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

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 **8**
2.1 Title of the first subchapter of the second chapter 8
2.2 Title of the second subchapter of the second chapter 8

Conclusion **10**

Bibliography **11**

A Attachments **12**
A.1 First Attachment 12

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-are technologies such as BERT from Devlin et al. [2018].

divná věta

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 Wikidata[1] with Czech Wikipedia[2]. 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.

.

.

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

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.

which methods, were they not?

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.

co víc tam je

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 Transformers[3] library which makes training well-known pre-trained models accesible.

je tam vizualizovatko

whatever prostě to nejdřív udělej, pak o tom piš

[1] https://www.wikidata.org/wiki/
[2] https://cs.wikipedia.org/wiki/
[3] https://github.com/huggingface/transformers/

1. Datasets

In this chapter, we will overwiev 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 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

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

CAUSE-EFFECT An event or object leads to an effect. Example: *The burst has been caused by water hammer pressure.*	12.4% (1331)
INSTRUMENT-AGENCY An agent uses an instrument. Example: *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. Example: *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. Example: *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). Example: *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. Example: *This book has transported readers into ancient times.*	10.6% (1137)
COMPONENT-WHOLE An object is a component of a larger whole. Example: *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 Example: *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. Example: *Cieply's story makes a compelling point about modern-day studio economics.*	8.4% (895)
OTHER Example: *The child was carefully wrapped and bound into the cradle by means of a cord.*	17.4% (1864)

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 predicting 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 ord: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 Example: *“ One step at a time , ” said Con Edison spokesman Chris Olert in Sunday editions of The Daily News .*	79.5% (84490)
ORG:ALTERNATE_NAMES Example: *The ARMM was established as a result of the peace agreement between the government and the Moro National Liberation Front -LRB- MNLF -RRB- in 1996 .*	1.3% (1358)
ORG:CITY_OF_HEADQUARTERS Example: *Once completed , the cuts will leave the Irvine , California-based Option One subsidiary with about 1,400 employees .*	0.5% (572)
ORG:COUNTRY_OF_HEADQUARTERS Example: *The Review based its report on a new survey conducted by the International Agency for Research on Cancer in Lyon , France .*	0.7% (752)
ORG:DISSOLVED Example: *News Corp. sold its satellite television service DirecTV in 2008 to Liberty Media .*	0.0% (32)
ORG:FOUNDED Example: *New York-based Zirh was founded in 1995 and makes products using natural oils and extracts .*	0.2% (165)
ORG:FOUNDED_BY Example: *The Jerusalem Foundation , a charity founded by Kollek 40 years ago , said he died of natural causes Tuesday morning .*	0.3% (267)

ORG:MEMBER_OF Example: *Lyons and the Red Sox say they are n't aware of any other Major League Baseball team with such an arrangement .*	0.2% (170)
ORG:MEMBERS Example: *The NFL refused to abandon the city , and the Saints won the NFC South in 2006 , their first season with Brees and Payton .*	0.3% (285)
ORG:NUMBER_OF_EMPLOYEES/MEMBERS Example: *Established in September 1969 , the organization now has 57 member states worldwide .*	0.1% (120)
ORG:PARENTS Example: *The initial offering of AIA raised $ 178 billion for AIG , while the sale of ALICO to MetLife reaped about $ 155 billion .*	0.4% (443)
ORG:POLITICAL/RELIGIOUS_AFFILIATION Example: *Manila signed a peace treaty with the MNLF in 1996 , ending a decades-old separatist campaign in return for limited Muslim self-rule .*	0.1% (124)
ORG:SHAREHOLDERS Example: *Stop the NAACP and Al Sharpton 's National Action Network from committing this disgrace in our community .*	0.1% (143)
ORG:STATEORPROVINCE_OF_HEADQUARTERS Example: *Learn More Chelsea District Library 221 S Main St Chelsea , MI 48118 -LRB- 734 -RRB- - 475-8732 Find it on a map*	0.3% (349)
ORG:SUBSIDIARIES Example: *The new law will also enable the government to take over Austral Lineas Aereas , an Aerolineas Argentinas subsidiary .*	0.4% (452)
ORG:TOP_MEMBERS/EMPLOYEES Example: *Earlier this year , Anatoly Isaikin , head of Rosoboronexport , said Russia still considers Iran a valuable arms customer .*	2.6% (2769)
ORG:WEBSITE Example: *Swiss Bankers Association : http://www.swissbanking.org*	0.2% (222)
PER:AGE Example: *Doctor Carolyn Goodman , Rights Champion , Dies at 91*	0.8% (832)
PER:ALTERNATE_NAMES Example: *Remy Ma , whose real name is Remy Smith , is charged with first - degree assault and other charges .*	0.1% (152)
PER:CAUSE_OF_DEATH Example: *The cause was kidney failure , said a spokesman for the Ali Akbar College of Music .*	0.3% (336)
PER:CHARGES Example: *Actor Danny Glover has been convicted in Canada for trespassing in a hotel during a union rally in 2006 .*	0.3% (279)
PER:CHILDREN Example: *Al-Hakim 's son , Ammar al-Hakim , has been groomed for months to take his father 's place .*	0.3% (346)
PER:CITIES_OF_RESIDENCE Example: *As part of a Navy family , she also lived in Long Beach , Calif. , San Diego and Annapolis .*	0.7% (741)
PER:CITY_OF_BIRTH Example: *Jane Matilda Bolin was born on April 11 , 1908 , in Poughkeepsie , NY .*	0.1% (102)
PER:CITY_OF_DEATH Example: *The statement was confirmed by publicist Maureen O'Connor , who said Dio died in Los Angeles .*	0.2% (226)

PER:COUNTRIES_OF_RESIDENCE Example: *His wife , who accompanied Yoadimnadji to Paris , will repatriate <u>his</u> body to <u>Chad</u> , the ambassador said .*	0.8% (818)
PER:COUNTRY_OF_BIRTH Example: *CARACAS , Jan 10 -LRB- Xinhua -RRB- <u>Hugo Chavez</u> , was born on July 28 , 1954 , in <u>Venezuela</u> 's Sabaneta .*	0.0% (52)
PER:COUNTRY_OF_DEATH Example: *Egypt 's state-owned Middle East News Agency said <u>Tantawi</u> died in <u>Saudi Arabia</u> , where he attended a religious ceremony .*	0.1% (60)
PER:DATE_OF_BIRTH Example: *Antonioni was born in <u>1912</u> in the northern Italian city of <u>Ferrara</u> .*	0.1% (102)
PER:DATE_OF_DEATH Example: *<u>December 6 , 2007</u> <u>Jefferson DeBlanc</u> , Hero Pilot , Dies at 86 By RICHARD GOLDSTEIN*	0.4% (393)
PER:EMPLOYEE_OF Example: *<u>He</u> and his group also joined in a legal battle challenging the <u>Washington Redskins</u> ' trademarked name .*	2.0% (2162)
PER:ORIGIN Example: *French media are reporting that <u>French</u> tennis player <u>Mathieu Montcourt</u> had died at the age of 24 .*	0.6% (666)
PER:OTHER_FAMILY Example: *In the interview <u>Cunningham</u> acknowledged the fragility of <u>his</u> choreographic record .*	0.3% (318)
PER:PARENTS Example: *The outgoing governor of Barinas is <u>Hugo de los Reyes Chavez</u> , father of <u>Hugo</u> and Adan Chavez .*	0.3% (295)
PER:RELIGION Example: *<u>He</u> closed out the quarter making seven payments to <u>Scientology</u> groups totaling $ 13,500 .*	0.1% (152)
PER:SCHOOLS_ATTENDED Example: *<u>She</u> graduated from <u>Mount Holyoke College</u> in 1941 and from the Yale School of Law in 1948 .*	0.2% (228)
PER:SIBLINGS Example: *<u>Raul Castro</u> , <u>Fidel</u> 's younger brother , has made several overtures toward Washington .*	0.2% (249)
PER:SPOUSE Example: *After returning to Dothan in 1946 , <u>Flowers</u> married <u>Mary Catherine Russell</u> .*	0.5% (482)
PER:STATEORPROVINCE_OF_BIRTH Example: *<u>Thomas Joseph Meskill</u> Jr was born in New Britain , <u>Conn</u> , on Jan 30 , 1928 .*	0.1% (71)
PER:STATEORPROVINCE_OF_DEATH Example: *Jessica Weiner says <u>Greenwich</u> died of a heart attack at St. Luke 's Roosevelt Hospital in <u>New York</u> .*	0.1% (103)
PER:STATEORPROVINCES_OF_RESIDENCE Example: *Sen. <u>Chris Dodd</u> of <u>Connecticut</u> has proposed taxing polluters for their carbon emissions .*	0.5% (483)
PER:TITLE Example: *<u>He</u> is the <u>founder</u> and leader of Architects and Engineers for 9/11 Truth -LRB- AE911Truthorg -RRB- .*	3.6% (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

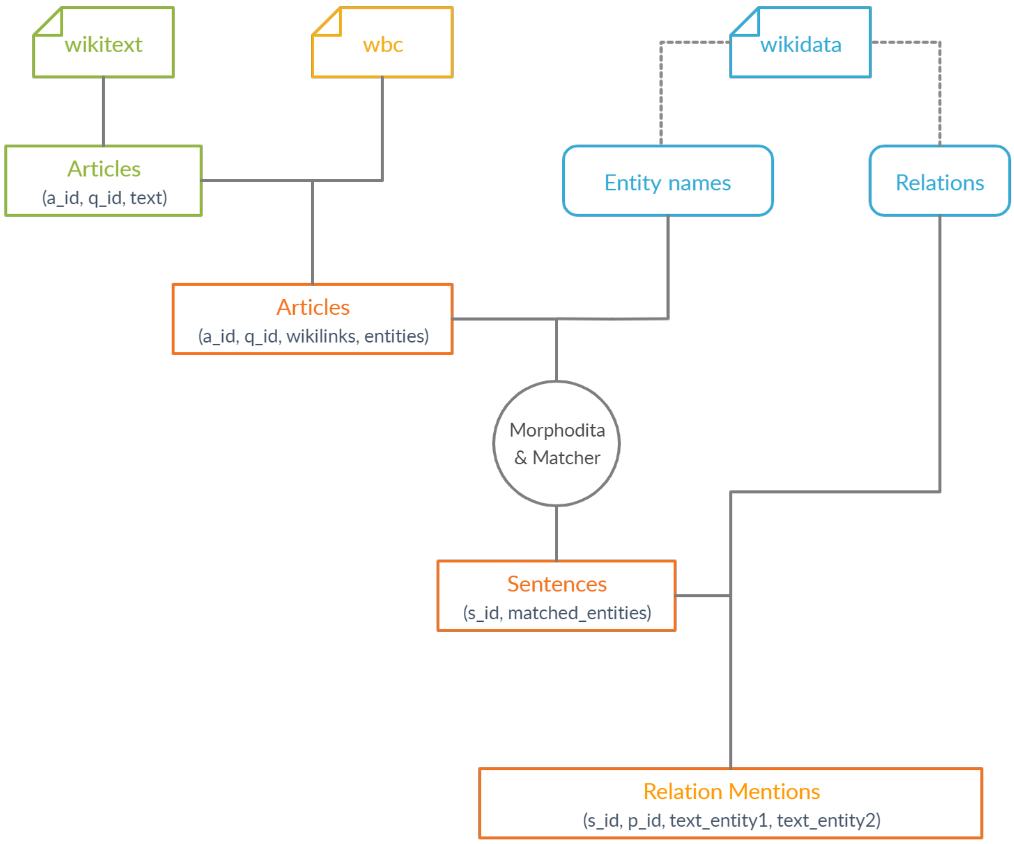

Figure 2.1: Zjednodušený diagram výroby korpusu

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