FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Zuzana Šimečková

Entity Relationship Extraction

Institute of Formal and Applied Linguistics

Supervisor of the master thesis:	RNDr. Milan Straka, Ph.D.
Study programme:	Computer Science
Study branch:	IUI

Prague 2020

This is not a part of the electronic version of the thesis, do not scan!

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....
Author's signature

Dedication.

Title: Entity Relationship Extraction

Author: Zuzana Šimečková

Institute: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Milan Straka, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Abstract.

Keywords: key words

Contents

Introduction

There has been made noticeable progress in natural language processing since the first deep neural networks attempts. With multiple new approaches and inventions such as multitask learning, word embeddings, RNN, attention and the transformer architecture. Last year Devlin et al. [2018] created BERT and managed to achieve state-of-the-art performance in eleven natural language processing tasks, including GLUE (7.7% point absolute improvement), MultiNLI accuracy (4.6% absolute improvement) and SQuAD problems.

previous work: Existing work on relation extraction (e.g., Zelenko et al., 2003; Mintz et al., 2009; Adel et al., 2016)

not a sentence

In this thesis, we will try to use those novel approaches to predict relation between two entities based on a Czech sentence. First part of this thesis will be focused on data. We will introduce some existing English datasets for Entity Relation Extraction. Than we will describe how we prepared data for Czech version of this task using distant supervision on Czech Wikipedia and Wikidata. Second part

o čem bude druhá část

1. Datasets

tady představíme existující dataesty

1.1 SEMEVAL 2010 task 8 dataset

The SemEval-2010 Task 8 dataset (S10T8) was introduced in SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals Hendrickx et al. [2010]. We will summarize how S10T8 was created and some other information from that article so that later we can compare different approaches.

nějak napsat, že nebudu citovat, ale je to hodně vykradené?

The authors started by choosing an inventory of semantic relations. They aimed for such a set of relations that it would be exhaustive (enable the description of relations between any pair of nominals) and mutually exclusive (given context and a pair of nominals only one relation should be selectable). Chosen relations with descriptions and examples are listed in table 1.1.

proč není table součást odkazu?

They decided to accept as relation arguments any noun phrases with common-noun heads not just named entities or some other specific class of noun phrases, mentioning 'Named entities are a specific category of nominal expressions best dealt with using techniques which do not apply to common nouns.' But they restricted noun phrases to single words with the exception to lexicalized terms (such as science fiction).

nechat ut tu citaci?

quote better

formát

The annotation process had three rounds. In the first round, authors manually collected around 1,200 sentences for each relation through pattern-based Web search (with at least a hundred patterns per relation). This way, they obtained around 1200 sentences for each relation. In the second round, each sentence was annotated by two independent annotators. In the third round disagreements were resolved and the dataset was finished. Every sentence was classified either as a true relation mention or was a near-miss and thus classified as "other", or was removed.

lepší uvozovky

The dataset contains of 10717 relation mentions. For the original competition, teams were given three training dataset of sizes 1000 (TD1), 2000 (TD2), 4000 (TD3), and 8000 (TD4). There was a notable gain TD3 →TD4 therefore the authors concluded that even larger dataset might be helpful to increase performance of models. But

> .. that is so much easier said than done: it took the organizers well in excess of 1000 person-hours to pin down the problem, hone the guidelines and relation definitions, construct sufficient amounts of trustworthy training data, and run the task.

1.2 TACRED dataset

The TAC Relation Extraction Dataset was introduced in Zhang et al. [2017]. TACRED is a supervised dataset obtained via crowdsourcing. It contains about 100 000 examples. Each example contains is in Authors claim so far used training data had often been too noisy for reliable training of relation extraction systems

Label	Freq
Cause-Effect An event or object leads to an effect. *The burst has been caused by water hammer pressure.*	12.4% (1331)
Instrument-Agency An agent uses an instrument. *The author of a keygen uses a disassembler to look at the raw assembly code.*	6.2% (660)
Product-Producer A producer causes a product to exist. *The factory's products have included flower pots, Finnish rooster-whistles, pans, trays, tea pots, ash trays and air moisturisers.*	8.8% (948)
Content-Container An object is physically stored in a delineated area of space. *This cut blue and white striped cotton dress with red bands on the bodice was in a trunk of vintage Barbie clothing.*	6.8% (732)
Entity-Origin An entity is coming or is derived from an origin (e.g., position or material). *The avalanches originated in an extensive mass of rock that had previously been hydrothermally altered in large part to clay.*	9.1% (974)
Entity-Destination An entity is moving towards a destination. *This book has transported readers into ancient times.*	10.6% (1137)
Component-Whole An object is a component of a larger whole. *The system as described above has its greatest application in an arrayed configuration of antenna elements.*	11.7% (1253)
Member-Collection A member forms a nonfunctional part of a collection *The student association is the voice of the undergraduate student population of the State University of New York at Buffalo.*	8.6% (923)
Message-Topic A message, written or spoken, is about a topic. *Cieply's story makes a compelling point about modern-day studio economics.*	8.4% (895)
Other *The child was carefully wrapped and bound into the cradle by means of a cord.*	17.4% (1864)

Table 1.1: S10T8 summary. List of relations, their official descriptions, a random example and both relative and absolute count.

... machine learning approaches have suffered from two key problems: (1) the models used have been insufficiently tailored to relation extraction, and (2) there has been insufficient annotated data available to satisfy the training of data-hungry models, such as deep learning models.

2. Title of the second chapter

2.1 Title of the first subchapter of the second chapter

2.2 Title of the second subchapter of the second chapter

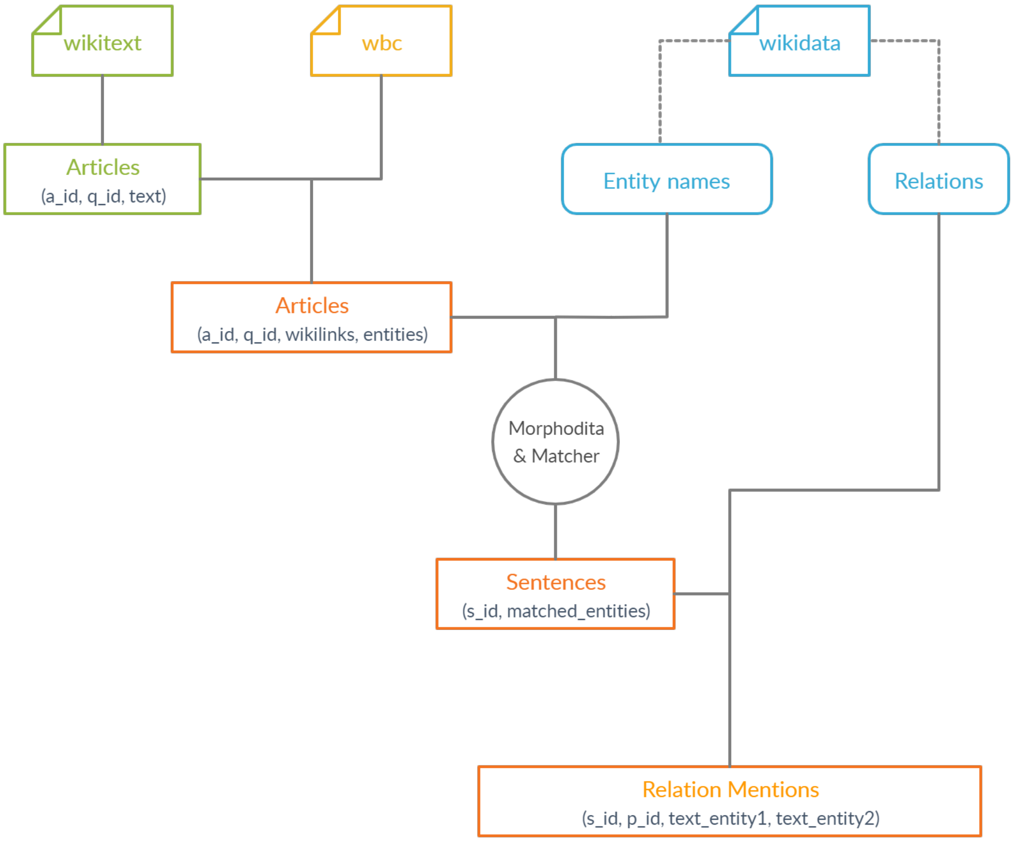

Figure 2.1: Zjednodušený diagram výroby korpusu

Conclusion

Bibliography

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, page 33–38, USA, 2010. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017. URL https://nlp.stanford.edu/pubs/zhang2017tacred.pdf.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment