Call for Discussion:
Building a New Standard Dataset for Relation Extraction Tasks

Teresa Martin and **Fiete Botschen** and **Ajay Nagesh** and **Andrew McCallum**
University of Massachusetts
140 Governors Drive
Amherst, MA 01003, USA
tmartin, fbotschen, ajaynagesh, mccallum@cs.umass.edu

Abstract

This paper is an attempt to raise pertinent questions and act as platform to generate fruitful discussions within the AKBC community about the need for a large scale dataset for relation extraction. For proper training and evaluation of relation extraction tasks, the weaknesses of datasets used so far need to be tackled: mainly the size (too small) and the amount of data that is actually labelled (unlabelled data leading to recall problems). We have the vision of building a new large and fully labelled dataset for entity pairs connected via binary relations from both Freebase as well as other datasets, such as Clueweb. Concerning the process of building, we present pioneering work on a roadmap which will serve as the foundation for the intended discussion within the community. Points to discuss arise within the following steps: first, the source data has to be preprocessed in order to ensure that the set of relations consists of valid relations only; second, we suggest a method to find the most relevant relations for an entity pair; and third, we outline approaches on how to actually label the data. It is necessary to discuss several key issues in the process of generating this dataset. This will enable us to thoroughly create a dataset that will have the potential to serve as a standard to the community.

1 Motivation

A challenging problem for artificial intelligence is Information Extraction (IE) - extracting structured facts from raw unstructured text. A particularly important instance of this is Relation Extraction (RE), the detection of mentions of semantic relationships between entities in text. A typical RE task is classifying relations as 'true' or 'false' when looking at pairs of entities: e.g., for the entity pair (ent_1, ent_2) = ('Barack Obama', 'Michelle Obama'), decide whether the relation rel = 'marriedTo' is 'true' or 'false'. Two of the most widely used datasets for this task are the NYTimes dataset (Riedel et al., 2013) and FB15k (Bordes et al., 2013) (and its extension, FB15k-237 (Toutanova et al., 2015)). The latter is based on Freebase (FB) (Bollacker et al., 2008), a manually constructed knowledge base (KB) of entity pairs linked via relations. Table 1 clarifies how the labelling structure of FB is used for binary RE tasks.

	rel_1	rel_2	...	rel_x
(ent_1, ent_2)	1	0	...	0
(ent_1, ent_2)	0	1	...	0
...	...	...	...	...
(ent_m, ent_n)	0	0	...	0

Table 1: Labelling structure of Freebase for binary RE tasks. Rows: all entity pairs. Columns: all relations. A cell is labelled with 1 as 'true' if its connection of entity pair and relation is contained in FB. The label 0 means that this combination of relation and entity pair is not contained in FB - which can be either that it is actually 'false' or that it is actually 'true' but not contained because no one added it.

The first problem is the small size of FB. Statistics on the incompleteness of FB can be found in (Min et al., 2013). For instance, 78% of person entities do not have nationalities. Also, when training on FB15k it might happen that there is only one single

	FB and NYTimes	FB and Toutanova CW
documents	1.8m	800m
entities	n.a.	14.5k
entity pairs	418k	2m
relation types	4k	2.7m
relation types from FB	1.5k	237
relation types from text	4k	2.7m

Table 2: Statistics over the data used in different works. Rows: the statistics. Columns: the different works.

case of a relation. If this case is in the test set it cannot be learned at all. Thus, evaluation is restricted to relation types within the FB schema. To extend and broaden the coverage of FB many approaches include unlabelled data from text corpora. However the choice of the unlabelled text varies between different contributions, see table 2 for a comparison of dataset statistics between two commonly used datasets. For RE and Universal Schemas (Riedel et al., 2013), data is used from the NYTimes corpus (Sandhaus, 2008) and FB. In comparison to other unlabelled corpora, the NYTimes corpus is quite small. To represent text, data could be used from FB and a (subset of) already existing datasets. For example, FB15k-237 (Toutanova et al., 2015) is a combination of Clueweb(CW) and FB data. CW is a webcrawl that comes annotated with FB entities (Gabrilovich et al., 2013). In the work with FB15k-237, relations are expressed as parse paths but not as entire text and in addition the majority of test entity pairs have no textual or other KB evidence. In terms of increasing quality of FB, we suggest to increase the number of relations by adding data from other datasets. As mentioned above, this could be CW or other already available datasets. By that, more relations should be covered as well as evaluating against entities which are not included in FB is possible.

The second problem is the so called 'recall problem' in FB with misleading results concerning recall when evaluating RE tasks. As illustrated in figure 1, the measure of recall is misleading when non-existent relations for entity pairs in FB are assumed to be 'false' just because they do not appear so far. This assumption is dangerous as non-existent relations in FB could indeed be 'true'. The recall problem occurs in the evaluation of a lot of work on RE due to incomplete labelling, e.g., in (Riedel et al., 2013) evaluating on (1.5k) FB relation types and in (Toutanova et al., 2015) evaluating on 237 FB relation types.

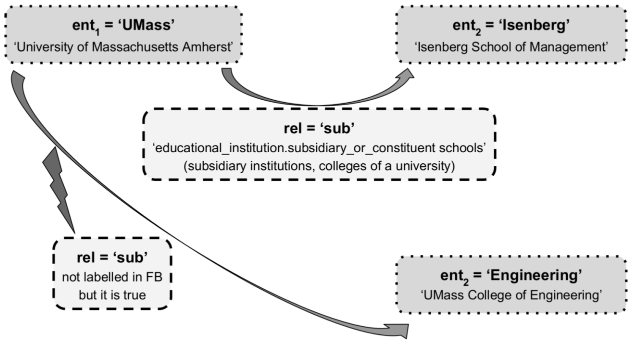
ent_1 = 'UMass'
'University of Massachusetts Amherst'
ent_2 = 'Isenberg'
'Isenberg School of Management'
rel = 'sub'
'educational_institution.subsidiary_or_constituent schools'
(subsidiary institutions, colleges of a university)
rel = 'sub'
not labelled in FB
but it is true
ent_2 = 'Engineering'
'UMass College of Engineering'

Figure 1: Situation leading to recall problem. The entities ent_1 = 'UMass' and ent_2 = 'Isenberg' are connected via the relation rel = 'sub'. UMass Amherst has several other subsidiary colleges which do exist in FB e.g., ent_2 = 'Engineering' but this relation does not exist for this entity pair in FB. Evaluation is misleading if non-existence is interpreted as 'false'.

In order to avoid the recall problem, a fully labelled version is needed, where for each entity pair all relations should be labelled as 'true' or 'false'. Another source evoking the recall problem are entity mentions not being identified, or being attached to wrong types and hence recall is decreased. The recall problem is also claimed and tackled in the context of work on logical background knowledge (Rocktäschel et al., 2015). There, it is approached with a method avoiding manual labelling: simple logic formulae over patterns and relations are used to incorporate additional domain knowledge. The authors mention manual incorporation of such additional knowledge as an avenue for further research. Our effort is similar in spirit to TAC KBP tasks [1]. The number of documents given as input in KBP tasks are fairly small (around 50,000 articles) since the evaluation set is created by human annotators by labelling the responses from all participating teams. The KBP tasks are aimed at benchmarking various approaches and not to create a benchmark dataset

[1] http://www.nist.gov/tac/2015/KBP/ColdStart/guidelines/TAC_KBP_2015_ColdStartTaskDescription_1.1.pdf

that can be used for training. A filtering approach as proposed by us is not adopted in KBP. Our filtering approach becomes essential due to the sheer size of CW corpus. Our proposal is similar to the FACC1 annotation on CW (Gabrilovich et al., 2013). The difference is that FACC1 annotations are aimed at entity linking where as our approach is aimed at relation extraction. We want to discuss a roadmap for a labelling procedure involving human labelling.

2 Roadmap for Creation of the Dataset

2.1 First: Preprocessing of Source Data

Working with a crawl of web data (e.g., CW) yields one preprocessing problem: how can one extract usable relations from this kind of source data? A usable relation is a word sequence that actually is a relation for at least one entity pair. Filtering usable relations out of crawled data can be approached in several ways: for example, the word sequence between the entity pair can be considered as the relation. This is the approach taken in (Verga et al., 2015) and in many other work on open-domain IE. But this presents another hurdle: not all such word sequences are valid relations. To give an example, the sequence *'as well as'* may occur between two entities in a sentence like *'Max as well as Peter are happy.'* however this sequence is not an interesting candidate for any binary relation in IE. As a solution to this, patterns which do not make sense can be excluded by a fixed set of rules, e.g. 'must contain a verb', 'must not contain personal pronouns', 'must contain at least n words' etc. A commonly used approach related to fixed rules is to detect appositives as in (Yao et al., 2013). Coming up with a fixed rule set is problematic though as there is always the risk of excluding patterns that are actually usable. As an alternative to fixed rule sets one can come up with learned models to select usable relations. Another approach is to use dependency parsers such as (Chen and Manning, 2014) or openIE style RE (Verga et al., 2015). However, dependency parsers often focus on the syntax of text which is not ideal given the need of propositional information. Hence, important information might get lost. Others use dependency trees (Stanovsky et al., 2016) in order to explore propositional structure of text. This in return might help to decide whether a certain fraction of the text is actually a usable relation. Data from web crawlers is not only noisy, but also there are a lot of relation candidates which are just too seldom to be learned. It would be an option to follow (Riedel et al., 2013) by excluding entity pairs and relations that occur less than 10 times in the corpus. To sum it up, selecting usable relations when working with big data plays an important role. More unusable relations lead to worse results inevitably. Also, openIE patterns are too noisy to be worthwhile for downstream processing tasks which require RE. Hence, finding good ways to filter unstructured texts is an integral part of constructing the proposed dataset.

2.2 Second: Reduction to Relevant Relations

The number of relations to label for each entity pair is immense: possibly more than a million, depending on how much of the non-valid relation are filtered out during preprocessing. This number is so high as it expresses the number of unique relation instances instead of classes of relations e.g., not the class 'marriedTo' but all mentions like 'married', 'married to', 'happily married', etc. are counted as relations. It is not feasible to label all relations for each entity pair manually. Therefore, other approaches to obtain a full labelling need to be discussed. It is easy to observe that the most of these relations do not make sense at all regarding a specific entity pair. Those would be labelled as 'false' anyway and therefore do not require manual checking. This is why the following question arises: can we find a reasonable way to automatically select the most relevant relations for an entity pair? This would reduce the effort of labelling drastically. To achieve this reduction, we suggest to discuss the following: once there is a representation of entity pairs as well as relations in a common space, the relations closest to an entity pair can be determined. The Universal Schema model as presented in (Riedel et al., 2013) and developed further in (Verga et al., 2015) seems to fit to this need of representation learning. By training the baseline Universal Schema model on the combination of FB and CW data, embeddings are learned jointly for the entity pairs and for the relations. To find the most relevant relations for a given entity pair of interest, the cosine distance between the vector embedding of this entity pair and the vector embeddings of all the relations can be calculated, respectively. The result is a relevance ranking over all the relations specific to this entity pair. From this point, a fixed number of most relevant relations can be selected for manual labelling.

2.3 Third: Labelling Procedure

Table 3 sketches how the data to label could be structured. Labelling of relations has to be done entity pairwise because both the observed relations as well as the ranking over relevance of non-observed relations is specific to an entity pair. Looking at one specific entity pair, some relations might exist in FB already and we suggest to label them 'true'; some relations might be observed in CW and we suggest to accept them as 'true' but probably still consider a checking. Most importantly, all the remaining relations (ranked by relevance for this pair) are not observed and therefore labelled with 'false' by default. All of these default false relations need to be checked and turned to 'true' accordingly. In order to obtain a first fully labelled subset to experiment with, we suggest the following: for $c = 3$ different chosen entities, take all their entity pairs and label their first $r = 300$ most relevant relations. It is difficult to estimate the time required to label even one entity pair's $r = 300$ most relevant relations because this depends on two factors: (a) the quality of the relations; e.g., if relations are not filtered cautiously and even non-valid relations are listed it is quick to tag those relations as 'false' (b) the depths to which the person who labels is familiar with the entity pair; e.g., someone who knows a lot about an entity pair has to do less research in order to decide whether a relation is 'true' or 'false'.

3 Topics for Discussion

First of all, feedback on the need of the proposed dataset is important in order to shape the procedure of building it to the actual needs of the community.

Furthermore, many points of discussion arise along the suggested roadmap for creating the dataset. Concerning the preprocessing, section 2.1:

- Which way to go to get potential relations out of the raw sentences from CW containing entity pairs?
- How to deal with entities annotated with confidence scores in CW and thus how to identify entity mentions reliable?
- Possibilities to map CW relations to corresponding FB relations?

Concerning the reduction to relevant relations, section 2.2:

- Weighing the danger of relying on ONE model trained on the dataset we want to improve (Universal Schema model) in order to learn the embeddings?
- Which measure to determine the closest relations for an entity pair?
- Possibilities of selecting valid relations from CW and reducing to most relevant relations in an end-to-end way?

Concerning the labelling, section 2.3:

- Is it sustainable to just label the non-observed relations even for CW?
- How to minimize influence of the labeller?
- Use crowd sourcing?

Finally, it will be difficult to compare the results on the new dataset with results on previous datasets. For comparison, the datasets should be the same with the only difference that the new one has some more relations labelled as 'true'.

4 Conclusion

Following the need for a large and fully labelled dataset for training and evaluating RE tasks, we presented pioneering explorations on how to build such a dataset out of FB and CW. The purpose of this paper is to provide a platform to facilitate discussions within the community to gather ideas, needs, opinions and feedback all of which will help in the further development of the suggested dataset.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

	observed in FB	**observed in CW**		**not observed in FB and CW**	
	(any number)	(any number)		(all remaining, most relevant)	
specific entity pair (e_1, e_2)	rel_{F1} $\|(e_1, e_2)$	rel_{C1} $\|(e_1, e_2)$	rel_{C2} $\|(e_1, e_2)$	rel_{not1} $\|(e_1, e_2)$	rel_{rnot2} $\|(e_1, e_2)$
(‘UMass’, ‘Isenberg’)	‘subsidiary school’	‘,’	-	‘educational institution’	‘released a documentary on’
(‘UMass’, ‘Boston’)	-	‘is located just 90 miles from’	‘at’	‘initially came to’	‘is located on the slopes of’

Table 3: Sketch of the necessary labelling. Rows: entity pairs (e_1, e_2). Blocks of columns listing relations for the row’s entity pair that exist in FB (first), in CW (second) or do not occur in neither of them (third). Regarding labelling, the first block (FB) is already ‘true’, the second block (CW) is ‘true’ but should be checked and the third block is where a labelling procedure is needed.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). *Note: http://lemurproject. org/clueweb09/FACC1/Cited by*, 5.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *HLT-NAACL*, pages 777–782.

Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. In *Joint Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL ’13)*, June.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Proceedings of the 2015 Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. *ACL Association for Computational Linguistics*.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2015. Multilingual relation extraction using compositional universal schema. *arXiv preprint arXiv:1511.06396*.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2013. Universal schema for entity type prediction. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 79–84. ACM.