Unsupervised Conceptualization and Semantic Text Indexing for Information Extraction

Eugen Ruppert(✉)

FG Language Technology, Technische Universität Darmstadt,
Darmstadt, Germany
ruppert@lt.informatik.tu-darmstadt.de

Abstract. The goal of my thesis is the extension of the Distributional Hypothesis [13] from the word to the concept level. This will be achieved by creating data-driven methods to create and apply conceptualizations, taxonomic semantic models that are grounded in the input corpus. Such conceptualizations can be used to disambiguate all words in the corpus, so that we can extract richer relations and create a dense graph of semantic relations between concepts. These relations will reduce sparsity issues, a common problem for contextualization techniques. By extending our conceptualization with named entities and multi-word entities (MWE), we can create a Linked Open Data knowledge base that is linked to existing knowledge bases like Freebase.

1 Motivation

The current NLP research is moving from the linear word/sentence/discourse representation towards semantic representations, where entities and their relations are made explicit. Even though NLP components are still being improved by emerging techniques like deep learning, the quality of existing components is sufficient to work on the semantic level – one level of abstraction up from surface text. We want to semantify text by assigning word sense IDs to the content words in the document. Working on the semantic level does not only provide us with entities like nouns, but also with their relations between each other. After semantification, we can use this representation to grasp the meaning of the document, e.g. what are the subjects in the document and to which class do they belong.

Semantic Web (SW) applications use entities, e.g. to disambiguate which *Turkey* the text refers to: the country or the animal[1]. However, knowledge bases like Freebase [6] or DBpedia [8] only relate concepts, not necessarily disambiguating senses. While linking words to such knowledge bases is useful in their current state, we propose an all-word conceptualization where all content words are identified by their senses.

[1] Throughout this proposal, we are using *italics* for text examples, underscores for hypernyms and `monospace text` for technical details, e.g. explicit context features.

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 853–862, 2016.
DOI: 10.1007/978-3-319-34129-3_54

With our symbolic conceptualizations we are able to identify concepts in a text (contextualization). The contextualization of a document text will improve the performance of entity linking for the 'classic Semantic Web' because we assume that concepts/entities also follow the Distributional Hypothesis – concepts co-occur with similar concepts in a document. Identification of all content words in a document will enable semantic applications like semantic indexing.

2 State of the Art

2.1 Conceptualizations

The creation of conceptualizations is related to Ontology Learning [2], which uses supervised and unsupervised methods. Supervised approaches use WordNet [11] (e.g. [32]), Wikipedia, DBpedia and Freebase for ontology induction or entity extraction [8,21,35]. DBpedia Spotlight [8] relies on explicit links between entities in Wikipedia. Entity extraction is performed by a pre-trained prefix tree, and the disambiguation is done with a language model (LM). Sense embeddings can also be trained for conceptualizations [15]. The proposed approach is related to [7] but does not use knowledge bases.

Unsupervised methods often rely on relation extraction (cf. Sect. 2.3). OntoUSP [27] extracts a probabilistic ontology for the medical domain using dependency path features. OntoGain [9] relies on multi-word expressions (MWEs) to construct a taxonomic graph using hierarchical clustering. This graph is expanded with non-taxonomic relations. The hybrid approach of [35], which accesses search engines, can also identify novel MWEs, which signify entities. Local taxonomic relations can be extracted from text by identifying certain syntactic patterns in a text [14,16]. There are also approaches that directly work on the SW graph and utilize distributional methods to establish concept similarities, e.g. [24].

2.2 Contextualization

The two most recent contextualization shared tasks are the Word Sense Disambiguation (WSD) tasks of SemEval 2010 [20] and SemEval 2013 [23]. The participating systems often use knowledge bases like YAGO [19], WordNet or other ontologies to assign sense identifiers to target words (usually nouns) in a sentence. Knowledge-free systems employ co-occurrence and distributional similarities together with language models.

2.3 Relation Extraction

TextRunner [36] extracts explicit relationship tuples (R, T_1, T_2) from POS-tagged text. These relations can be seen as 'facts' and aid question answering. GraBTax [34] can build taxonomies by utilizing co-occurrence and lexical similarity of n-gram topics from document titles.

OntoUSP [27] focuses on identifying specific relations in the medical domain. It can identify nominal MWEs and group different spellings into clusters. Also, it performs hierarchical clustering on verbs. [31] present a relation learning approach. By extracting known facts, they create triggers to extract similar relations of different entities. [28] uses distributional statistics to extract relations between nouns. This produces highly precise relations, but the recall is quite low. We believe that we can alleviate this problem by extracting relations not on the word level (leaf in the taxonomy graph) but between taxonomic concepts (nodes).

2.4 Linked Open Data

There are many useful data collections available on the Web. Freebase [6] or DBpedia [17] offer a large number of relations, usually as RDF triples, that can be queried using APIs. On top of that, there are applications like DBPedia Spotlight [8] or Babelfy [22] that can annotate texts with e.g. DBPedia entities, which in turn can link to Wikipedia.

We plan to extract and display information similar to the Weltmodell [1] or ConceptNet [33]. By utilizing disambiguated concepts, we believe that we can extract more (higher recall on concept level vs. word level) and more precise relations (handling polysemy).

3 Problem Statement and Contributions

The most trivial sentences and phrases can be difficult to understand and process for computers. Supervised ontologies and dictionaries help to a large extent, however, often they do not fit the textual domain to which they are applied. Search engines can be viewed as semantic applications, as they are able to identify word senses by using the provided keywords, e.g. *throw a ball* vs. *attend a ball*. However, it is not shown how many senses of the provided query terms the search engine knows about. We believe that this is important information and making this information available would help many users. Semantic resources like WordNet are often too fine-grained, which reduces usability in semantic applications[2].

To alleviate such problems, we plan to investigate the following research questions:

- How can we construct a semantic model that improves WSD performance? In this task we are going to use Distributional Semantics (DS) methods [13] to create conceptualizations from an input corpus. Afterwards, we are going to structure the concepts in a global taxonomy graph by utilizing the hypernymy structure.
- Which unsupervised, knowledge-free methods can we use to obtain state-of-the-art WSD? This task involves identifying concepts in their context to obtain an explicit semantic representation of the corpus. To the best of our knowledge, there have been no attempts to create a fully disambiguated corpus.

[2] WordNet identifies 12 senses for *ball*, some are highly domain-specific, like sense 12, “a pitch that is not in the strike zone.” Humans intuitively identify fewer senses.

– How can we identify significant relations between concepts to enrich our conceptualization? Using the contextualized corpus, we can extract relations on the concept and hypernym level, allowing us to extract more relations. We need to make sure that we only add significant relations to our concept graph.

Furthermore, we plan to create demonstrator applications based on the conceptualizations to exemplify the semantic annotation capability. We are also going to release the created applications as free, open-source applications with a focus on usability.

4 Research Methodology and Approach

4.1 Conceptualizations

To be able to annotate and semantify text, we need a knowledge model. Therefore, we are going to use the JoBimText framework [5] to create symbolic conceptualizations. We believe that having an explicit symbolic representation is an advantage to vector-based models like deep learning because of direct interpretability. We are going to create JoBimText models [30] and extend those to interconnected graphs, where we introduce new semantic relations between the nodes. A JoBimText model consists of a Distributional Thesaurus (DT) with sense-disambiguated entries. We induce word senses using Chinese Whispers [3], a knowledge-free graph clustering algorithm. The senses are labeled with the most frequent hypernym terms that were obtained using lexico-syntactic patterns [14,16], producing local taxonomies. In addition, the models contain significant context features for each word and a DT of such context features. The combination of entities like named entities and multi-word expressions [29] with common words will create an all-word knowledge base. Since it is fully based on the input corpus, there is no need for domain adaptation.

By utilizing the hypernymy structure, we can aggregate context features on concept levels. E.g. *jaguars*, *tigers* and *wolves* are all animals, but in our corpus, we only find sentences where *tigers* and *wolves* `hunt`. From this information, we can infer that *jaguars* probably can `hunt` as well, thus projecting contextual information through aggregation into the animal concept.

We apply contextualization to obtain a sense-disambiguated corpus. Then we compute the similarity graph once again, this time using word sense IDs instead of words. This should result in a DT, where each entry is fully disambiguated, allowing us to create a more detailed and more precise model (more entries per word sense, disambiguated context features).

4.2 Contextualization

With our sense-disambiguated semantic models, we can perform semantic text annotation. We put a strong focus on the contextualization technique, since it is going to connect the conceptualized knowledge to text. Using word sense disambiguation on the input text, we are going to infer the senses of the words.

By using similar context features from the model, we are able to identify the word sense, even if the term–context feature combination has never been observed in the corpus. Preliminary experiments have shown that utilizing similar features improves recall, with a slight decline in precision. To further increase recall, we will use co-occurrence features and a language model.

Once we have established a conceptualization with relations between concepts and aggregated context features per concept, we can even infer the concept for yet unseen words – a zero-shot contextualization, e.g. by matching the context features of concepts to the unseen word, we can assume that X in the phrase X *hunts its prey* is an animal.

4.3 Relation Extraction

The conceptualization yields a taxonomy graph with sense-labeled leaf nodes. We want to extend such a 'taxonomic skeleton' into a dense graph with many types of relations. Our semantic model already contains a large number of facts, like *jaguar* is-an animal or *cars* can be *driven*. While such facts seem trivial, in our model there are many facts that are not considered common knowledge, e.g. *impala* is-a car (indicating a Chevrolet Impala). We employ Open Information Extraction (OIE) techniques [26] to extract additional semantic relations.

Using the disambiguated corpus, we propagate the dependency relations in the hypernym graph to extend our conceptualization. If we take the input phrase *jaguar kills deer*, we can extract a multitude of facts:

basic *jaguar kills deer*
agent expansion *tiger kills deer*
object expansion *jaguar kills prey*
verb expansion *jaguar wounds deer*
hypernymies animal kills animal
combinations animal *kills deer, jaguar wounds* animal, etc.

Especially to identify relations between named entities [12], we need a larger, richer set of relations. Therefore, we are going to use supervised resources like WordNet to extract examples of a relation tuple (R, T_1, T_2) and – based on this input relation – find patterns that can be used to extract such relations. This bootstrapping method is similar to [31].

4.4 Linked Open Data

To make our approach usable to other researchers, as well as to incorporate our conceptualizations into the SW, we are going to publish the models as Linked Open Data. We are going build a semantic network similar to the hyper graph that is available in JoBimViz [30] and extend it with the concept taxonomies and contextualization techniques. This will allow to browse our conceptualizations with unique identifiers.

We are going to offer our contextualization technique through an open API. The annotated text would consist of a sequence of sense IDs, e.g. *jaguar#NN_2 hunt#VB_1 deer#NN_1* for the source sentence *jaguar hunts deer*. To bridge the gap between our inferred knowledge base to existing knowledge bases, we can create an alignment using Lesk [18]. This will increase the usability, since users can obtain identifiers of established SW resources and use this information to semantify their texts.

5 Preliminary Results

5.1 Conceptualization

Extending [5], we are now able to create semantic models that contain word senses and (unstructured) hypernyms. We call these models JoBimText models [30] and already use them for contextualization. The next step is to create a concept taxonomy with aggregated context features.

5.2 Contextualization

We have implemented a contextualization system that we are now extending with new features for a publication in the near future. Currently, it performs sense annotation based on a context feature extractor, e.g. trigram or dependency features. Using large language model with and word co-occurrences, we achieve a performance comparable to the systems in SemEval 2013, task 13 [23].

5.3 Relation Extraction

This task has not yet started, because it relies on a contextualized corpus.

5.4 Web Demonstrators/LOD

Our JoBimViz[3] web application is used to exemplify our semantic models [30]. It already features word identifiers, consisting of the model and a word representation (e.g. `lemma#POS`). It can be browsed as a semantic network, by following the links, similarly to LOD repositories. Furthermore, it features a transparent Java API for machine access. As a demonstrator for contextualized corpora, we have created a semantic search demo based on Apache Solr and PHP. It incorporates keyword search as well as search for concepts and displays possible MWE expansions.

6 Evaluation Plan

6.1 Conceptualizations

Using a path based measure [25], we can assess the structural similarity of our conceptualizations with WordNet. The sense clustering method is flexible and

[3] http://maggie.lt.informatik.tu-darmstadt.de/jobimviz/.

allows for different granularities. We want to identify the best granularity settings extrinsically, by evaluating the performance of several clusterings (with different granularities) using the contextualization technique. We use the Turk Bootstrap Word Sense Inventory (TWSI) 2.0 dataset [4]. It contains sense-annotated sentences from Wikipedia and a crowdsourced sense inventory with substitutions for about 1,000 nouns.

To verify our intuition that a model computed on a domain-specific corpus outperforms general or foreign-domain models, we plan to compute several models and cross-evaluate them.

6.2 Contextualization

To evaluate the performance of the contextualization system, we are going to use the TWSI dataset [4] here as well. It contains contextualized substitutions for about 150,000 sentences, a larger collection than used for SemEval WSD tasks. The TWSI dataset is mostly used for parameter tuning and determining the best feature configuration. Once the best feature set is established, we are going to evaluate our contextualization on the SemEval 2010 [20] and SemEval 2013 [23] datasets. This allows us to compare our unsupervised contextualization technique to state-of-the-art techniques, and possibly to participate in a future WSD challenge.

To evaluate the zero-shot contextualization, we can remove sentences with certain (even polysemous) target terms from the input corpus and create the conceptualization. Then we can input the sentences with the "unknown" words and evaluate the concept identification. To demonstrate improvements of the complex structured semantic model, we compare it with a simple distributional model.

6.3 Relation Extraction

The evaluation of relation extraction is challenging. To evaluate our approach, we are going to apply the relation extraction on a slightly different task. We are going to extract named entities like politicians (news data) and use our conceptualization to identify the events and relations in which they are involved. Most other open information systems rely on manually created datasets [10] to evaluate their systems.

7 Conclusion

In this proposal we have presented a framework for unsupervised conceptualizations based on unstructured text collections. Its advantage is that the resulting models are tied to the input text, thus allowing for applications without domain adaptation. The generated models are data-driven and can therefore be created for every domain where large amounts of texts are available. Using a contextualization technique, the framework creates a fully semantified sentence and document representation. This representation is tied to Linked Open Data resources.

Acknowledgments. This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the context of the Software Campus project LiCoRes under grant No. 01IS12054. The author would like to thank his mentor Simone Paolo Ponzetto, his advisers Chris Biemann and Martin Riedl, and the reviewers for their valuable feedback.

References

1. Akbik, A., Michael, T.: The weltmodell: a data-driven commonsense knowledge base. In: Proceedings of LREC 2014, Reykjavik, Iceland, pp. 3272–3276 (2014)
2. Biemann, C.: Ontology learning from text: a survey of methods. LDV forum **20**(2), 75–93 (2005)
3. Biemann, C.: Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In: Proceedings of TextGraphs-1, New York City, NY, USA, pp. 73–80 (2006)
4. Biemann, C.: Turk bootstrap word sense inventory 2.0: a large-scale resource for lexical substitution. In: Proceedings of LREC 2012, Istanbul, Turkey, pp. 4038–4042 (2012)
5. Biemann, C., Riedl, M.: Text: now in 2D! a framework for lexical expansion with contextual similarity. J. Lang. Model. **1**(1), 55–95 (2013)
6. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of ACM SIGMOD 2008, Vancouver, Canada, pp. 1247–1250 (2008)
7. Chen, X., Liu, Z., Sun, M.: A unified model for word sense representation and disambiguation. In: Proceedings of EMNLP 2014, Doha, Qatar, pp. 1025–1035 (2014)
8. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of I-SEMANTICS 2013, Graz, Austria, pp. 121–124. ACM (2013)
9. Drymonas, E., Zervanou, K., Petrakis, E.G.M.: Unsupervised ontology acquisition from plain texts: the *OntoGain* system. In: Hopfe, C.J., Rezgui, Y., Métais, E., Preece, A., Li, H. (eds.) NLDB 2010. LNCS, vol. 6177, pp. 277–287. Springer, Heidelberg (2010)
10. Etzioni, O., Fader, A., Christensen, J., Soderland, S., Mausam, M.: Open information extraction: the second generation. In: IJCAI, Barcelona, Spain, vol. 11, pp. 3–10 (2011)
11. Fellbaum, C.: Wordnet. An Electronic Lexical Database. MIT Press, Cambridge (1998)
12. Feuerbach, T., Riedl, M., Biemann, C.: Distributional semantics for resolving bridging mentions. In: Proceedings of RANLP 2015, Hissar, Bulgaria, pp. 192–199 (2015)
13. Harris, Z.S.: Methods in Structural Linguistics. University of Chicago Press, Chicago (1951)
14. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of COLING-1992, Nantes, France, pp. 539–545 (1992)
15. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Sensembed: learning sense embeddings for word and relational similarity. In: Proceedings of ACL 2015, Beijing, China, pp. 95–105 (2015)
16. Klaussner, C., Zhekova, D.: Lexico-syntactic patterns for automatic ontology building. SRW at RANLP **2011**, 109–114 (2011)

17. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semant. Web J. **6**(2), 167–195 (2015)
18. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of SIGDOC 1986, pp. 24–26. ACM, Toronto, Ontario, Canada (1986)
19. Mahdisoltani, F., Biega, J., Suchanek, F.: YAGO3: a knowledge base from multilingual wikipedias. In: Proceedings of CIDR 2015, Asilomar, CA, USA (2015)
20. Manandhar, S., Klapaftis, I.P., Dligach, D., Pradhan, S.S.: SemEval-2010 task 14: word sense induction & disambiguation. In: Proceedings of SemEval-2010, Uppsala, Sweden, pp. 63–68 (2010)
21. Medelyan, O., Manion, S., Broekstra, J., Divoli, A., Huang, A.-L., Witten, I.H.: Constructing a focused taxonomy from a document collection. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 367–381. Springer, Heidelberg (2013)
22. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. TACL **2**, 231–244 (2014)
23. Navigli, R., Vannella, D.: SemEval-2013 task 11: word sense induction and disambiguation within an end-user application. In: Proceedings of *SEM 2013, Atlanta, GA, USA, vol. 2, pp. 193–201 (2013)
24. Nováček, V., Handschuh, S., Decker, S.: Getting the meaning right: a complementary distributional layer for the web semantics. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 504–519. Springer, Heidelberg (2011)
25. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity: measuring the relatedness of concepts. In: Demonstration Papers at HLT-NAACL 2004, Boston, MA, USA, pp. 38–41 (2004)
26. Piskorski, J., Yangarber, R.: Information extraction: past, present and future. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing, pp. 23–49. Springer, Heidelberg (2013)
27. Poon, H., Domingos, P.: Unsupervised ontology induction from text. In: Proceedings of ACL 2010, Uppsala, Sweden, pp. 296–305 (2010)
28. Remus, S.: Unsupervised relation extraction of in-domain data from focused crawls. In: SRW at EACL 2014, Gothenburg, Sweden, pp. 11–20 (2014)
29. Riedl, M., Biemann, C.: A single word is not enough: ranking multiword expressions using distributional semantics. In: Proceedings of EMNLP 2015, Lisboa, Portugal, pp. 2430–4440 (2015)
30. Ruppert, E., Kaufmann, M., Riedl, M., Biemann, C.: JoBimViz: a web-based visualization for graph-based distributional semantic models. In: System Demonstrations at ACL 2015, Beijing, China, pp. 103–108 (2015)
31. Shinyama, Y., Sekine, S.: Preemptive information extraction using unrestricted relation discovery. In: Proceedings of HLT-NAACL 2006, New York, NY, USA, pp. 304–311 (2006)
32. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogenous evidence. In: Proceedings of COLING/ACL 2006, Sydney, Australia, pp. 801–808 (2006)

33. Speer, R., Havasi, C.: Conceptnet 5: a large semantic network for relational knowledge. In: Gurevych, I., Kim, J. (eds.) The Peoples Web Meets NLP. Theory and Applications of Natural Language Processing, pp. 161–176. Springer, Heidelberg (2013)
34. Treeratpituk, P., Khabsa, M., Giles, C.L.: Graph-based approach to automatic taxonomy generation (grabtax). CoRR abs/1307.1718 (2013)
35. Wong, W., Liu, W., Bennamoun, M.: Acquiring semantic relations using the web for constructing lightweight ontologies. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 266–277. Springer, Heidelberg (2009)
36. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: Textrunner: open information extraction on the web. In: System Demonstrations at NAACL 2007, Rochester, NY, USA, pp. 25–26 (2007)