



**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
0.1 Thesis organization	2
1 Datasets	3
1.1 SEMEVAL 2010 task 8 dataset	3
1.2 TACRED dataset	3
2 Title of the second chapter	9
2.1 Title of the first subchapter of the second chapter	9
2.2 Title of the second subchapter of the second chapter	9
Conclusion	11
Bibliography	12
A Attachments	13
A.1 First Attachment	13

Introduction

This thesis researches relationship extraction in Czech. Relationship extraction is the task of extracting semantic relationship from a text. It is closely connected to named entity recognition, the task of tagging entities in text with their corresponding type, and entity linking, the task of disambiguating named entities to a knowledge base. If all those task are used together, we could gain knowledge databases automatically from text.

For English multiple attempts were made to solve or at least advance in relationship extraction, varying both in task assignment and in used technologies.

To be able to approach this set of tasks, we will focus on pure relationship extraction and thus the following restriction: we will only extract relations from sentences with labeled subject and object for the potential relation. We will benefit from the state-of-the-art technologies such as BERT from Devlin et al. [2018].

A key role in modern machine learning play datasets. In major part of this thesis, we will address the absence of a Czech dataset for relationship extraction. We will generate our dataset by aligning Wikidata¹ with Czech Wikipedia². This type of aligning is sometimes referred to as distant supervision. We will also need to recognize entities includes other . We will than be able to train different models and we will also be able to discuss how choices made in dataset generation affect the ability of a model to learn.

Given the absence of a dataset, we also deal with an absence of a baseline for model performance. To show that, at least the proposed architecture and training method we used, are comparable to state of the art result we will perform the same training with English BERT and we will evaluate it on some well known English datasets.

0.1 Thesis organization

This thesis is split in two parts. Before we dive into the first part, we will provide information that is relevant for this thesis, but is not part-specific, such as more details on relationship extraction, connected terminology and further motivation. We will briefly introduce the Czech language to explain why existing distant supervision methods were most likely not applied on Czech.

The first part will focus on datasets. We will present some existing supervised datasets, we will propose methodology for generating the dataset via distant supervision and elaborate on the process of implementation.

In the second part, we will finally talk about the modern technologies, we will try to pinpoint the important aspects of models, etc. we are using. We will use the Transformers³ library which makes training well-known pre-trained models accesible.

¹<https://www.wikidata.org/wiki/>

²<https://cs.wikipedia.org/wiki/>

³<https://github.com/huggingface/transformers/>

divná
věta

previous
work:
Ex-
isting
work on
relation
extrac-
tion
(e.g.,
Zelenko
et al.,
2003;
Mintz
et al.,
2009;
Adel
et al.,
2016)

which
meth-
ods,
were
they
not?

co víc
tam je

je tam
vizuali-
zovatko

whatever
prostě
to
nejdřív
udělej,
pak o
tom piš

1. Datasets

In this chapter, we will overview well-known datasets related to Entity Relationship Extraction. We will start with supervised datasets (SEMEVAL 2010 task 8 and TACRED), then we will focus on distant supervision.

tady představíme existující data

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 součástí 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 úvozkovky

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

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

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

100 000 examples.

The authors are relatively brief about the data collection process:

We create TACRED based on query entities and annotated system responses in the yearly TAC KBP evaluations. In each year of the TAC KBP evaluation (2009–2015), 100 entities (people or organizations) are given as queries, for which participating systems should find associated relations and object entities. We make use of Mechanical Turk to annotate each sentence in the source corpus that contains one of these query entities. For each sentence, we ask crowd workers to annotate both the subject and object entity spans and the relation types.

TACRED relation inventory captures only relations with subject being an organization or a person. Objects are of following types: cause of death, city, country, criminal charge, date, duration, ideology, location, misc (used for alternative name relation and no_relation only), nationality, number, organization, person, religion, state or province, title and url.

TACRED was designed to be highly unbalanced. 79.5% of data is the no_relation relation, which should be closer to real-world text and supposedly should help with not prediction false positive. However even if we look only at actual relations, there are vast differences in frequency: top six relations make up half the dataset and bottom six less than 2%. In absolute numbers the least common org:dissolved relation has only 33 examples and median is only 286 examples.

Table 1.2: TACRED summary. List of relations, their official descriptions, a random example and both relative and absolute count.

no_relation	79.5%
<i>“ <u>One</u> step at a time , ” said Con Edison spokesman Chris Olert in Sunday editions of The <u>Daily News</u> .</i>	(84490)
org:alternate_names	1.3%
<i>The ARMM was established as a result of the peace agreement between the government and the <u>Moro National Liberation Front</u> -LRB- <u>MNLF</u> -RRB- in 1996 .</i>	(1358)
org:city_of_headquarters	0.5%
<i>Once completed , the cuts will leave the <u>Irvine</u> , California-based <u>Option One</u> subsidiary with about 1,400 employees .</i>	(572)
org:country_of_headquarters	0.7%
<i>The Review based its report on a new survey conducted by the <u>International Agency for Research on Cancer</u> in Lyon , <u>France</u> .</i>	(752)
org:dissolved	0.0%
<i>News Corp. sold its satellite television service <u>DirecTV</u> in <u>2008</u> to Liberty Media .</i>	(32)
org:founded	0.2%
<i>New York-based <u>Zirh</u> was founded in <u>1995</u> and makes products using natural oils and extracts .</i>	(165)

org:founded_by	0.3%
<i>The <u>Jerusalem Foundation</u> , a charity founded by <u>Kollek</u> 40 years ago , said he died of natural causes Tuesday morning .</i>	(267)
org:member_of	0.2%
<i>Lyons and the <u>Red Sox</u> say they are n't aware of any other <u>Major League Baseball</u> team with such an arrangement .</i>	(170)
org:members	0.3%
<i>The NFL refused to abandon the city , and the <u>Saints</u> won the <u>NFC South</u> in 2006 , their first season with Brees and Payton .</i>	(285)
org:number_of_employees/members	0.1%
<i>Established in September 1969 , the <u>organization</u> now has <u>57</u> member states worldwide .</i>	(120)
org:parents	0.4%
<i>The initial offering of AIA raised \$ 178 billion for AIG , while the sale of <u>ALICO</u> to <u>MetLife</u> reaped about \$ 155 billion .</i>	(443)
org:political/religious_affiliation	0.1%
<i>Manila signed a peace treaty with the <u>MNLF</u> in 1996 , ending a decades-old separatist campaign in return for limited <u>Muslim</u> self-rule .</i>	(124)
org:shareholders	0.1%
<i>Stop the NAACP and <u>Al Sharpton</u> 's <u>National Action Network</u> from committing this disgrace in our community .</i>	(143)
org:stateorprovince_of_headquarters	0.3%
<i>Learn More <u>Chelsea District Library</u> 221 S Main St Chelsea , <u>MI</u> 48118 -LRB- 734 -RRB- - 475-8732 Find it on a map</i>	(349)
org:subsidiaries	0.4%
<i>The new law will also enable the government to take over <u>Austral Lineas Aereas</u> , an <u>Aerolineas Argentinas</u> subsidiary .</i>	(452)
org:top_members/employees	2.6%
<i>Earlier this year , <u>Anatoly Isaikin</u> , head of <u>Rosoboronexport</u> , said Russia still considers Iran a valuable arms customer .</i>	(2769)
org:website	0.2%
<i><u>Swiss Bankers Association</u> : http://www.swissbanking.org</i>	(222)
per:age	0.8%
<i>Doctor <u>Carolyn Goodman</u> , Rights Champion , Dies at <u>91</u></i>	(832)
per:alternate_names	0.1%
<i><u>Remy Ma</u> , whose real name is <u>Remy Smith</u> , is charged with first - degree assault and other charges .</i>	(152)
per:cause_of_death	0.3%
<i>The cause was <u>kidney failure</u> , said a spokesman for the <u>Ali Akbar</u> College of Music .</i>	(336)
per:charges	0.3%
<i>Actor <u>Danny Glover</u> has been convicted in Canada for <u>trespassing</u> in a hotel during a union rally in 2006 .</i>	(279)
per:children	0.3%
<i><u>Al-Hakim</u> 's son , <u>Ammar al-Hakim</u> , has been groomed for months to take his father 's place .</i>	(346)

per:cities_of_residence	0.7%
<i>As part of a Navy family , <u>she</u> also lived in Long Beach , Calif. , San Diego and <u>Annapolis</u> .</i>	(741)
per:city_of_birth	0.1%
<i>Jane <u>Matilda Bolin</u> was born on April 11 , 1908 , in <u>Poughkeepsie</u> , NY .</i>	(102)
per:city_of_death	0.2%
<i>The statement was confirmed by publicist Maureen O'Connor , who said <u>Dio</u> died in <u>Los Angeles</u> .</i>	(226)
per:countries_of_residence	0.8%
<i>His wife , who accompanied Yoadimnadjì to Paris , will repatriate <u>his</u> body to <u>Chad</u> , the ambassador said .</i>	(818)
per:country_of_birth	0.0%
<i>CARACAS , Jan 10 -LRB- Xinhua -RRB- <u>Hugo Chavez</u> , was born on July 28 , 1954 , in <u>Venezuela</u> 's Sabaneta .</i>	(52)
per:country_of_death	0.1%
<i>Egypt 's state-owned Middle East News Agency said <u>Tantawi</u> died in <u>Saudi Arabia</u> , where he attended a religious ceremony .</i>	(60)
per:date_of_birth	0.1%
<i>Antonioni was born in <u>1912</u> in the northern Italian city of <u>Ferrara</u> .</i>	(102)
per:date_of_death	0.4%
<i><u>December 6 , 2007</u> <u>Jefferson DeBlanc</u> , Hero Pilot , Dies at 86 By RICHARD GOLDSTEIN</i>	(393)
per:employee_of	2.0%
<i><u>He</u> and his group also joined in a legal battle challenging the <u>Washington Redskins</u> ' trademarked name .</i>	(2162)
per:origin	0.6%
<i>French media are reporting that <u>French</u> tennis player <u>Mathieu Montcourt</u> had died at the age of 24 .</i>	(666)
per:other_family	0.3%
<i>In the interview <u>Cunningham</u> acknowledged the fragility of <u>his</u> choreographic record .</i>	(318)
per:parents	0.3%
<i>The outgoing governor of Barinas is <u>Hugo de los Reyes Chavez</u> , father of <u>Hugo</u> and Adan Chavez .</i>	(295)
per:religion	0.1%
<i><u>He</u> closed out the quarter making seven payments to <u>Scientology</u> groups totaling \$ 13,500 .</i>	(152)
per:schools_attended	0.2%
<i><u>She</u> graduated from <u>Mount Holyoke College</u> in 1941 and from the Yale School of Law in 1948 .</i>	(228)
per:siblings	0.2%
<i><u>Raul Castro</u> , <u>Fidel</u> 's younger brother , has made several overtures toward Wash- ington .</i>	(249)
per:spouse	0.5%
<i>After returning to Dothan in 1946 , <u>Flowers</u> married <u>Mary Catherine Russell</u> .</i>	(482)
per:stateorprovince_of_birth	0.1%
<i><u>Thomas Joseph Meskill Jr</u> was born in New Britain , <u>Conn</u> , on Jan 30 , 1928 .</i>	(71)

per:stateorprovince_of_death	0.1%
<i>Jessica Weiner says <u>Greenwich</u> died of a heart attack at St. Luke 's Roosevelt Hospital in <u>New York</u> .</i>	(103)
per:stateorprovinces_of_residence	0.5%
<i>Sen. <u>Chris Dodd</u> of <u>Connecticut</u> has proposed taxing polluters for their carbon emissions .</i>	(483)
per:title	3.6%
<i><u>He</u> is the <u>founder</u> and leader of Architects and Engineers for 9/11 Truth -LRB- <u>AE911Truthorg</u> -RRB- .</i>	(3861)

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

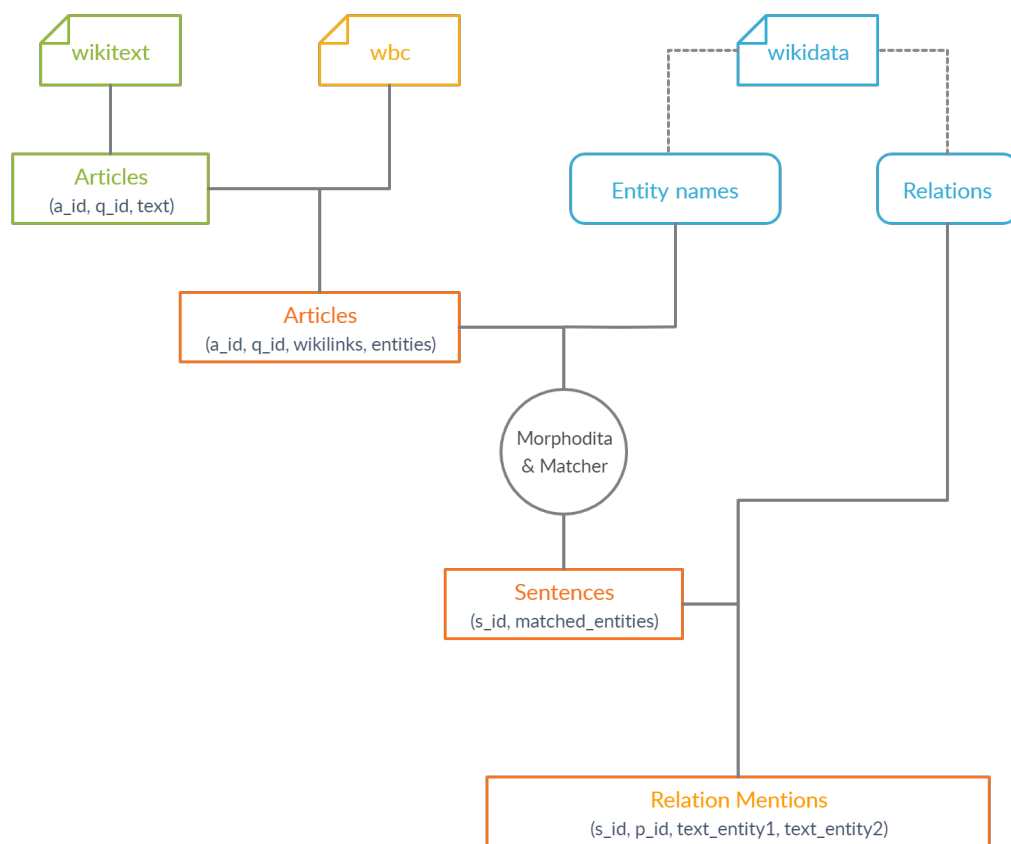


Figure 2.1: Zjednodušený diagram výroby korpusu

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*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/S10-1006>.
- 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>.

A. Attachments

A.1 First Attachment