HUMBOLDT-UNIVERSITÄT ZU BERLIN
MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT
INSTITUT FÜR INFORMATIK

Exposé

Typ der Arbeit:	Bachelorarbeit Informatik (4 Monate)
Arbeitstitel:	Investigating weak supervision for the extraction of mobility relations and events in German text
Kandidat:	Truong, Phuc Tran
Matrikelnummer:	558919
Gutachter:	Prof. Dr. Ulf Leser
Betreuer:	Dr. Leonhard Hennig
Geplanter Zeitraum:	Februar 2020 bis Juni 2020

1 Scientific Background

This work aims to investigate weak supervision methods to accelerate labeled training data generation for training of a supervised machine learning model for mobility related event and relation extraction from German text.

Weakly Supervised Machine Learning. The goal of supervised learning is to learn a mathematical model given a training set that contains pairs of inputs and the desired label. The model can be evaluated on labeled test examples by comparing its predictions with the true labels (Chapelle et al., 2010). Deep learning methods not only learn to predict but also to represent data with non-linear combinations of arbitrary features, such that it is suitable for classification (LeCun et al., 2015). This in turn amplifies the need for training data because typical deep learning systems for text or images contain millions of parameters. Each of these parameters has to be adjusted by looking at training examples to minimise the model's prediction error using an optimisation procedure such as Stochastic Gradient Descent.

The traditional approach to get more training data is to do manual annotation, which is expensive and time consuming. Building a text-based dataset requires instructing human annotators to follow specific annotation guidelines and revising annotations as they can be prone to errors (Pustejovsky and Stubbs, 2012). In practice, researchers may spend months to create these hand-labeled training sets.[1]

Weak supervision aims to get lower-quality, but larger training datasets faster, for instance through heuristic patterns (Hearst, 1992), distant supervision with a knowledge base (Mintz et al., 2009) or crowd-sourcing (Zhang et al., 2017). Recent work by Ratner et al. (2016) and Bach et al. (2017) proposes to unify multiple, potentially overlapping weak supervision sources in a single weak supervision model. By modelling correlations and other dependencies between the weak supervision sources and learning their estimated accuracies they aim to reduce their noise. A recent example of such strategies is the Snorkel framework (Ratner et al., 2017), where users express weak supervision sources as labeling functions. For each example these labeling functions either cast a vote (e.g. this example belongs to class C) or abstain. The learned label model combines the votes of all labeling functions weighted by their estimated accuracies and outputs a set of probabilistic labels that can be used to train machine learning models. The end goal is for these models to generalise beyond the information expressed in the individual labeling functions.

[1] `https://www.snorkel.org/blog/weak-supervision`

Relation Extraction and Event Extraction. Relation Extraction (RE) is the task of extracting relations between two or more named entities of a certain type in text. Relations correspond to semantic categories such as *married-to*, *lives-in*. RE is a key part of natural language processing tasks such as knowledge base population (Ji and Grishman, 2011) or question answering (Xu et al., 2016).

It is a component of event extraction, which according to the ACE Event Detection and Characterization task[2] (Doddington et al., 2004) involves

1. identifying and characterising the type of event (e.g. *Obstruction*, *TrafficJam*) through its trigger (i.e. the word that most clearly expresses the event's occurrence),

2. as well as identifying the event trigger's arguments (entity mentions, temporal expressions or values) and labeling their roles (e.g. location and duration of a *TrafficJam* event), i.e. their relation to the event.

Pattern-based approaches construct event specific patterns and perform pattern matching to extract an event with its arguments (Xiang and Wang, 2019). While patterns can be manually constructed, recent approaches obtain patterns automatically via bootstrapping using only a few annotated examples or seed patterns (Xu et al. (2007), Grusdt et al. (2018)). For machine learning methods there are two main approaches: (1) A pipelined approach treats each subtask, namely trigger prediction and argument extraction, as its own classification problem with its own model. (2) A joint approach extracts event triggers and arguments simultaneously (Xiang and Wang, 2019). Liu et al. (2018) use the latter approach. They take word embeddings as input, which are low-dimensional and real-valued vector representations (Goldberg, 2017), and enrich these word representations with local sequential context information using a BiLSTM layer and syntactic information using a graph convolution network layer. They use these syntactic contextual representations in a joint extraction layer for trigger classification and argument role labeling.

The supervised machine learning approaches are reliant on large annotated datasets. However, well-known datasets for relation extraction such as SemEval-2010 Task 8 dataset (Hendrickx et al., 2009), TACRED (Zhang et al., 2017) or the ACE dataset for the Event Detection and Characterization task (Doddington et al., 2004) are less useful for domain- and language-specific tasks like event extraction on German mobility data. It requires different mobility-related event types and other types of relations between

[2] `https://www.ldc.upenn.edu/collaborations/past-projects/ace`

events and its arguments. For example, from the sentence *"S5 Mahlsdorf <> Friedrichsfelde Ost, noch bis 27.01. (Mo), ca. 1.30 Uhr, Ersatzverkehr"* we would like to extract a *RailReplacementService* event (*trigger="Ersatzverkehr*) with the arguments *location="S5"*, *start-location="Mahlsdorf"*, *end-location="Friedrichsfelde Ost"* and *end-date="27.01. (Mo), ca. 1.30 Uhr"*.

Weak supervision methods such as distant supervision (Mintz et al., 2009) have been used to combat the lack of training data for RE by leveraging relational information from knowledge bases to automatically identify extraction examples. Zeng et al. (2018) use distant supervision to extract event extraction training instances and develop a neural model to automatically label them by leveraging existing structured knowledge bases such as Freebase[3]. However, there is no standard data base available for German mobility events such as *TrafficJam* (Schiersch et al., 2018).

2 Goals of the Thesis

The main goal of this work is to adapt weak supervision tools provided by the Snorkel framework to the task of extracting mobility related events and relations from German text. We aim to assess the quality and the suitability of the generated weakly labeled data for training of a supervised machine learning model.

The dataset of interest in this thesis is provided by the German Research Center for Artificial Intelligence and consists of more than 3500 German-language documents from mobility related tweets and RSS feeds, which have been annotated with fine-grained geo-entities and standard named entities. Of those documents more than 670 originated from the "SmartData" corpus[4] and have annotations for the following traffic related relations and events (see Schiersch et al. (2018) for detailed descriptions): *Accident*, *CanceledRoute*, *CanceledStop*, *Delay*, *Obstruction*, *RailReplacementService* and *TrafficJam*. Each of these events uses named entities as arguments. All of the events are identified by a trigger, require a *location* argument and have optional arguments. For example, the *Accident* event has the following optional arguments: *delay*, *direction*, *start-location*, *end-location*, *start-date*, *end-date* and *cause*. We will refer to these documents with relation and event annotation as the "SmartData" dataset and refer to the remaining documents as the "Daystream" dataset.

As we are interested in training a supervised machine learning model to detect the events mentioned above and extract their arguments, we want

[3] `https://developers.google.com/freebase/`

[4] `https://github.com/DFKI-NLP/smartdata-corpus`

to increase our training data by designing Snorkel labeling functions to probabilistically label the event types and the relations between the events and their arguments in the "Daystream" documents based on the existing NER annotation. In the process, we want to explore the following questions:

1. Can we apply Snorkel to the extraction of mobility related relations and events from German text?
2. How does a model trained with a small hand-labeled training set (from the SmartData dataset) compare to a model trained with a bigger weakly labeled training set (Snorkel labeled Daystream dataset)?
3. Does augmenting with weakly labeled data improve model performance?
4. What limitations does this approach have?

Throughout our experiments we plan to use an existing BiLSTM-based model from the DFKI loosely based on Liu et al. (2018), that jointly detects events and extracts their arguments. As a baseline we will train and evaluate this model on an existing split of the SmartData documents. We measure its performance on trigger (event type) classification and event argument role classification using precision, recall and F-measure. We then train the model on the newly created "Daystream" training set to see how well it performs on the "SmartData" test set, when it is only trained on Snorkel labeled data. Finally, we will train the model on a combination of the "SmartData" training set and the Snorkel labeled "Daystream" training set to see how data augmentation affects performance. These different configurations will be used to investigate whether the Snorkel labeling quality is sufficient or whether a manual correction is necessary.

3 Plan of Implementation

This work is mainly based on the Python library Snorkel[5]. As the "Daystream" data already includes NER annotation including a generic trigger type, we will limit ourselves to designing labeling functions for the trigger classification and argument role classification. We write these labeling functions based on heuristic patterns and refine the labeling functions by measuring their performance and coverage on the "SmartData" training set.

[5] https://www.snorkel.org

A starting point for a trigger classification labeling function is to use a keyword-based heuristic. If the trigger matches some keyword associated with *TrafficJam* such as "Stau" or "stockender Verkehr", the labeling function outputs *TrafficJam* and abstains otherwise. However, "Stau" can be used in other contexts to denote slow or halting progress (Hennig et al., 2016), which suggests the use of the surrounding words in the sentence and the existing NER annotation in addition to the trigger tokens. For the argument role classification we could use pattern-based approaches. For the design of our labeling functions we could revisit patterns of prior work at the German Research Center for Artifical Intelligence. For instance, Hennig et al. (2016) extracted dependency patterns from a set of event-specific training sentences, as proposed by Xu et al. (2007). They reported that those patterns worked best for RSS feeds, since traffic information RSS feeds tend to be well structured and use very formalised language. Grusdt et al. (2018) used surface patterns, that were automatically extracted, ranked using a bootstrapping approach and then manually refined.

References

Bach, S. H., He, B., Ratner, A., and Ré, C. (2017). Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 273–282. JMLR.org.

Chapelle, O., Schlkopf, B., and Zien, A. (2010). *Semi-Supervised Learning*. The MIT Press, 1st edition.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Grusdt, B., Nehring, J., and Thomas, P. (2018). Bootstrapping patterns for the detection of mobility related events. In *14th Conference on Natural Language Processing. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS-2018), September 19-21, Vienna, Austria*. Verlag der Österreichischen Akademie der Wissenschaften.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2009). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, SEW '09, pages 94–99, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hennig, L., Thomas, P., Ai, R., Kirschnick, J., Wang, H., Pannier, J., Zimmermann, N., Schmeier, S., Xu, F., Ostwald, J., and Uszkoreit, H. (2016). Real-time discovery and geospatial visualization of mobility and industry events from large-scale, heterogeneous data streams. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Annual Meeting of the Association for Computational Linguistics (ACL-16), 54th, August 7-12, Berlin, Germany*. ACL.

Ji, H. and Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1148–1158, Stroudsburg, PA, USA. Association for Computational Linguistics.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Liu, X., Luo, Z., and Huang, H. (2018). Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Brussels, Belgium. Association for Computational Linguistics.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* " O'Reilly Media, Inc.".

Ratner, A. J., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. (2016). Data programming: Creating large training sets, quickly. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 3567–3575. Curran Associates, Inc.

Schiersch, M., Mironova, V., Schmitt, M., Thomas, P., Gabryszak, A., and Hennig, L. (2018). A german corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. In *Proceedings of the 11th International Conference on Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-18), 11th, May 7-12, Miyazaki, Japan*. European Language Resources Association.

Xiang, W. and Wang, B. (2019). A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Xu, F., Uszkoreit, H., and Li, H. (2007). A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *Proceedings of the 45th annual meeting of the Association of Computational Linguistics*, pages 584–591.

Xu, K., Reddy, S., Feng, Y., Huang, S., and Zhao, D. (2016). Question answering on Freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2326–2336, Berlin, Germany. Association for Computational Linguistics.

Zeng, Y., Feng, Y., Ma, R., Wang, Z., Yan, R., Shi, C., and Zhao, D. (2018). Scale up event extraction learning via automatic training data generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, Y., Zhong, V., Chen, D., Angeli, G., and Manning, C. D. (2017). Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.