Final Report for CENIIT Project 15.02

Semantic Parsing for Text Analytics

Marco Kuhlmann

1 Most important scientific results

This project has developed new techniques for *text analytics*, the process of transforming textual data into structured information. One of the key component technologies in text analytics is *semantic parsing*, the automatic mapping of a sentence into a formal representation of its meaning. One such representation are so-called dependency graphs, which spell out the predicate–argument structure of the analysed sentence, the 'who did what to whom'. This information has been shown to be useful for many downstream applications, including information extraction, question answering, and machine translation.

In the project, we have presented some of the first dynamic programming algorithms for semantic parsing to dependency graphs (Kuhlmann and Jonsson, 2015; Kurtz and Kuhlmann, 2017), and contributed to a better understanding of the neural network architectures that define the state of the art for this task (Kurtz and Kuhlmann, 2019; Kurtz et al., 2019; Kunz and Kuhlmann, 2020). Additionally, we have compiled benchmark data sets for the development and evaluation of semantic dependency parsers (Flickinger et al., 2016), and coordinated community-building efforts targeted at the comparison of different meaning representations and semantic parsers (Oepen et al., 2015; Kuhlmann and Oepen, 2016; Oepen et al., 2019).

2 Degrees and promotions the project has contributed to

- PhD student Robin Kurtz, who was funded by the Swedish Graduate School for Computer Science (CUGS), supervised by Marco Kuhlmann, and associated with the project since 2015, successfully defended his doctoral thesis *Contributions to Semantic Dependency Parsing: Search, Learning, and Application* on 2020-09-25.

- Marco Kuhlmann was promoted to the rank of Associate Professor (*Biträdande professor*) in November 2016. In February 2021, the Academic Appointments Board of the Institute of Technology recommended him for promotion to the rank of Professor; the final decision by the Vice-Chancellor is pending.

3 Master's theses connected to the project

During the course of the project, we supervised and examined a total number of 30 Master's thesis on topics related to the project. For the complete list, see Appendix B.

4 Persons funded by the project

- Marco Kuhlmann, Senior Lecturer (2013–2016), then Associate Professor (since 2016). The project funded up to 30% of his salary in the period 2015–2019.
- Jenny Kunz, PhD student since 2019. Principal supervisor: Marco Kuhlmann. The project funded the major part of her salary in the period 2019–2020.

5 Industrial connections

- Collaboration on Master's theses on project-related topics with many different companies at the regional and national level, including Ericsson, Etteplan, Fodina Language Technology, Gavagai, IamIP, iMatrics, Sectra, Storytel.
- Collaboration with Google Inc. in the research project *Adaptive Algorithms for Semantic Dependency Parsing*, funded by a Google Faculty Research Award, 2016. Funding amount 0.5 MSEK.
- Collaboration with IamIP, Stockholm, and RISE on a research proposal *AI-assisted IP Coordinator*, submitted to Vinnova's call *From AI Research to Innovation*, 2019. Went to the final round of interviews; not funded.
- Collaboration with iMatrics, Linköping, on a research proposal *Relation Extraction with Deep Neural Language Models*, submitted to ELLIIT's call for pre-projects, 2020. Funded; funding amount 2 MSEK.
- Marco Kuhlmann has acted (2015–2020) as a technical advisor and mentor (*pro bono*) for Worldish AB, a start-up company that won the 2017 Swedbank Rivstart competition for entrepeneurs from all over Sweden.

6 Connections with other CENIIT projects

Joint seminars and supervision of PhD students Jenny Kunz and Riley Capshaw with CENIIT Project 12.10 'Semantic Technologies for Decision Support' (PI: Eva Blomqvist) and its follow-up projects.

7 New research group

This project and the funding from CENIIT have been essential for the creation of Marco Kuhlmann's research group *Natural Language Processing and Text Mining* at the Department of Computer and Information Science. As of May 2021, this group consists of 3 PhD students, 1 postdoc, and 1 Associate Professor.

Appendix A: List of publications

1. Jenny Kunz and Marco Kuhlmann. Classifier Probes May Just Learn from Linear Context Features. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 5136–5146, Barcelona, Spain, 2020.

2. Robin Kurtz, Stephan Oepen, and Marco Kuhlmann. End-to-End Negation Resolution as Graph Parsing. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 14–24, Online, 2020.

3. Riley Capshaw, Marco Kuhlmann, and Eva Blomqvist. Probing a Semantic Dependency Parser for Translational Relation Embeddings. In *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2020) Co-located with the 17th Extended Semantic Web Conference 2020 (ESWC 2020)*, Heraklion, Greece – moved online, 2020.

4. Fredrik Sand Aronsson, Marco Kuhlmann, Vesna Jelić, and Per Östberg. Is Cognitive Impairment Associated with Reduced Syntactic Complexity in Writing? Evidence from Automated Text Analysis. *Aphasiology*, 2020.

5. Stephan Oepen, Omri Abend, Jan Hajič, Daniel Hershcovich, Marco Kuhlmann, Tim O'Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, and Zdeňka Urešová. MRP 2019: Cross-Framework Meaning Representation Parsing. In *Proceedings of the CoNLL 2019 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 1–27, Hong Kong, China, 2019.

6. Marco Kuhlmann, Andreas Maletti, and Lena Katharina Schiffer. The Tree-Generative Capacity of Combinatory Categorial Grammars. In *Proceedings of the IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 44:1–44:14, Mumbai, India, 2019.

7. Robin Kurtz and Marco Kuhlmann. The Interplay Between Loss Functions and Structural Constraints in Dependency Parsing. *Northern European Journal of Language Technology*, 6:43–66, 2019.

8. Robin Kurtz, Daniel Roxbo, and Marco Kuhlmann. Improving Semantic Dependency Parsing with Syntactic Features. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 12–21, Turku, Finland, 2019.

9. Marco Kuhlmann, Giorgio Satta, and Peter Jonsson. On the Complexity of CCG Parsing. *Computational Linguistics*, 44(3):447–482, 2018.

10. Robin Kurtz and Marco Kuhlmann. Exploiting Structure in Parsing to 1-Endpoint-Crossing Graphs. In *Proceedings of the 15th International Conference on Parsing Technologies (IWPT)*, pages 78–87, Pisa, Italy, 2017.

11. Per Fallgren, Jesper Segeblad, and Marco Kuhlmann. Towards a Standard Dataset of Swedish Word Vectors. In *Proceedings of the Sixth Swedish Language Technology Conference (SLTC)*, Umeå, Sweden, 2016.

12. Marco Kuhlmann and Stephan Oepen. Towards a Catalogue of Linguistic Graph Banks. *Computational Linguistics*, 42(4):819–827, 2016.

13. Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Zdeňka Urešová. Towards Comparability of Linguistic Graph Banks for Semantic Parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 3991–3995, Portorož, Slovenia, 2016.

14. Dan Flickinger, Jan Hajič, Angelina Ivanova, Marco Kuhlmann, Yusuke Miyao, Stephan Oepen, and Daniel Zeman. SDP 2014 & 2015: Broad Coverage Semantic Dependency Parsing LDC2016T10. Linguistic Data Consortium, 2016.

15. Marco Kuhlmann and Peter Jonsson. Parsing to Noncrossing Dependency Graphs. *Transactions of the Association for Computational Linguistics*, 3:559–570, 2015.

16. Frank Drewes, Kevin Knight, and Marco Kuhlmann. Formal Models of Graph Transformation in Natural Language Processing (Dagstuhl Seminar 15122). *Dagstuhl Reports*, 5(3):143–161, 2015.

17. Marco Kuhlmann, Alexander Koller, and Giorgio Satta. Lexicalization and Generative Power in CCG. *Computational Linguistics*, 41(2):187–219, 2015.

18. Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. SemEval 2015 Task 18: Broad-Coverage Semantic Dependency Parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, CO, USA, 2015.

Appendix B: List of Master's theses

1. Marc Pàmies Massip. Multilingual Identification of Offensive Content in Social Media. MSc in Computer Science, 2020. External project at Helsinki University.

2. Alexander Häger. Contextualizing Music Recommendations. MSc in Computer Science and Engineering, 2020. External project at Spotify, Boston, USA.

3. Min-Chun Shih. Exploring Cross-lingual Sublanguage Classification with Multilingual Word Embeddings. MSc in Statistics and Machine Learning, 2020.

4. Robin Ellgren. Exploring Emerging Entities and Named Entity Disambiguation in News Articles. MSc in Information Technology, 2020. External project at iMatrics, Linköping.

5. Ludvig Westerdahl. Predicting the Financial Impact of the CEO's Comments in Quarterly Reports. MSc in Computer Science, 2020. External project at Redeye, Stockholm.

6. Jesper Hedlund and Emma Nilsson Tengstrand. A Comparison between Different Recommender System Approaches for a Book and an Author Recommender System. MSc in Computer Science and Engineering, 2020. External project at Storytel Sweden AB, Stockholm.

7. Pontus Svensson. Automated Image Suggestions for News Articles: An Evaluation of Text and Image Representations in an Image Retrieval System. MSc in Computer Science and Engineering, 2020. External project at Consid, Linköping.

8. Rebecca Lindblom. News Value Prediction with Textual Features and Machine Learning. MSc in Computer Science and Engineering, 2020. External project at iMatrics, Linköping.

9. Ludvig Noring. Predicting Swedish News Article Popularity. MSc in Computer Science and Engineering, 2020. External project at Schibsted Sverige AB, Stockholm.

10. Harald Pettersson. Sentiment Analysis and Transfer Learning Using Recurrent Neural Networks: An Investigation of the Power of Transfer Learning. MSc in Computer Science and Engineering, 2019. External project at Findwise AB, Stockholm.

11. Milda Pocevičiūtė. Machine Learning Framework for Automated Case Assignment of Radiology Report Requests. MSc in Statistics and Machine Learning, 2019. External project at Sectra AB, Linköping.

12. Anna-Katharina Fürgut. Mining Symptom Phrases within Free-Text Answers to Anamnesis Questionnaires. MSc in Statistics and Machine Learning, 2019. External project at Doctrin AB, Stockholm.

13. Harald Grant. Extractive Multi-Document Summarization of News Articles. MSc in Computer Science, 2019. External project at Schibsted Sverige AB, Stockholm.

14. Max Lund. Duplicate Detection and Text Classification on Simplified Technical English. MSc in Computer Science, 2019. External project at Etteplan, Linköping.

15. Johannes Palm Myllylä. Domain Adaptation for Hypernym Discovery via Automatic Collection of Domain-Specific Training Data. MSc in Computer Science and Engineering, 2019. External project at Fodina Language Technology AB, Linköping.

16. Gustav Gränsbo. Word Clustering in an Interactive Text Analysis Tool. MSc in Computer Science and Engineering, 2019. External project at Gavagai AB, Stockholm.

17. Daniel Roxbo. A Detailed Analysis of Semantic Dependency Parsing with Deep Neural Networks. MSc in Computer Science, 2019.

18. Sanne Ingvarsson. Using Machine Learning to Learn from Bug Reports: Towards Improved Testing Efficiency. MSc in Electrical Engineering, 2019. External project at Sectra AB, Linköping.

19. Sijin Cheng. Relevance Feedback-based Optimization of Search Queries for Patents. MSc in Computer Science, 2019. External project at IamIP Sverige AB, Sundbyberg.

20. Alice Reinaudo. Hierarchical Text Classification of Fiction Books. MSc in Computer Science, 2019. External project at Storytel Sweden AB, Stockholm.

21. Fredrik Öhrström. Cluster Analysis with Meaning: Detecting Texts that Convey the Same Message. MSc in Computer Science, 2019. External project at Etteplan, Linköping.

22. Jesper Bäck. Domain Similarity Metrics for Predicting Transfer Learning Performance. MSc in Computer Science, 2018. External project at Consid, Linköping.

23. Lina Gunnarsson. Semiautomatic De-Identification of Patient Data. MSc in Biomedical Engineering, 2018. External project at Sectra AB, Linköping.

24. Simon Lindblad. Labeling Clinical Reports with Active Learning and Topic Modeling. MSc in Computer Science, 2018. External project at Sectra AB, Linköping.

25. Justus Johansson Lindkvist. Automatic De-Identification of Personally Identifiable Information. MSc in Electrical Engineering, 2018. External project at Sectra AB, Linköping.

26. Riley Capshaw. Relation Classification using Semantically-Enhanced Syntactic Dependency Paths: Combining Semantic and Syntactic Dependencies for Relation Classification using Long Short-Term Memory Networks. MSc in Computer Science, 2018.

27. Francesco Cucari. Development of an Artificial Intelligence System for Localizing Bugs in Large Industrial Software Projects. MSc in Artificial Intelligence and Robotics, 2017. External project at Ericsson AB, Linköping.

28. Nils Axelsson. Dynamic Programming Algorithms for Semantic Dependency Parsing. MSc in Computer Science and Engineering, 2017.

29. Jesper Segeblad. Putting a Spin on SPINN: Representations of Syntactic Structure in Neural Network Sentence Encoders for Natural Language Inference. MSc in Cognitive Science, 2017.

30. Zonghan Wu. Neural Networks for Dependency Parsing. MSc in Statistics and Data Mining, 2016.