



**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	2
1 Datasets	3
1.1 SEMEVAL 2010 task 8 dataset	3
1.2 TACRED dataset	3
2 Title of the second chapter	5
2.1 Title of the first subchapter of the second chapter	5
2.2 Title of the second subchapter of the second chapter	5
Conclusion	7
Bibliography	8
List of Figures	9
List of Tables	10
List of Abbreviations	11
A Attachments	12
A.1 First Attachment	12

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.

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

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

not a sentence

o čem bude druhá část

1. Datasets

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.

First the authors decided on 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).

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).

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

... 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.

tady
představíme
exis-
tující
dataesty

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

připravil
tabulku
nebo
něco,
kam se
odkázat
na
jejich
seznam

nechat
ut tu
citaci?

quote
better

formát

Label	Freq	num2
Cause-Effect An event or object leads to an effect. <i>The <u>burst</u> has been caused by water hammer <u>pressure</u>.</i>	12.4% (1331)	•
Instrument-Agency An agent uses an instrument. <i>The <u>author</u> of a <u>keygen</u> uses a <u>disassembler</u> to look at the raw assembly code.</i>	6.2% (660)	•
Product-Producer A producer causes a product to exist. <i>The <u>factory</u>'s products have included flower pots, Finnish rooster-whistles, pans, <u>trays</u>, tea pots, ash trays and air moisturisers.</i>	8.8% (948)	•
Content-Container An object is physically stored in a delineated area of space. <i>This cut blue and white striped cotton <u>dress</u> with red bands on the bodice was in a <u>trunk</u> of vintage Barbie clothing.</i>	6.8% (732)	•
Entity-Origin An entity is coming or is derived from an origin (e.g., position or material). <i>The <u>avalanches</u> originated in an extensive <u>mass</u> of rock that had previously been hydrothermally altered in large part to clay.</i>	9.1% (974)	•
Entity-Destination An entity is moving towards a destination. <i>This book has transported <u>readers</u> into <u>ancient times</u>.</i>	10.6% (1137)	•
Component-Whole An object is a component of a larger whole. <i>The system as described above has its greatest application in an arrayed configuration of antenna <u>elements</u>.</i>	11.7% (1253)	•
Member-Collection A member forms a nonfunctional part of a collection <i>The <u>student association</u> is the voice of the undergraduate student population of the State University of New York at Buffalo.</i>	8.6% (923)	•
Message-Topic A message, written or spoken, is about a topic. <i>Cieply's <u>story</u> makes a compelling <u>point</u> about modern-day studio economics.</i>	8.4% (895)	•
Other <i>The <u>child</u> was carefully wrapped and bound into the <u>cradle</u> by means of a cord.</i>	17.4% (1864)	•

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

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

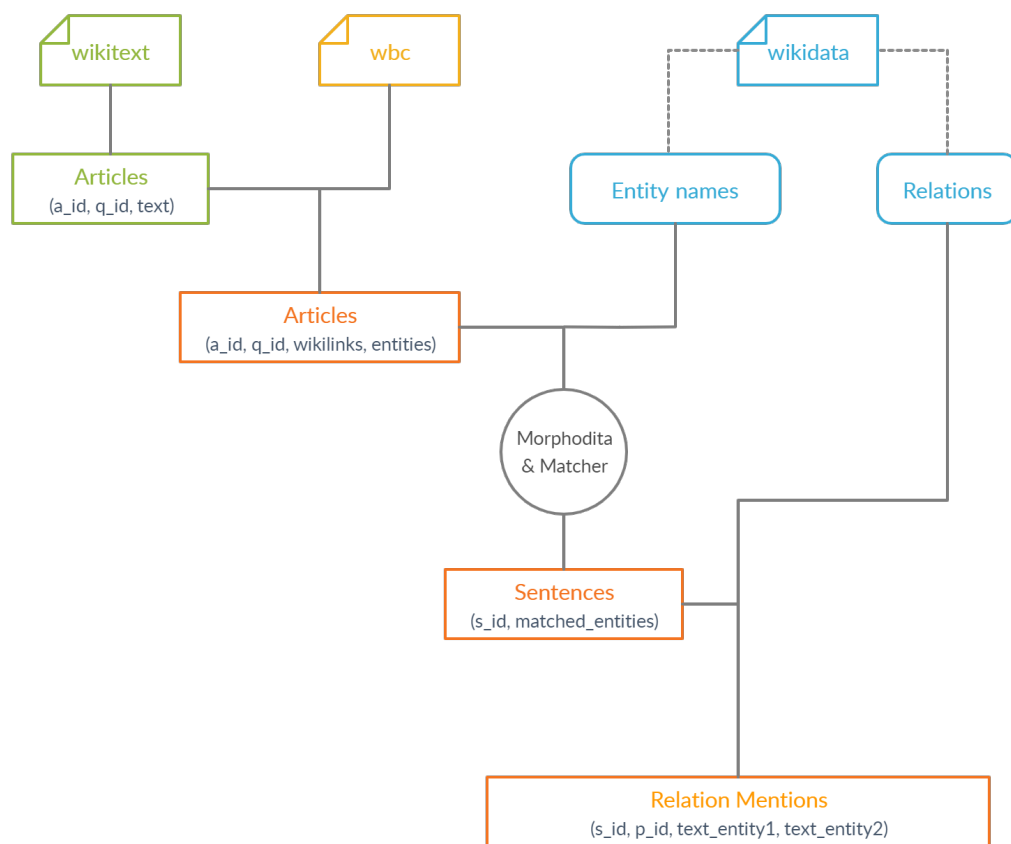


Figure 2.1: Zjednodušený diagram výroby korpusu

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

2.1	Zjednodušený diagram výroby korpusu	6
-----	---	---

List of Tables

1.1	S10T8 summary. List of relations, their official descriptions, a random example and both relative and absolute count.	4
-----	---	---

List of Abbreviations

A. Attachments

A.1 First Attachment