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

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.

.

.

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

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. which methods, were they not?

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. co víc tam je 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. Relationship extraction intro

1.1 Terminology

Terminology in NLP subtasks is often not exact or non-standardized. We will attempt to introduce following concepts as exactly as possible and respecting the terms that seem to be established by majority.

Relation in this context is semantic (not grammatical etc.). It has a type, is binary and oriented and describes relationship between a relation subject and a relation object.

Relation subject and **relation object**. Subject is the first argument of relation, object the second. In the sentence "Albus Severus is Harry Potters's son." a relation of typeSON is captured, subject is John and object is Eric. The reasoning for this choice of direction is as follows: suppose we are gathering information about Harry, than we would probably have both the information that his son is Albus Severus and his father is James. So we are gathering information about the subject (Harry Potter), even though in most sentences like "James is Harry's father." Harry is a grammatical object. We will use the notation RELATION TYPE(subject,object): SON(Harry Potter,Albus Severus Potter).

Both subject and object can generally be any word or sequence of words that have the ability to form relations. In some cases subjects, objects or both are limited to entities or named entities.

Named entity is a real-world object, such as persons, locations, organizations, products, etc., that can be denoted with a proper name. It can be abstract or have a physical existence. Named entities can simply be viewed as entity instances (e.g., New York City is an instance of a city). Sometimes, numeric data is considered in this category as well (for example by NER tools).

Relation inventory is the set of types of relations, that are considered valid for given dataset or model.

Relation mention is a sentence, that captures a relation, together with type of the relation and tagged subject and object.

Negative mention is close to relation mention in the sense that it is a sentence with tagged subject and object, but the relation type is one of these types:

- OTHER - human annotator would classify a relation of type, that is not in relation inventory.

- NO RELATION - in this case, human annotator should feel an absence of relation between subject and object.

NO RELATION comes with difficulties. Since there is no semantic relation between subject and object, it makes it harder to choose subject-object pairs. It is probably desirable to have subject-object pairs, that could be related in a different sentence.

Relationship Extraction

Lemma

1.2 Czech language

One of the objective of this thesis is to work with Czech language, therefore we find it useful to make some notes on Czech (for non Czech speaking readers). Czech is a Slavic language with rich morphology and relatively free word order. Most of Czech morphology can be treated with a morphological analyzer, still, it might be useful to have a better understanding of the language we will work with.

1.2.1 Declanation

In Czech, nouns, adjectives, pronounce and numerals are inflected. The inflection expresses (not necessarily unambiguously) one of seven cases and a number (singular or plural). Any inflected word in Czech has a grammatical gender, for words, that have natural gender, those two genders align: "žena" (*woman*) is feminime and "muž" (*man*) is masculine. Inflection of each inflective word follows a pattern. This all means that a single word (lemma) can have a lexeme of size

Verbs are conjugated, the conjugation expresses person, numeral, tense, voice and mode. Verbs follow one of 14 patterns and average Czech either finds the theory about Czech verbs and tenses confusing, or is unaware there even are verb patterns. With that, we will not elaborate on conjugation.

An important aspect of declanation for us is agreement. In English, subject and verb agrees (limited just to third person). In Czech subject and verb also agree, but in noun phrases there needs to be an agreement as well.

An abstract entity; the set of all forms related by inflection (but not derivation).

příklad: toho, jak je nějaká noun phrase, počet lexémů, počet validních ..

odkaz dopředu, kde řeším, jak matchovat

2. Existing 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.

2.1   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.

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 2.1.

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).

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.

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.

2.2   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 2.1: 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)
Example: *The <u>burst</u> has been caused by water hammer <u>pressure</u>.*	
INSTRUMENT-AGENCY	6.2%
An agent uses an instrument.	(660)
Example: *The <u>author</u> of a keygen uses a <u>disassembler</u> to look at the raw assembly code.*	
PRODUCT-PRODUCER	8.8%
A producer causes a product to exist.	(948)
Example: *The <u>factory</u>'s products have included flower pots, Finnish rooster-whistles, pans, <u>trays</u>, tea pots, ash trays and air moisturisers.*	
CONTENT-CONTAINER	6.8%
An object is physically stored in a delineated area of space.	(732)
Example: *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.*	
ENTITY-ORIGIN	9.1%
An entity is coming or is derived from an origin (e.g., position or material).	(974)
Example: *The <u>avalanches</u> originated in an extensive <u>mass</u> of rock that had previously been hydrothermally altered in large part to clay.*	
ENTITY-DESTINATION	10.6%
An entity is moving towards a destination.	(1137)
Example: *This book has transported <u>readers</u> into <u>ancient times</u>.*	
COMPONENT-WHOLE	11.7%
An object is a component of a larger whole.	(1253)
Example: *The system as described above has its greatest application in an arrayed <u>configuration</u> of antenna <u>elements</u>.*	
MEMBER-COLLECTION	8.6%
A member forms a nonfunctional part of a collection	(923)
Example: *The <u>student</u> <u>association</u> is the voice of the undergraduate student population of the State University of New York at Buffalo.*	
MESSAGE-TOPIC	8.4%
A message, written or spoken, is about a topic.	(895)
Example: *Cieply's <u>story</u> makes a compelling <u>point</u> about modern-day studio economics.*	
OTHER	17.4%
	(1864)
Example: *The <u>child</u> was carefully wrapped and bound into the <u>cradle</u> by means of a cord.*	

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

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

ORG:MEMBER_OF	0.2%
Example: *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 .*	(170)
ORG:MEMBERS	0.3%
Example: *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 .*	(285)
ORG:NUMBER_OF_EMPLOYEES/MEMBERS	0.1%
Example: *Established in September 1969 , the <u>organization</u> now has <u>57</u> member states worldwide .*	(120)
ORG:PARENTS	0.4%
Example: *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 .*	(443)
ORG:POLITICAL/RELIGIOUS_AFFILIATION	0.1%
Example: *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 .*	(124)
ORG:SHAREHOLDERS	0.1%
Example: *Stop the NAACP and <u>Al Sharpton</u> 's <u>National Action Network</u> from committing this disgrace in our community .*	(143)
ORG:STATEORPROVINCE_OF_HEADQUARTERS	0.3%
Example: *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*	(349)
ORG:SUBSIDIARIES	0.4%
Example: *The new law will also enable the government to take over <u>Austral Lineas Aereas</u> , an <u>Aerolineas Argentinas</u> subsidiary .*	(452)
ORG:TOP_MEMBERS/EMPLOYEES	2.6%
Example: *Earlier this year , <u>Anatoly Isaikin</u> , head of <u>Rosoboronexport</u> , said Russia still considers Iran a valuable arms customer .*	(2769)
ORG:WEBSITE	0.2%
Example: *<u>Swiss Bankers Association</u> : <u>http://www.swissbanking.org</u>*	(222)
PER:AGE	0.8%
Example: *Doctor <u>Carolyn Goodman</u> , Rights Champion , Dies at <u>91</u>*	(832)
PER:ALTERNATE_NAMES	0.1%
Example: *<u>Remy Ma</u> , whose real name is <u>Remy Smith</u> , is charged with first - degree assault and other charges .*	(152)
PER:CAUSE_OF_DEATH	0.3%
Example: *The cause was <u>kidney failure</u> , said a spokesman for the <u>Ali Akbar College of Music</u> .*	(336)
PER:CHARGES	0.3%
Example: *Actor <u>Danny Glover</u> has been convicted in Canada for <u>trespassing</u> in a hotel during a union rally in 2006 .*	(279)
PER:CHILDREN	0.3%
Example: *<u>Al-Hakim</u> 's son , <u>Ammar al-Hakim</u> , has been groomed for months to take his father 's place .*	(346)
PER:CITIES_OF_RESIDENCE	0.7%
Example: *As part of a Navy family , <u>she</u> also lived in <u>Long Beach</u> , Calif. , <u>San Diego</u> and <u>Annapolis</u> .*	(741)
PER:CITY_OF_BIRTH	0.1%
Example: *<u>Jane Matilda Bolin</u> was born on April 11 , 1908 , in <u>Poughkeepsie</u> , NY .*	(102)
PER:CITY_OF_DEATH	0.2%
Example: *The statement was confirmed by publicist Maureen O'Connor , who said <u>Dio</u> died in <u>Los Angeles</u> .*	(226)

PER:COUNTRIES_OF_RESIDENCE Example: _His wife , who accompanied Yoadimnadji to Paris , will repatriate his body to Chad , the ambassador said ._	0.8% (818)
PER:COUNTRY_OF_BIRTH Example: _CARACAS , Jan 10 -LRB- Xinhua -RRB- Hugo Chavez , was born on July 28 , 1954 , in Venezuela 's Sabaneta ._	0.0% (52)
PER:COUNTRY_OF_DEATH Example: _Egypt 's state-owned Middle East News Agency said Tantawi died in Saudi Arabia , where he attended a religious ceremony ._	0.1% (60)
PER:DATE_OF_BIRTH Example: _Antonioni was born in 1912 in the northern Italian city of Ferrara ._	0.1% (102)
PER:DATE_OF_DEATH Example: _December 6 , 2007 Jefferson DeBlanc , Hero Pilot , Dies at 86 By RICHARD GOLDSTEIN_	0.4% (393)
PER:EMPLOYEE_OF Example: _He and his group also joined in a legal battle challenging the Washington Redskins ' trademarked name ._	2.0% (2162)
PER:ORIGIN Example: _French media are reporting that French tennis player Mathieu Montcourt had died at the age of 24 ._	0.6% (666)
PER:OTHER_FAMILY Example: _In the interview Cunningham acknowledged the fragility of his choreographic record ._	0.3% (318)
PER:PARENTS Example: _The outgoing governor of Barinas is Hugo de los Reyes Chavez , father of Hugo and Adan Chavez ._	0.3% (295)
PER:RELIGION Example: _He closed out the quarter making seven payments to Scientology groups totaling $ 13,500 ._	0.1% (152)
PER:SCHOOLS_ATTENDED Example: _She graduated from Mount Holyoke College in 1941 and from the Yale School of Law in 1948 ._	0.2% (228)
PER:SIBLINGS Example: _Raul Castro , Fidel 's younger brother , has made several overtures toward Washington ._	0.2% (249)
PER:SPOUSE Example: _After returning to Dothan in 1946 , Flowers married Mary Catherine Russell ._	0.5% (482)
PER:STATEORPROVINCE_OF_BIRTH Example: _Thomas Joseph Meskill Jr was born in New Britain , Conn , on Jan 30 , 1928 ._	0.1% (71)
PER:STATEORPROVINCE_OF_DEATH Example: _Jessica Weiner says Greenwich died of a heart attack at St. Luke 's Roosevelt Hospital in New York ._	0.1% (103)
PER:STATEORPROVINCES_OF_RESIDENCE Example: _Sen. Chris Dodd of Connecticut has proposed taxing polluters for their carbon emissions ._	0.5% (483)
PER:TITLE Example: _He is the founder and leader of Architects and Engineers for 9/11 Truth -LRB- AE911Truthorg -RRB- ._	3.6% (3861)

3. CERED

In this chapter we will describe the process of generating **C**zech **R**elationship **E**xtraction **D**ataset (CERED). We will discuss various decisions that were made during this process and their impacts. We will start by characterizing available data and technological resources. ┌─────────────────── a co dál

3.1 Czech Wikipedia and Wikidata

Wikipedia is a multilingual online encyclopedia created and maintained as an open collaboration project by a community of volunteers according to Wikipedia contributors [2020]. From out point of view, Wikipedia is a corpus of text

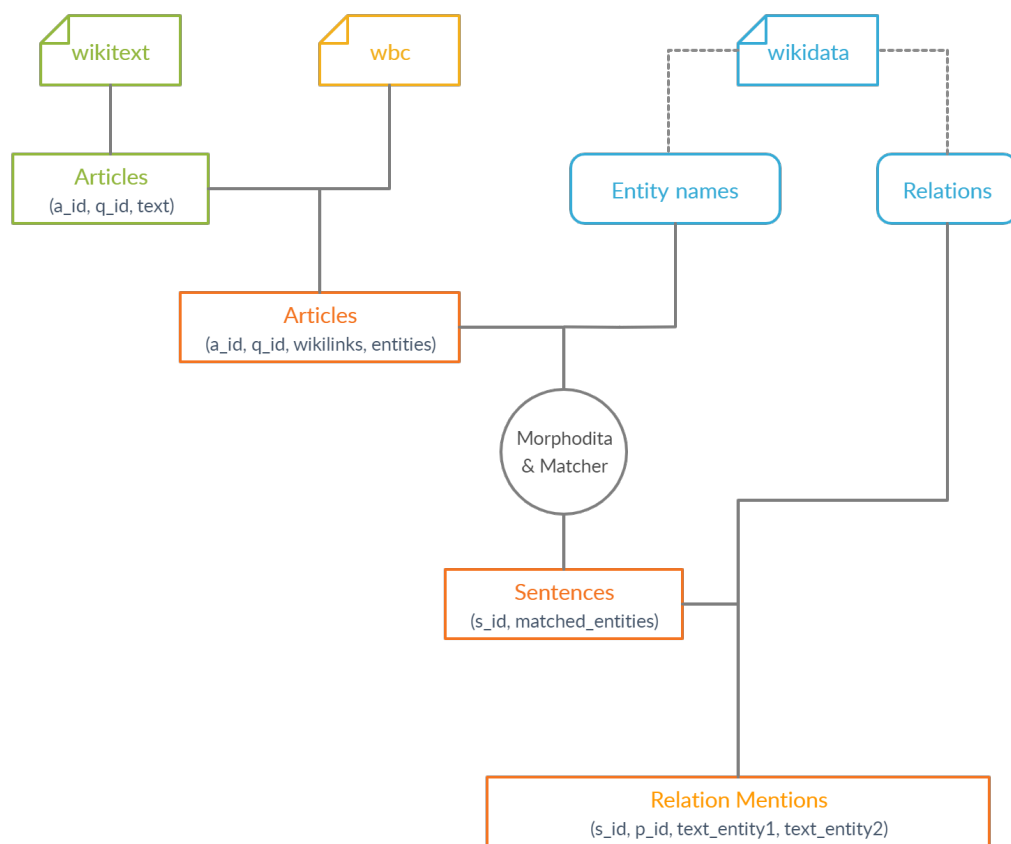## 3.2 Title of the second subchapter of the second chapter

Figure 3.1: Zjednodušený diagram výroby korpusu

4. Existing ML věci

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`.

Wikipedia contributors. Wikipedia — Wikipedia, the free encyclopedia, 2020. URL `https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=947302871`. [Online; accessed 28-March-2020].

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