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 Relationship extraction intro **4**
1.1 Terminology . . . 4
1.2 Czech language . . . 5
1.2.1 Inflection . . . 5
1.2.2 Free word order . . . 5

2 Existing datasets **6**
2.1 SEMEVAL 2010 task 8 dataset . . . 6
2.2 TACRED dataset . . . 6

3 CERED **11**
3.1 Overview . . . 11
3.2 Data sources . . . 11
3.2.1 Czech Wikipedia . . . 12
3.2.2 Wikidata . . . 12
3.3 Analysis . . . 13
3.3.1 Dataflow . . . 13
3.3.2 Entity matching . . . 16
3.3.3 silver data . . . 16
3.3.4 Distant supervision assumption . . . 16
3.3.5 možná to bude chtít zobrazovátko? aneb jak hodnotit kvalitu? . . . 16
3.3.6 Jaké kategorie? . . . 16
3.4 Used technologies . . . 16
3.4.1 Python . . . 16
3.4.2 Spark . . . 17
3.4.3 MorphoDiTa . . . 17
3.4.4 Streamlit . . . 17
3.5 Implementation . . . 17
3.5.1 WikiManipulace . . . 17
3.6 Viewer . . . 17
3.7 Results . . . 17

4 Existing ML věci **19**

Conclusion **20**

Bibliography **21**

A Attachments **22**
A.1 First Attachment . . . 22

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 and obtained results.

In the second part, we will finally talk about the modern deep learning 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

co víc tam je

je tam vizualizovatko

[1] https://www.wikidata.org/wiki/

[2] https://cs.wikipedia.org/wiki/

[3] https://github.com/huggingface/transformers/

pre-trained models accesible.

whatever, prostě to nejdřív udělej, pak o tom piš

1. Relationship extraction intro

změnit název

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.

hloupá věta

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.

někam zapracovat příklad na větě, jak je nejdřív najití rozštíření a tak všechno

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).

will we, lepší vzhled

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).

ukradeno z wiki, ocitovat nebo ukráct od jinud

rozepsat ner

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:

odrážky

- 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.

doplnit příklady, včetně vyloženě ne příkladu

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

znímit entity z názvu

Lemma

A form from a lexeme chosen by convention (e.g., nomi-

Lexeme
Noun phrase

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 Inflection

In Czech, nouns, adjectives, pronounce and numerals are declined. 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 declinable word follows a pattern. This all means that a single word (lemma) can have a lexeme of size

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

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.

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

odkaz dopředu, kde řeším, jak matchovat

https://www.aclw
5003.pdf, statistiky o češtině

1.2.2 Free word order

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.

tady představíme existující dataesty

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.

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 2.1.

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.

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 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 2.2: 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 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 **Cz**ech **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

nějaká návaznost - říct, že popisuju, že budu spojovat wikipedie

3.1 Overview

The objective is to create a Relationship Extraction dataset for Czech language using distant supervision. This section is a quick summary for easier orientation in this chapter. Each of these paragraphs is a teaser for one section of this chapter.

link zpátky

First we researched available knowledge bases and Czech corpora to determine which ones will best suit our purpose. We chose Wikimedia projects Wikidata and Czech Wikipedia.

Next we analysed how we will find mentions of Wikidata relations in Czech Wikipedia. We sketched out first dataflow diagrams and thought about all the different complex aspects of this task.

We continued with choosing technologies that we will use. Aware of the volume and other characteristics of chosen data, we chose Python as the main programming language, spark as a way to speed up the computations and MorphoDita to deal with Czech language.

Změnit, když už máme analýzu

Than we started the implementation and realized that this seemingly simple problem is rather complex. Even though all that we wanted was to get sentences from Wikipedia, find words or phrases, that can have relations, and link those relations from Wikidata, the number of decisions we had to make and obstacles we had to overcome was rather surprising.

lepší formulace

As a side project, we implemented a simple viewer, that can present the dataset

results

3.2 Data sources

To be able to perform distant supervision, we need to find suitable data - Czech text corpus and a knowledge base. We will explain the requirements and constraints we have on such data and present our options. In this section, we will provide more information on the chosen ones.

The main constraint is quite straightforward, there has to be nontrivial shared set of entities and relations mentioned in text and stored in knowledge base. We expect more fact based texts to be more suitable, leaning towards encyclopedic or journalistic genre . One option is to focus on some subset of Czech National Corpus [1], for example SYN2013PUB, SYN2009PUB and SYN2009PUB are corpora of written journalism. The other option is to lean in the direction of encyclopedic text with Czech Wikipedia.

tak špatná věta

[1] https://www.korpus.cz/

Our options for knowledge base are limited, to the best of our knowledge, to Wikidata or Google Knowledge graph [2].

We decided to use Czech Wikipedia and Wikidata, mostly because the intersection of information expressed in text data and in structured data seems promising.

Další důvody.. zmínit že třeba celkem multilinqual? že jde stáhnout? že není black-box? lepší disambiguita

3.2.1 Czech Wikipedia

Wikipedia is a multilingual online encyclopedia created and maintained as an open collaboration project by a community of volunteersWikipedia contributors [2020] and we believe anyone reading this article is familiar with Wikipedia. From out point of view Wikipedia is a corpus of text with tagged topics of articles and some entity mentions. Czech Wikipedia contains approximately 440 000 articles and ranks top 30 across all the different language editions of Wikipedia. [3]

A dump of Czech Wikipedia is about 1,6GB and 770MB when compressed.

3.2.2 Wikidata

Wikidata is a knowledge base, which acts as central storage of the structured data of Wikipedia and other Wikimedia projects. Just like Wikipedia, this project is freely available and edited by users (and bots). It provides the option to query the database online (for small enough queries), but it is also possible to download the database in standard formats.

The database focuses on **items**, which represent objects, entities, concepts, etc. The first data collected in Wikidata were links to multilingual version of Wikipedia articles on the same topic, the same Wikidata item. Each item was assigned an identifier, prefix Q and unique number, referred to as **QID**. A label together with a description of an item should serve as a human readable identifier. Labels, descriptions a optional aliases are language dependant.

Properties, another big concept of Wikidata, can be thought of as categories of items (mother *P25* implies a category of all mothers) or as relations between items (Ron Weasley *Q173998* has a mother *P25* Molly Weasley *Q3255012*). Each property has its **PID**, an identifier consisting of a prefix P and an unique number, and a data type for a value it can be paired with (such as an item, string, url, number or media file).

hezčí formátování wiki-itemm

Information about any item is recorded in statements. Statement is a key-value pair of an property and a value of prescribed data type. For example, for Ron Weasley *Q173998* there are seven statements about his siblings:

- sibling *P3373* Ginny Weasley *Q187923*,
- sibling *P3373* Fred Weasley *Q13359612*,
- sibling *P3373* George Weasley *Q13359613* and so on

.

formating

[2] https://developers.google.com/knowledge-graph

[3] As of March 2020 according to https://en.wikipedia.org/wiki/List_of_Wikipedias

Wikidata project contains over 80 000 000 items, which raises requirements on technological resources, that we will need to work efficiently with such data. Json dump of Wikidata takes 110GB of disk space or 37GB if bzip2 compressed.

3.3 Analysis

The process of creation of CERED is mostly an attempt to execute the first two part of a pipeline we mention in the first chapter. To the best of our knowledge, there is no suitable entity linking tool for Czech. There are tools for Named Entity Recognition that we could theoretically use to our advantage if we decided to focus on named entity only.

Therefore we need to find a way, to get to similar results as the pipeline would get. We do not expect that our CERED generator will be as powerful as respective dedicated tools would be. We will not try to create a general entity recognition and linking tools - on the contrary we will exploit any extra information that chosen Wikimedia projects provide.

There are several aspect that we need to think through.

- Dataflow - jaké info kde vezmeme a s čím ho v jakém pořadí spojíme?
- Entity Matching - jak matchoat?
- Silver data - konečně přejmenovat btw! Snad čiscí?
- Vztahy - jak matchovat, jaké kategorie?
- zobrazovátko?

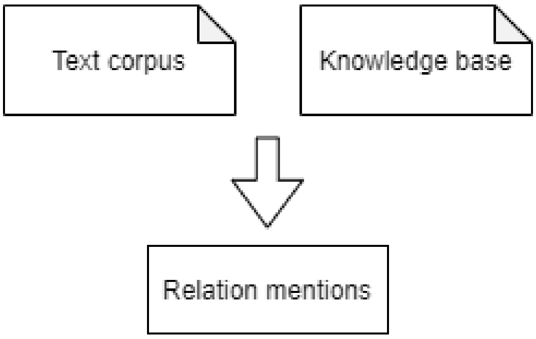
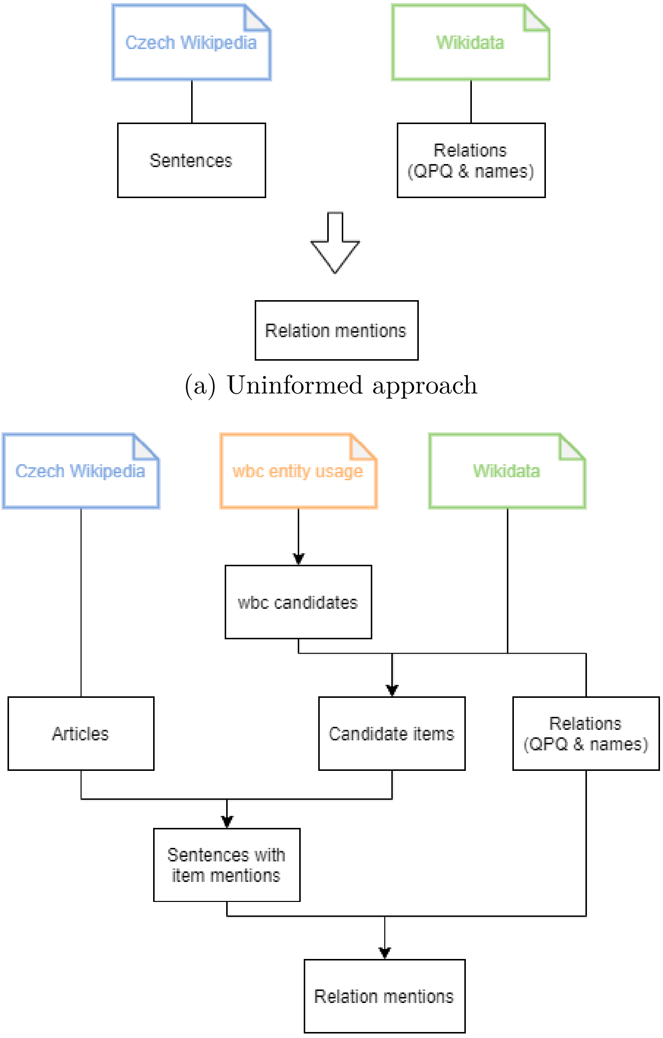


Figure 3.1: Distant supervision diagram

3.3.1 Dataflow

We are starting with two files. One being a Czech Wikipedia dump: it is a collection of articles. Each article has, among other information, its title, id and text. The other is a Wikidata dump. The simpliest way of processing those files would be to process them separately and thus obtaining sentences on one side and relations (a relation type with two items) on the other, see 3.2. This approach comes with a clear disadvantage. We would lose any additional information to the sentences, that could be potentially useful (for example article title might



(a) Uninformed approach

(b) Informed approach

be helpful to determine which items are mentioned in a sentence). To solve this we could precompute something for each article and attach it to each sentence, risking a massive increase in required capacity to work with such data. On a similar note, if we were to follow the diagram exactly, we would probably store item names (labels and aliases) in each relation, worsening the situation even further.

We decided to update the dataflow to address those issues. We will preprocess Wikidata dump to contain only the data we will use. An item will be kept only if it has a Czech name and we will significantly reduce its statements: we will keep title of its Czech Wikipedia article and create a list of (QID,PID,QID) triples - **QPQ**, representing statements that contained information about relations between between this and other items. This way, we have all the necessary information - article title to be able to connect article to item, names for each item to be able to find mentions of items and finally QPQ triples to connect relations and sentences.

One approach to finding item mentions in text could be called uninformed. We could assume that any item can be mentioned in any sentence. This approach seems to have two main issues: the computation would likely take quite some time but mainly we expect a huge amount of ambiguous mentions. An example of this ambiguity, that we seen as problematic might be children named after their parents. In this case, not only that the entity might get confused, moreover, if we then assign the relation, we might easily confuse a sentence mentioning a spouse relation for a parent relation mention.

On the other side, we can use the extra information that Wikimedia projects provide and opt for a more informed approach. The topic of most Czech Wikipedia articles is a Wikidata item, therefore this item is nearly certainly mentioned in the article. Some Wikidata statements were based on relevant articles and thus it seem rational to expect items, that are related to the main item of an article, to be mentioned. We decided to look only for a tiny subset of all Wikidata items in each article - **candidate items**. As we just discussed, if an article is based on a item, than this item and all items, that are connected to it by a statement, are considered candidate items.

Czech Wikipedia maintains a wbc entity usage table, which contains information about which article uses which item. If we use this table, we are able to obtain a list of items, that should be mentioned in an article, lets call this list a **wbc candidates**. A wbc candidate is at the same time a candidate item.

We might consider adding even a second level of relatives (items related to items that are related to the main item) but the branching factor might be relatively high and cause unwanted ambiguity. Consider an instance item like a concrete country, all countries would be second level relatives and thus a candidate item. Since countries tend to be of certain type (kingdom, republic, state etc.) there might be simply the type or some other more general name amongst their names (Q30 United States of America are also known as America or United States) and more countries might share this name.

typografie

So far we mostly discussed advantages of proposed informed approach, mainly a hope for higher precision specifically higher precision for item mentions. We should elaborate on some disadvantages as well. We are not trying to fully do entity linking. In the end, we will only use item mentions, if there are two of them in one sentence and if a QPQ that connects them exists. It is debatable whether we need an informed approach to increase relation mention precision, the necessity and improbability that this condition will be fulfilled for false positive item mentions might in fact be sufficient.

One more way to locate item mentions are **wikilinks**. A wikilink links a page to another page within same-language Wikipedia. First additional information this brings is simply the item mention (if the linked page or article has its main item). We can also consider the textual part of the link to be another name for the linked item. The quality and suitability of this name is to be examined and if we will find these names useful, they can be added to official Wikidata names.

sice by šlo neparalelně, ale co rychlost?

Mluvit o tom, proč nejdřív najdeme, co v článku hledat, pak to nasekáme na věty, pak matchujeme. Zmínit, kolik je jiných možností, že teoreticky by šlo ještě před rozsekáním na věty dělat entity linking ...

Detailně popsat, co kdy kam poteče + diagram

3.3.2 Entity matching

We have text on one side, gathered candidate items on the other and the goal is to label words that represent those items - entity mentions. There is a wide spectrum of complexity we might aim for, but once again, we are not trying to create a tool for entity recognition and linking. We will describe some of those complexity tears, but before that we should address some basic text manipulation that all of the matching methods will benefit from.

String equality. Definitely the easiest method which just test whether **String similarity Morphological analysis zájmena a více vět a tak**

but the choice of final method was mostly based on performance measured once the implementation was done.

Napsat, že v aj dělají často jen exact modulo zkratky a malé přípony, což tady nejde.

Ukázat nápady se sebráním linků z wikipedie a zavrhnout to

připomenout, jak moc se dá čeština skloňovat říct, že nemá nejspíš smysl snažit se najít jen validní tvary, protože stejně v textu nejspíš nebudou nevalidní

asi mluvit o word order? a možná i implementovat

Říct, že jako kontrolní dataset budou přímo z linků

3.3.3 silver data

Co to je proč je nejspíš větší šance na kvalitu možná i ručně udělat měření kvality na těhle datech v druhé části

3.3.4 Distant supervision assumption

aneb jak matchovat vztahy

3.3.5 možná to bude chtít zobrazovátko? aneb jak hodnotit kvalitu?

možná navrhout nějaké random dotazy na odhalení nevalidních dat?

3.3.6 Jaké kategorie?

3.4 Used technologies

We chose Python to be our main programming language. To be able to work faster with bigger volume of data, we wanted to use a cluster, which leads to Spark and occasionally to some small scripts in shell/bash. To top it, we will use MorphoDita to work with Czech language. Later, we will mention a simple Streamlit app we used to comfortably see results of our Spark queries.

koukli jsme se na diagram a mysleli, že celé ve spark a tak

co je co?

In this section we will briefly introduce those technologies.

3.4.1 Python

zmiň pandy a numpy a tak

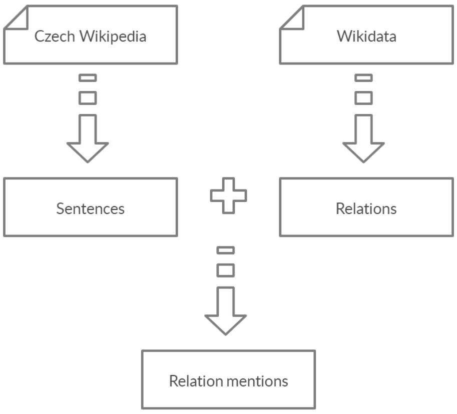


Figure 3.2: A very simple diagram of dataset generation.

3.4.2 Spark

3.4.3 MorphoDiTa

MorphoDiTa Straková et al. [2014] (Morphological Dictionary and Tagger) is an open-source tool for morphological analysis of natural language texts. It is designed to work well on inflective languages and achieves state-of-the-art results for Czech language. Internally, during training tries are built to represent patterns for declension. Externally, MorphoDiTa API provides functionalities such as splitting text into sentences, tokenization and lemmatization.

3.4.4 Streamlit

3.5 Implementation

3.5.1 WikiManipulace

Popsat, že wikitext má sám o sobě strukturovaná a divná data a že je to třeba nejspíš řešit Přesunout do implementace?

3.6 Viewer

3.7 Results

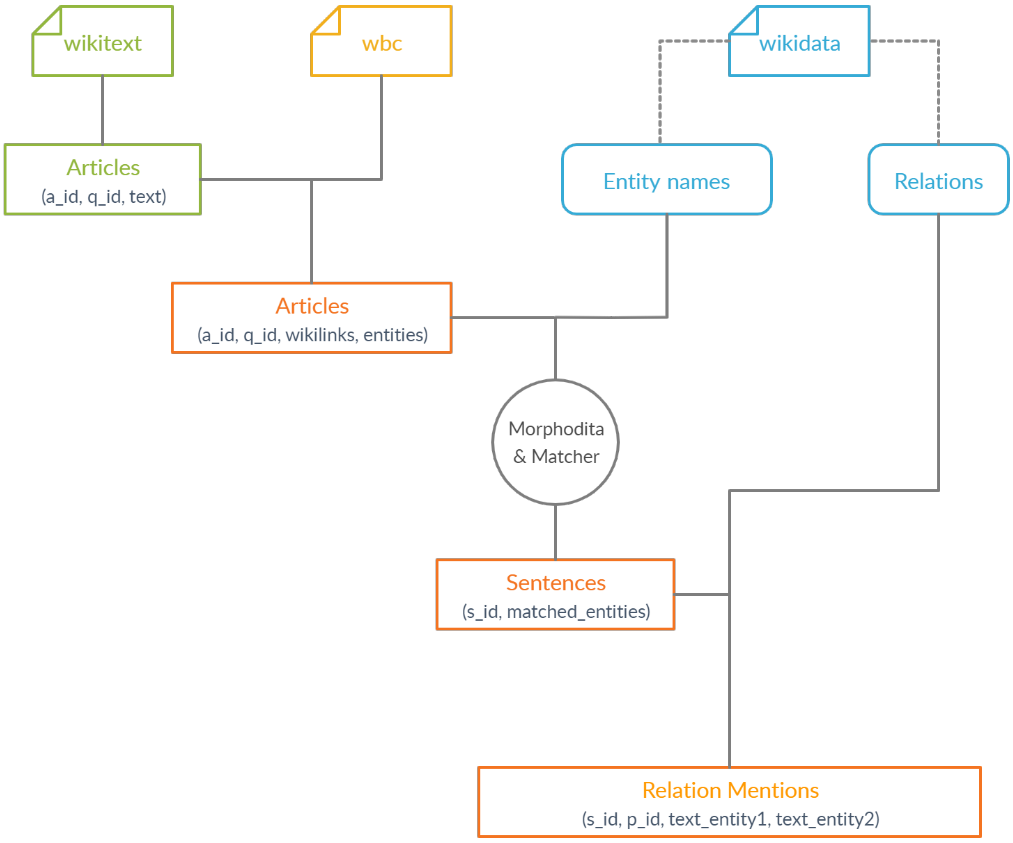


Figure 3.3: Zjednodušený diagram výroby korpusu

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`.

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P14/P14-5003.pdf`.

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