

triple(s) they describe, discarding those which describe no relation. For example, consider two sentences: *Bill gates founded Microsoft* and *Microsoft was founded in 1975* and two RDF triples: `founder(Microsoft, Bill Gates)` and `location(Microsoft, Redmond, USA)`. From all of this information we know that Microsoft was founded in 1975 by Bill Gates, and has its headquarters in Redmond, USA. However, only the fact that Bill Gates is the founder is present in both the sentences and the triples. So the only alignment that can be made is *Bill gates founded Microsoft* with `founder(Microsoft, Bill Gates)`.

To perform this task we used the alignment tool[4] built in the context T-REx [9]. T-REx is a large aligned dataset of 3.09 million Wikipedia abstracts (6.2 million sentences) with 11 million Wikidata[5] triples. The tool aligns the sentences with triples using distantly supervised learning triple aligners, more specifically those specified in [5].

4. Experiments

In this section, we present the corpora that have been used, the result of the alignment process of one corpus and the performance of the RE models presented in Section 2.

4.1. The Corpora

To assess the performances of RE on company texts, we used the Wikipedia Company Corpus (WCC)[6] [10,11]. This was constructed for the automatic generation of company descriptions from a set of relations. The WCC consists of a 43 980 companies extracted from Wikipedia. Each company example comes with an abstract and a list of at least two attribute-value pairs. The total amount of sentences in the corpus is 159 710, an average of just under four sentences per abstract. However, the attribute-value pairs were not aligned with the text. Even worse, since it is a real noisy corpus, some attribute-value pairs are not present in the Wikipedia text and vice-versa. Hence, this corpus needs to be aligned.

To evaluate the models on clean conditions, we also used the dataset released with the WebNLG 2020 challenge[7], which already comes aligned. The dataset contains a total of 16 categories including the *Company* category. An example of text and its corresponding triples is shown Figure 2. The amount of triples per sentence ranges from 1 to 7. As of the time of writing, the test set has not yet been released, only a training and a development set. The training set consists of 13 229 entries (in the form of a set of triples) with 35 415 texts (3 to 5 texts per entry). The development set consists of 1669 set of triples with 4 468 texts. The amount of instances per category ranges from 299 for *Monument* to 1591 for *Food*.

4.2. Results of the alignment

The output produced by the T-Rex pipeline on WCC consists of 193 203 triples over 108 227 sentences, concerning 34 299 companies. The distantly supervised approach was

[4] `https://github.com/hadyelsahar/RE-NLG-Dataset`
[5] `https://www.wikidata.org/wiki/Wikidata:Main_Page`
[6] `https://gricad-gitlab.univ-grenoble-alpes.fr/getalp/wikipediacompanycorpus`
[7] `https://webnlg-challenge.loria.fr/challenge_2020/`

Table 1. Relation distribution of the aligned Wikipediacompanycorpus.

Relation	Percentage	Relation	Percentage
location	33.8%	services	3.9%
industry	27.5%	ownedBy	3.4%
products	19.3%	defunct	1.6%
foundedIn	10.2%	numberOfEmployees	0.2%

able to recognize and align the majority of the sentences and the semantic facts. As the WCC corpus is noisy, perfect alignment was not expected. Looking at the distribution of the triples in Table 1, it is unsurprising that the biggest part, describes the *location* relation, as it is available for almost every company. At the other end of the spectrum, the *numberOfEmployees* relation is often present as fact but rarely in the abstract, so there is very little alignment possible.

Hereafter, when referring to the Wikipediacompanycorpus (or WCC), we are referring to the aligned version described in this section.

4.3. Evaluation

The models that were evaluated are the ones mentioned in Section 2, with three different variants of the BERT-based model. **BERT** does not use the entity markings that **BERT-Entity** uses, and only **BERT-Entity + MTB** uses the MTB pre-trained model.

The results are reported in Table 2. The WCC corpus was randomly split into 64/16/20 % for training/development and testing. For the WebNLG 2020 challenge, since the test set has not yet been released, the reported results are those of the development set. This means that the result presented should be higher than when using a true test set. For WebNLG, it can be seen that BERT-Entity performs the best, followed by BERT-Entity+MTB then BERT and then CNN. The results obtained from experimenting on WebNLG points to BERT-Entity as the best model to use. For WCC, CNN + Softmax

Table 2. Overview of the obtained results for WebNLG 2020 and WCC.

Model	WebNLG 2020				WCC			
	Accuracy	P	R	F1	Accuracy	P	R	F1
CNN + Softmax	97.70%	93.62%	93.12%	93.02%	87.77%	86.99%	82.84%	84.64%
BERT	98.47%	94.08%	94.89%	94.30%	91.08%	89.95%	89.25%	89.58%
BERT-Entity	**98.92%**	96.60%	**97.33%**	**96.81%**	91.16%	90.12%	89.51%	89.79%
BERT-Entity + MTB	98.74%	**96.74%**	96.79%	96.58%	**91.19%**	**90.26%**	**89.52%**	**89.87%**

performs the worst overall as well, while BERT-Entity + MTB performs the best over all metrics.

Trane, which was founded on January 1st 1913 in La Crosse, Wisconsin, is based in Ireland. It has 29,000 employees.

```
<entry category="Company" eid="Id21" shape="(X (X) (X) (X) (X))" shape_type="sibling" size="4">
    <modifiedtripleset>
        <mtriple>Trane | foundingDate | 1913-01-01</mtriple>
        <mtriple>Trane | location | Ireland</mtriple>
        <mtriple>Trane | foundationPlace | La_Crosse,_Wisconsin</mtriple>
        <mtriple>Trane | numberOfEmployees | 29000</mtriple>
    </modifiedtripleset>
</entry>
```

Figure 2. Example of text and triples extracted from WebNLG2020 dataset [12]

5. Discussion and further work

In this study, we have explored the performance of several RE models on corpus about company information. For this purpose, we have created an aligned corpus using the Wikipediacompanycorpus and augmenting it through the T-REx alignment pipeline. This gave a set of aligned sentences and RDF-triples, that are specific to companies. A big shortcoming of the corpus in its current form is the absence of negative training samples, sentences that do not contain any relations, or irrelevant ones. This is needed because there is an overwhelming amount of noise as wells as irrelevant information in company web pages.

After evaluating on WebNLG and WCC, BERT-Entity comes forward as the most accurate and most consistent model. In some cases, using the MTB pre-trained model improves results, but not enough to warrant its use (taking into account the need for memory and time intensive pre-training).

Future work for this project includes the improvement of the constructed corpus (with e.g. negative training samples and data from Skopai database), the implementation of the best model to be used by Skopai as well as the possibility to change language focus with different variants of BERT and a way to automatically handle ontology population and ontology evolution.

References

[1] Kumar S. A survey of deep learning methods for relation extraction; 2017. ArXiv preprint arXiv:1705.03645.

[2] Trisedya BD, Weikum G, Qi J, Zhang R. Neural Relation Extraction for Knowledge Base Enrichment. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 229–240.

[3] Li X, Yin F, Sun Z, Li X, Yuan A, Chai D, et al. Entity-Relation Extraction as Multi-Turn Question Answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 1340–1350.

[4] Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation Classification via Convolutional Deep Neural Network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics; 2014. p. 2335–2344.

[5] Augenstein I, Maynard D, Ciravegna F. Distantly supervised web relation extraction for knowledge base population. Semantic Web. 2016;7(4):335–349.

[6] Elsahar H, Demidova E, Gottschalk S, Gravier C, Laforest F. Unsupervised open relation extraction. In: European Semantic Web Conference; 2017. p. 12–16.

[7] Hendrickx I, Kim SN, Kozareva Z, Nakov P, Ó Séaghdha D, Padó S, et al. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden: Association for Computational Linguistics; 2010. p. 33–38. Available from: `https://www.aclweb.org/anthology/S10-1006`.

[8] Baldini Soares L, FitzGerald N, Ling J, Kwiatkowski T. Matching the Blanks: Distributional Similarity for Relation Learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 2895–2905.

[9] Elsahar H, Vougiouklis P, Remaci A, Gravier C, Hare J, Laforest F, et al. T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA); 2018. p. 3448–3452.

[10] Qader R, Jneid K, Portet F, Labbé C. Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In: Proceedings of the 11th International Conference on Natural Language Generation; 2018. p. 254–263.

[11] Qader R, Portet F, Labbe C. Semi-Supervised Neural Text Generation by Joint Learning of Natural Language Generation and Natural Language Understanding Models. In: 12th International Conference on Natural Language Generation (INLG 2019). Tokyo, Japan; 2019. Available from: https://hal.archives-ouvertes.fr/hal-02371384.

[12] Gardent C, Shimorina A, Narayan S, Perez-Beltrachini L. Creating Training Corpora for NLG Micro-Planners. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2017. p. 179–188. Available from: http://www.aclweb.org/anthology/P17-1017.