Construction and Alignment of Multilingual Entailment Graphs for Semantic Inference

Sabine Weber
University of Edinburgh
s.weber@sms.ed.ac.uk

Mark Steedman
University of Edinburgh
steedman@inf.ed.ac.uk

Abstract

This paper presents ongoing work on the construction and alignment of predicate entailment graphs in English and German. We extract predicate-argument pairs from large corpora of monolingual English and German news text and construct monolingual paraphrase clusters and entailment graphs. We use an aligned subset of entities to derive the bilingual alignment of entities and relations, and achieve better than baseline results on a translated subset of a predicate entailment data set (Levy and Dagan, 2016) and the German portion of XNLI (Conneau et al., 2018).

1 Introduction

Semantic inference is necessary whenever a target meaning is to be inferred from a text. Entailment and paraphrasing are of key importance in question answering and semantic parsing. The question "Who owns Yahoo?" might be answered by a sentence like "Verizon acquired Yahoo in 2016.",which does not correspond with the question but rather entails or paraphrases the answer. A form-independent semantic representation for natural language is needed to bridge this gap.

A new layer of complexity is added to this problem when considering a multilingual case, where a question in one language can only be answered from text in a different language. An alternative to translating either questions or answers is parsing both questions and answer candidates into a form- and language-independent semantic representation that allows for semantic inference.

Previous approaches (Berant et al., 2015) (Hosseini et al., 2018) have concentrated on learning predicate entailment relations (e.g. buying entails owning) from large amounts of monolingual text, creating predicate entailment graphs to enhance form dependent semantics. In this meaning representation each predicate is represented by a cluster containing all its paraphrases and entailments.

In this paper we create paraphrase clusters and entailment graphs in both English and German from a large corpus of non-parallel news text. We align them using link prediction methods for knowledge graphs and present further ideas how alignment and entailment can inform each other. We then address methods for testing the aligned graphs.

2 Related Work

(Lewis and Steedman, 2013a) approach the problem of paraphrase in a multilingual context by creating aligned paraphrase clusters. They use Wikipedia articles describing the same topic, which can be considered a form of parallel text. Even though there is no perfect alignment between sentences, the fact that the same topics are covered is helpful for the construction of paraphrase clusters. Lewis and Steedman use the inter-language links between named entities provided in the Wikipedia articles to align predicates in different languages. Similarly they use the types that were provided with the Wikipedia links.

Lewis and Steedman could work this way with perfect alignment of named entities to derive the alignment of predicates. Working with no parallel data poses new challenges. Because the linking of German named entities to an external data base like Freebase (Bollacker et al., 2008) or DBPedia (Lehmann et al., 2015) is slow and incomplete, only a partial overlap between named entities can be achieved (only about 20% of all named entities in the corpus can be linked to DBPedia). Where types can not be derived from DBPedia links other methods have to be employed.

Paraphrase clusters can be seen as a precursor to entailment graphs. Due to the use of cosine similarity as a metric to construct the clusters, they contain both entailments and paraphrases. Typed predicate entailment graphs were first introduced by (Berant et al., 2015), but were not scalable to a large amount of data. (Hosseini et al., 2018) apply a scalable method that learns globally consistent similarity scores.

3 Method

The construction of entailment graphs can be divided into three steps: (1) extraction, (2) linking and typing, and (3) the calculation of similarities and construction of paraphrase clusters or entailment graphs. Those steps are completed monolingually for German and English. For English we use the pipeline introduced by (Hosseini et al., 2018).

For German we first use dependency parsing and named entity recognition to extract triples of binary relations (e.g. the triple "besuchen::Angela Merkel::Staatsoper" (visit::Angela Merkel::State Opera House)). The linking and typing step poses difficulties for German because the use of state-of-the-art named entity linkers is not feasible for our data size. Instead, we use string matches with DBPedia URIs. Typing is performed using the the Stanford Named Entity recognition component (Finkel et al., 2005) for named entities and GermaNet (Hamp and Feldweg, 1997) for general entities. We create paraphrase clusters using the method introduced by (Lewis and Steedman, 2013b) and construct entailment graphs using the method introduced by (Hosseini et al., 2018).

While Lewis and Steedman used the linked named entities of different languages to align predicates, we only have a partial linking. One way to approach this problem is to model the triples of two entities and their respective relation as triple in a knowledge graph. We use a link prediction method introduced by (Trouillon et al., 2016). We employ the partial linking of entities to get an alignment of both entities and relations. This link prediction method creates an embedding of entities and relations. Instead of embedding a monolingual knowledge graph we pass all German and English relation triples to the method. Entities that are linked are only represented once. This way the learning algorithm sees linked entities with both German and English relations and therefore will embed them close to each other in vector space.

4 Results and Discussion

Evaluating the alignment of predicate entailment graphs poses a challenge. Existing work has not yet explored multilinguality. One approach is to evaluate entailment and alignment separately.

There is still no sufficiently large predicate entailment data set for German. We run preliminary experiments using paraphrase clusters on the German portion of the XNLI corpus (Conneau et al., 2018), which is only of limited use because most of its entailments are not predicate entailments. Nevertheless our method performs 2.7% over random. Hosseini et al. use the predicate entailment data set introduced in (Levy and Dagan, 2016). We translate 1946 sentence pairs from it and perform 5.6% over random.

One way to assess the alignment is to use a data set for bilingual dictionary alignment like the MUSE data set (Conneau et al., 2017) although verbs only make up a small portion of it. Also the relations extracted by us contain verb modifier constructions such as "try to reach" or light verb constructions like "have a border with", which are not represented in such data sets.

Preliminary qualitative results appear promising. In the intermediate step of paraphrase clustering, we see paraphrase clusters like "formulieren, schreiben" (formulate, write) and "darstellen, zeigen" (represent, show). For the full entailment results, we get e.g. "warten zu kommen" and "hoffen zu gelangen" entails "versuchen zu gelangen" (waiting to come, hoping to reach, trying to reach).

This paper presented ongoing work on construction and alignment of predicate entailment graphs from different languages without parallel text. Aligned entailment graphs are a preliminary step to a form- and language-independent semantic that can be used for various natural language understanding task like for example question answering from multilingual text. As a next step we plan to incorporate the entailment links between predicates to support the alignment. Preliminary tests show that the incorporation of entailment links for link prediction in knowledge graphs leads to significant improvement in performance. Translating the full data sets or creating a similar sets for German is another avenue for future work.

References

Jonathan Berant, Noga Alon, Ido Dagan, and Jacob Goldberger. 2015. Efficient global learning of entailment graphs. *Computational Linguistics*, 41(2):249–291.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. *Automatic information extraction and building of lexical semantic resources for NLP applications*.

Mohammad Javad Hosseini, Nathanael Chambers, Siva Reddy, Xavier R Holt, Shay B Cohen, Mark Johnson, and Mark Steedman. 2018. Learning typed entailment graphs with global soft constraints. *Transactions of the Association for Computational Linguistics*, 6:703–717.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Omer Levy and Ido Dagan. 2016. Annotating relation inference in context via question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 249–255, Berlin, Germany, August. Association for Computational Linguistics.

Mike Lewis and Mark Steedman. 2013a. Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.

Mike Lewis and Mark Steedman. 2013b. Unsupervised induction of cross-lingual semantic relations. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 681–692.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.