



**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

I would like to thank Profinit EU s.r.o. that they allowed me to create this thesis as part of their Big Data department, and Ing. Marek Sušický, the head of the department, for supervision. I would also like to thank my supervisor RNDr. Milan Straka, Ph.D. for his guidance and expertise.

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: entities named entities entity relationship entity relationship extraction

Contents

Introduction	3
1 Background	4
1.1 Terminology	4
1.2 Relationship Extraction	5
1.3 Czech language	5
1.3.1 Inflexion	5
1.3.2 Word order	7
I Datasets	8
2 Existing datasets	9
2.1 SEMEVAL 2010 task 8 dataset	9
2.2 TACRED dataset	11
2.3 Riedel NYT dataset	12
3 CERED	14
3.1 Overview	14
3.2 Data sources	14
3.2.1 Constraints and requirements	14
3.2.2 Czech Wikipedia	15
3.2.3 Wikidata	15
3.3 Analysis	16
3.3.1 Dataflow	17
3.3.2 Entity matching	19
3.3.3 Wikilink Mentions	24
3.3.4 Relation Matching	24
3.3.5 Relation Inventory	25
3.4 Used Technologies	27
3.4.1 Python	28
3.4.2 Spark	28
3.4.3 MorphoDiTa	28
3.4.4 Streamlit	28
3.5 Implementation	28
3.5.1 Wikidata Preprocessing	28
3.5.2 Wikitext Parsing	29
3.5.3 Entity Matching	30
3.5.4 Relation Matching	31
3.5.5 Characteristics of the Generated Dataset	32
3.6 CERED Versions	33
3.6.1 CERED0	33
3.6.2 CERED1	33
3.6.3 CERED2	34
3.6.4 CERED3	34

II	Training	36
4	Previous Work on Relationship Extraction	37
4.1	Deep NLP models	37
4.1.1	The Transformer Architecture	37
4.1.2	Pre-training	38
4.1.3	BERT	39
4.2	Metrics	40
4.2.1	Binary classification	40
4.2.2	Multiclass classification	41
4.2.3	Relationship Extraction Metrics	43
4.3	Relationship Extraction Models	44
5	Relationship Extraction on CERED	45
5.1	Model	45
5.2	Results	45
5.2.1	S10T8	46
5.2.2	TACRED	46
5.2.3	Riedel NYT	46
5.2.4	CEREDs	46
	Conclusion	48
5.3	Future work	48
5.3.1	Other Languages	48
5.3.2	Wikidata ontology	48
	Bibliography	49
A	Attachments	51
A.1	First Attachment	51

Introduction

Relationship extraction is the task of extracting semantic relationships 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 these task are used together, we can, for example, construct a knowledge database automatically from text.

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

In this thesis, we will focus on relationship extraction in the Czech language. Our goal is to construct a neural network model that will extract relationships from sentences with labelled subject and object entities.

In modern machine learning, the quality of datasets plays a key role. In the first 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 referred to as distant supervision.

Given the absence of a dataset, we also deal with an absence of a baseline our model could be compared to. We evaluate the performance of the proposed model architecture on some well known English datasets to show that it is comparable to the state-of-the-art results.

Thesis Organization

This thesis is split into two parts. Before we dive into the first part, we will provide background information that is relevant to this thesis, such as more details on relationship extraction, related terminology and further motivation. We will briefly introduce the Czech language.

The first part focuses on datasets. First, we will present some existing supervised and unsupervised datasets. Second, we will propose methodology for generating a Czech relationship extraction dataset via distant supervision. Using the methodology we will elaborate on the implementation process and on the generated dataset - the Czech Relationship Extraction Dataset (CERED).

In the second part, we will overview some concepts, that are used in modern natural language processing models. Then we describe the architecture of our model and compare its performance with reported results on popular English relationship extraction datasets. We will also report our results on CERED.

¹<https://www.wikidata.org/wiki/>

²<https://cs.wikipedia.org/wiki/>

1. Background

This thesis is split into two parts that focus on different aspects of relationship extraction in the Czech language. In this section, we provide a glossary of some NLP and relation extraction terms, we elaborate on the task itself, and we include an introduction to the Czech language for non-Czech readers.

1.1 Terminology

Terminology in NLP subtasks is often non-standardized or not exact. We will attempt to introduce the most important concepts for our work as exactly as possible while respecting the terms that seem to be already established.

Relation in this context is an abstraction of a semantic relation, for example, a father relation. A relation is of a type (father) is binary (between the son and the father) and oriented (the father and the son are not interchangeable), and describes the relationship between a subject (the son) and an object (the father). We will use the term relation as an equivalent for its type and the term relationship for an instance of the relation.

Subject and **Object**. The subject is the first argument of a relation, and the object is the second. In the sentence “*Albus Severus is Harry Potters’s son.*” a relation of type SON is captured, the subject is Harry and the object is Albus Severus. The reasoning for this choice of direction is as follows: suppose we are gathering information about Harry, then we would probably have both the information that his son is Albus Severus and that his father is James. So we are gathering information about the subject (Harry Potter), even though in most sentences Harry is likely to be the grammatical object: “*James is Harry’s father.*” We will use the notation `RELATION(subject, object)`: `SON(Harry Potter, Albus Severus Potter)`.

Both the subject and the object can generally be any word or sequence of words that represent concepts that can 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, and products, that can be denoted with a proper name. Named entities can be viewed as instances (e.g., New York City is an instance of a city) of some concept - **class**. Sometimes, numeric data is considered in this category as well (for example, by Named Entity Recognition tools). An **entity** is a named entity whose proper name is unknown or unimportant but still is an instance. (The word book can represent an abstract concept - a class - as well as an entity.)

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

Positive relation mention is a sentence, that captures a relationship: a relation together with a tagged subject and object. We will omit the word positive unless we want to emphasize the fact.

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

obrázek
pro
příklad
v
dalším
odstavci

lepší
vzhled

- OTHER - human annotator would classify a relation, that is not in the inventory.
- NO RELATION - in this case, human annotators should feel an absence of a relationship between the subject and the object.

NO RELATION comes with difficulties. Since there is no semantic relationship 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.

doplnit
příklady,
včetně
vyložene
ne
příkladu

1.2 Relationship Extraction

The relationship extraction task concentrates on the prediction of relationships. In the typical setup, the goal is to predict a relation type based on a sentence with two tagged entities. Variations of this exist, for example, in real-life applications, the goal might be to output a set of relationships based on a longer piece of text (and therefore the extractor would have several sentences mentioning the same entities).

Since the input for a relationship extraction model should contain tagged entities, a pipeline of an entity tagging tool and relationship extraction model would be necessary to perform relationship extraction on real data.

We are aware of two sources of motivation for relationship extraction. First, it can be an alternative to a summarization tool. Therefore it could be used anywhere where people are required to read long texts in a short amount of time. In the second application, the extractor could convert texts into structured data and therefore build a database of relationships. For both applications, an entity linking tool would be beneficial (such a tool disambiguates named entities to a knowledge base) to eliminate the confusion of similarly named entities.

Since many applications are likely to benefit from entity disambiguation, a pipeline of named entity recognizer, entity linker and finally relationship extraction would potentially be useful.

1.3 Czech language

One of the objectives of this thesis is to work with the 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 the 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.

1.3.1 Inflexion

In Czech, nouns, adjectives, pronouns and numerals are declined. The declination expresses (not necessarily unambiguously) one of seven cases and a number (singular or plural). Any inflected word in Czech has grammatical gender. For words, that have natural gender, those two genders align: “žena” (woman) is

singular				
	masculine animate	masculine inanimate	feminine	neuter
nominative	příčný		příčná	příčné
genitive	příčného		příčné	příčného
dative	příčnému		příčné	příčnému
accusative	příčného	příčný	příčnou	příčné
vocative	příčný		příčná	příčné
locative	příčném		příčné	příčném
instrumental	příčným		příčnou	příčným
plural				
	masculine animate	masculine inanimate	feminine	neuter
nominative	příční	příčné		příčná
genitive	příčných			
dative	příčným			
accusative	příčné			příčná
vocative	příční	příčné		příčná
locative	příčných			
instrumental	příčnými			

(a) Declension of “příčný” (diagonal)

	singular	plural
nominative	ulice	ulice
genitive	ulice	ulic
dative	ulici	ulicím
accusative	ulici	ulice
vocative	ulice	ulice
locative	ulici	ulicích
instrumental	ulicí	ulicemi

(b) Declension of “ulice” (street)

Figure 1.1: Examples of czech declension, taken from Wiktionary.

feminine and “muž” (man) is masculine. The inflexion of each declinable word follows a pattern. This all means that a single word can have relatively many different forms.

Verbs are conjugated, the conjugation expresses person, numeral, tense, voice, and mode. Verbs follow one of 14 patterns. An average Czech either finds the theory about Czech verbs and tenses confusing or is even unaware of the existence of the verb patterns. That likely contributes to common use of incorrect forms of verbs even in the official language.

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

We include an example to help grasp these concepts to readers who do not speak any inflexive language. In Figure 1.1 words “příčný” (diagonal) and “ulice” (street) are declined, the noun phrase “Příčná ulice” is the Czech equivalent of

the Diagon Alley. The lexemes sizes are 11 and 7, so combinatorically, the noun phrase could have 77 forms. As we explained, in Czech, there is an agreement in noun phrases, and only 9 different forms of the noun phrase “Příčná ulice” are valid.

1.3.2 Word order

Unlike in English, the sentence structure is relatively relaxed in Czech. The basic sentence structure is of SVO (subject verb object) type, but even when the word order is entirely different, the sentence might still be understandable thanks to the inflected forms of words. At the same time, the word order is not arbitrary; some orderings of words do not form a valid sentence. Those that are valid can carry a different message (completely different, with a different emotion or a different emphasis).

The position of attributes tends to be mostly fixed. Some attributes are prepositive, some postpositive, but most of them stay at the same position (within the noun phrase) unless some enumeration is used. If we return to the “Příčná ulice” example, we are not able to recollect a sentence, where the reversed order is used.

Part I

Datasets

2. Existing datasets

In this chapter, we will overview three 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 the article so that we can compare it later with other datasets.

The authors started by choosing an inventory of semantic relations. They aimed for such a set of relations that is exhaustive (enables the description of relations between any pair of nominals) and mutually exclusive (given the context and the pair of nominals, only one relation should be selectable). Chosen relations with descriptions and examples are listed in Table 5.1.

They decided to accept as relation arguments any noun phrases with common-noun heads not just, for example, named entities mentioning: “Named entities are a specific category of nominal expressions best dealt with using techniques which do not apply to common nouns.” They restricted noun phrases to single words with the exception to lexicalized terms (such as science fiction).

The annotation process consisted of 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. The near-miss sentences were classified as OTHER, or were removed.

The relations inventory contains nine positive relations and one negative. The authors decided to include the directionality into the relation and therefore the inventory size is $9 * 2 + 1$ in total.

The dataset contains 10717 relation mentions. For the original competition, teams were given three training dataset of sizes 1000 (TD1), 2000 (TD2), 4000 (TD3), and 8000 (TD4). Since there was a notable gain TD3 \rightarrow TD4, the authors concluded that even larger dataset might be helpful to increase the performance of models. On the topic, the creators have written:

.. 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.

Table 2.1: S10T8 summary. List of relations, their official descriptions, a random relation mention and both the relative and the absolute count of mentions when the directionality is ignored.

CAUSE-EFFECT	12.4%
An event or object leads to an effect.	(1331)
Example: <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)
Example: <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)
Example: <i>The <u>factory</u>'s products have included flower pots, Finnish rooster-whistles, pans, <u>trays</u>, tea pots, ash trays and air moisturisers.</i>	
CONTENT-CONTAINER	6.8%
An object is physically stored in a delineated area of space.	(732)
Example: <i>This cut blue and white striped cotton <u>dress</u> with red bands on the bodice was in a <u>trunk</u> of vintage Barbie clothing.</i>	
ENTITY-ORIGIN	9.1%
An entity is coming or is derived from an origin (e.g., position or material).	(974)
Example: <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)
Example: <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)
Example: <i>The system as described above has its greatest application in an arrayed <u>configuration</u> of antenna <u>elements</u>.</i>	
MEMBER-COLLECTION	8.6%
A member forms a nonfunctional part of a collection	(923)
Example: <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)
Example: <i>Cieply's <u>story</u> makes a compelling <u>point</u> about modern-day studio economics.</i>	
OTHER	17.4%
	(1864)
Example: <i>The <u>child</u> was carefully wrapped and bound into the <u>cradle</u> by means of a cord.</i>	

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 100 000 examples, which makes it about ten times bigger than S10T8 dataset.

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. ... 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 41 relations with the subject being an organization or a person and a negative relation. Objects are of the 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. The choice of subjects and objects is therefore very different from the S10T8 dataset.

TACRED was designed to be highly unbalanced. 79.5% of mentions represents the NO RELATION relation. This ratio of negative relation should be closer to real-world text and supposedly should help avoid false-positive predictions. However, even if we look only at positive relations, there are vast differences in frequency: the top six relations make up half the dataset and the bottom six less than 2%. In absolute numbers, the least common ORD:DISSOLVED relation has only 33 mentions, and the median is only 286 mentions.

Table 2.2: TACRED summary. List of relations, a random example, and both the relative and the absolute count. The table is restricted to ORG:* relations.

NO_RELATION	79.5%
Example: “ <u>One</u> step at a time , ” said Con Edison spokesman Chris Olert in Sunday editions of The <u>Daily News</u> .	(84491)
ORG:ALTERNATE_NAMES	1.3%
Example: 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 .	(1359)
ORG:CITY_OF_HEADQUARTERS	0.5%
Example: Once completed , the cuts will leave the <u>Irvine</u> , California-based <u>Option One</u> subsidiary with about 1,400 employees .	(573)
ORG:COUNTRY_OF_HEADQUARTERS	0.7%
Example: 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> .	(753)
ORG:DISSOLVED	0.0%
Example: News Corp. sold its satellite television service <u>DirecTV</u> in <u>2008</u> to Liberty Media .	(33)
ORG:FOUNDED	0.2%
Example: New York-based <u>Zirh</u> was founded in <u>1995</u> and makes products using natural oils and extracts .	(166)

ORG:FOUNDED_BY	0.3%
Example: The <u>Jerusalem Foundation</u> , a charity founded by <u>Kollek</u> 40 years ago , said he died of natural causes Tuesday morning .	(268)
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 .	(171)
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 .	(286)
ORG:NUMBER_OF_EMPLOYEES/MEMBERS	0.1%
Example: Established in September 1969 , the <u>organization</u> now has <u>57</u> member states worldwide .	(121)
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 .	(444)
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 .	(125)
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 .	(144)
ORG:STATEORPROVINCE_OF_HEADQUARTERS	0.3%
Example: Romney 's investments are in Novo Nordisk , a Danish company , and <u>Millipore Corp.</u> , based in Billerica , <u>Mass.</u> .	(350)
ORG:SUBSIDIARIES	0.4%
Example: Moody 's said it has growing concerns about the control wielded by Johnson at <u>Fidelity</u> 's parent company , <u>FMR LLC</u> .	(453)
ORG:TOP_MEMBERS/EMPLOYEES	2.6%
Example: <u>Frank Gut</u> , the chief financial officer , previously oversaw the Swiss brokerage operations at <u>Bank Julius Baer</u> in Zurich .	(2770)
ORG:WEBSITE	0.2%
Example: <u>Swiss Bankers Association</u> : <u>http://www.swissbanking.org</u>	(223)

2.3 Riedel NYT dataset

The previous two datasets were obtained through a tedious human labour - human annotators went through texts and manually annotated the data. This process is slow and expensive, which explains the relatively small data volume of the datasets. In this section, we will introduce a dataset that was presented in Riedel et al. [2010], and that was created without the need for any additional manual annotation.

This dataset was generated with the distant supervision approach. This approach is based on aligning structured data (knowledge base) with text, i.e. automatically tagging mentions of the structured data in the text. In distant supervision, we usually expect that if there are two entity mentions in a sentence that are related, then the sentence expresses their relationship. The authors acknowledge that this assumption is often violated and they propose a methodology that attempts to predict whether the assumption is violated in a sentence. Using this methodology, they generated a dataset from the The New York Times Annotated

Table 2.3: Random 10 relations from the Riedel NYT dataset. For each relation the name, the relative and the absolute number of mentions and a random mention is shown.

/LOCATION/LOCATION/CONTAINS	8.6%
Example: <i>But John Traugott , 68 , a hospital chaplain in <u>Rockaway Park</u> , <u>Queens</u> hinted at some of Chinatown 's problems .</i>	(58625)
/PEOPLE/PERSON/NATIONALITY	1.5%
Example: <i>We were unable to reach agreement , " Foreign Minister <u>Frank-Walter Steinmeier</u> of <u>Germany</u> announced tersely to reporters .</i>	(10464)
/PEOPLE/PERSON/PLACE_LIVED	1.2%
Example: <i>Camane -LRB- Emily Blunt -RRB- runs around <u>Rome</u> putting up signs declaring that Octavius is <u>Julius Caesar</u> 's rightful heir .</i>	(8275)
/BUSINESS/COMPANY/FOUNDERS	0.1%
Example: <i><u>Nick Grouf</u> , president and chief executive at <u>Spot Runner</u> in Los Angeles , is scheduled to announce the investments today .</i>	(999)
/PEOPLE/DECEASED_PERSON/PLACE_OF_DEATH	0.3%
Example: <i><u>Marie Antoinette</u> finally did arrive in <u>Paris</u> , at Christian Dior , where the designer John Galliano proclaimed her his platinum muse .</i>	(2190)
/BUSINESS/PERSON/COMPANY	0.9%
Example: <i><u>Christopher Bailey</u> , with his light-handed take on <u>Burberry</u> 's heritage , could add some military backbone and useful outerwear .</i>	(6455)
/LOCATION/US_COUNTY/COUNTY_SEAT	0.0%
Example: <i>Mr. Perhacs was taken to <u>Jersey City</u> Medical Center , where he died , said Edward J. De Fazio , the <u>Hudson County</u> prosecutor .</i>	(125)
/BUSINESS/COMPANY/PLACE_FOUNDED	0.1%
Example: <i>Next month in <u>Paris</u> , Ms. Tilbury will direct makeup at the spring fashion shows of Lanvin , <u>Chloé</u> and Alexander McQueen .</i>	(537)
/PEOPLE/PERSON/PLACE_OF_BIRTH	0.5%
Example: <i><u>Preston Robert Tisch</u> was born in the Bensonhurst section of <u>Brooklyn</u> on April 29 , 1926 , to parents who came from Russia .</i>	(3603)
/FILM/FILM/FEATURED_FILM_LOCATIONS	0.0%
Example: <i><u>Half Nelson</u> , " a new independent film about an idealistic young <u>Brooklyn</u> teacher , takes this claim at face value . "</i>	(19)

Corpus and Freebase. We will refer to this dataset as Riedel NYT.

The dataset is divided into two parts, each of them containing a test set and a train set. We will use both train sets for training and the "heldout" test set for evaluation in the second part of this thesis. We were unable to find any documentation that would describe the expected use of each part, nor did we find an official instruction concerning the metric for reporting results.

Relation inventory for the part of the dataset that we used contains 58 relations. The best represented is the NA relation with over 80%. The representation of relations varies between the train and test set. For example, two relations are present only in the test set.

možná
citace?
nebo
není
třeba?

3. CERED

In the previous chapter, we concluded that creating relationship extraction manually is expensive. We introduced Riedel NYT dataset that shows, that automatic generation of relationship extraction datasets is possible. In this chapter we will describe our process of generating **Czech Relationship Extraction Dataset** (CERED). We will discuss various decisions that we made during this process and their impacts.

3.1 Overview

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

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

Next we analyze how we will find mentions of Wikidata relations in Czech Wikipedia. We sketch out dataflow diagrams and we think about all the different complex aspects of this task.

We continue by choosing technologies that we use. Aware of the volume and other characteristics of chosen data, we choose Python as the main programming language, Spark as a way to speed up the computations and MorphoDita to deal with the specifics of the Czech language.

The questions and options that rose from the analysis get at least partially answered and decided during implementation. We tested different configurations and went through the data to determine what will work best. As a result, we generated CERED0.

The generated dataset have a huge relation inventory and other properties that would be impractical for training. We describe variations of the generated dataset CERED1-4, that were designed to be more practical.

3.2 Data sources

To be able to perform distant supervision we need to find suitable data - Czech text corpus and a knowledge base (Figure 3.1). In the first subsection, we will explain the requirements and constraints we have on such data and present our options. In the next two subsections, we will provide more information on the chosen ones.

3.2.1 Constraints and requirements

The main constraint is quite straightforward, there has to be a nontrivial shared set of entities and relations mentioned in the text and stored in the knowledge base. We expect fact-based text to be more suitable than fiction literature.

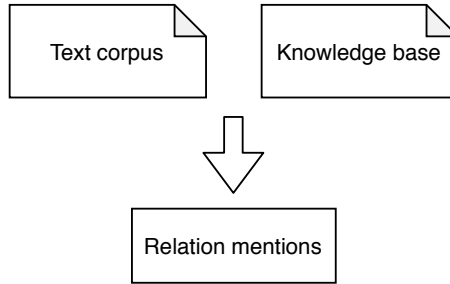


Figure 3.1: Distant supervision diagram

Therefore we will prefer encyclopedic or journalistic genre. One option is to focus on some subset of Czech National Corpus¹, 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.

To the best of our knowledge, our options for knowledge base are limited to Wikidata or Google Knowledge graph².

We decided to use Czech Wikipedia and Wikidata, mostly because the intersection of information expressed in text data and in structured data seems promising because they are build on each other. Another advantage could be the multilingualism of Wikimedia projects, and therefore the transferability of this work will be higher. Moreover, the Wikimedia projects are downloadable, which makes them easier to work with.

3.2.2 Czech Wikipedia

Wikipedia is a multilingual online encyclopedia created and maintained as an open collaboration project by a community of volunteers as defined in Wikipedia contributors [2020]. From our 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.

3.2.3 Wikidata

Wikidata is a knowledge base which acts as a central storage of the structured data of 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 the multilingual versions of Wikipedia articles on the same topic - on the same Wikidata item. Each item is assigned an identifier, prefix Q and a unique number, referred to as **QID**. A

¹<https://www.korpus.cz/>

²<https://developers.google.com/knowledge-graph>

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

label together with a description of an item should serve as a human readable identifier. Labels, descriptions and 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 a unique number, and a data type for a value it can be paired with (such as an item, string, url, number or media file).

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

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

Wikidata project contains over 80 000 000 items, which raises requirements on technological resources so that we can 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 parts of the pipeline we mention in the Relationship Extraction section. 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 entities only.

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

There are several aspects that we need to think through. We introduce them in the following list for better orientation. A subsection in this section is dedicated to each of those aspects.

- Dataflow - we chose Wikidata and Czech Wikipedia, but we did not discuss how to connect them and what exactly should be the outcome so that we can proceed to locating mentions.
- Entity Matching - suppose we collected a piece of text together with a set of entities that could be mentioned in the text. The process of entity matching attempts to mark words in the text that mention an entity.
- Wikilink Mentions - Wikilinks are (mostly) human labeled entity mentions. Utilizing them is the closest we can get to a supervised dataset without actually supervising the dataset.

hezčí
formátování
wiki-
itemm

formatování

mám
zade-fi-
no-
vanou
jinou
ty-
pografii
pro vz-
tahy,
použit

úplně
odstranit
ten
list?

- Relation Matching - entities are matched in the Wikipedia texts, relationships are extracted from the Wikidata dump, we use distant supervision assumption to locate relation mentions in the texts.
- Relation Inventory - we generated CERED, but the relation inventory is overly diverse. Moreover, the dataset is extremely unbalanced - the number of mentions per relation varies. And no “no relation” relation is obtained.
- Result Evaluation - every time we generate CERED during development, we need to evaluate its quality. We propose methods for this evaluation.

3.3.1 Dataflow

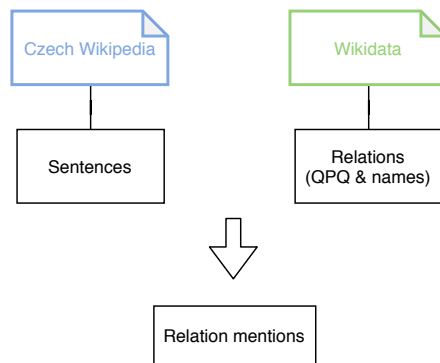
We start the whole process of creating our dataset with two files. The first file is a Czech Wikipedia dump. It is a collection of articles where each article has its title, id and text. And the other file is a Wikidata dump.

The simplest way of processing those files would be to process them separately and thus obtaining sentences on one side and relationships (a relation type with two items) on the other. This approach comes with a clear disadvantage. We would lose any additional information about the sentences that could be potentially useful (for example article title might be helpful to determine which items are mentioned in the sentence).

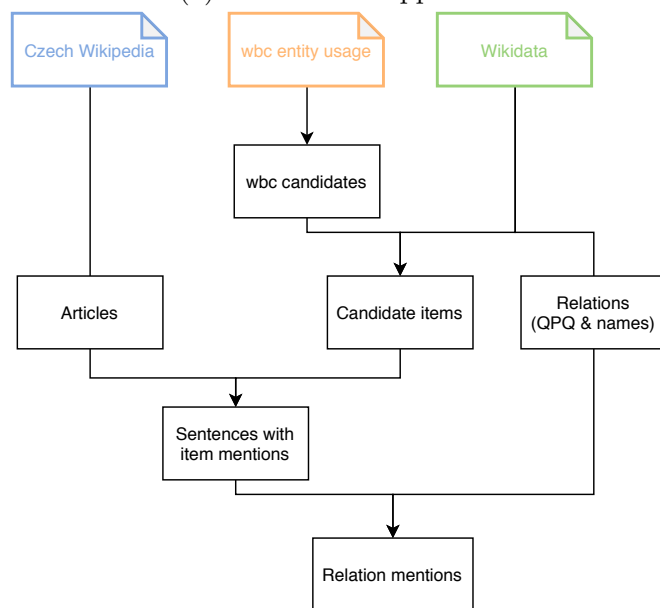
To solve this we could precompute something for each article and attach it to each sentence (article title, all Wikilinks in the article, etc.) risking a massive increase in required capacity to work with such data. On a similar note, we would probably process Wikidata to store item names (labels and aliases) for each relationship, 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. We will refer to this processed version of Wikidata as **custom Wikidata**. An item will be kept only if it has a Czech name and we will significantly reduce its statements: we will keep the title of its Czech Wikipedia article and create a list of (QID,PID,QID) triples - **QPQ**, representing statements that contain information about relations between this item and the other items. This way, we have all the necessary information - article title to be able to connect an article to an item, names for each item to be able to find mentions of items and finally QPQ triples to connect relations and sentences. Moreover, custom Wikidata size is closer to traditional RAM size. Therefore we could for example load item names into memory, which will come in handy during the implementation.

One approach to finding item mentions in text could be called uninformed (Figure 3.2a). We could assume that any item can be mentioned in any sentence. This approach seems to have two 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 see as problematic might be children named after their parents. In this case, not only that the entities might get confused but also if we then assign the relation, we might easily confuse a sentence mentioning a spouse relation for a parent relation, which, unfortunately, is extremely challenging to solve.



(a) Uninformed approach



(b) Informed approach

On the other side, we can use the extra information that Wikimedia projects provide and opt for a more informed approach. A diagram of this approach is captured in 3.2b 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 seems 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 an item, then 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, let us 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 specific country, all countries would be second level relatives and thus a candidate item. Since countries tend to be of a certain type (kingdom, republic, state etc.) there might be simply the type or some other more general name amongst their names (*United States of America Q30* are also known as America or United States) and more countries might share this name.

So far we mostly discussed the advantages of the 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 the following condition holds: there are two entity mentions in one sentence and there exists a QPQ that connects them. It is debatable whether we need an informed approach to increase relation mention precision. The improbability that this condition will be fulfilled for false-positive item mentions might in fact be sufficient.

One more way to locate item mentions is through **Wikilinks**. A Wikilink links a page to another page within the 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 are to be examined and if we will find these names useful, they can be added to the item names we use.

podívat se do nakreslených obrázků, jestli tam není něco použitelného sem

3.3.2 Entity matching

We have text on one side, gathered candidate items on the other and our goal is to find occurrences of these items in the text. We call this process **entity matching** and each found occurrence is an **entity mention**.

No matter how the matching will be done, it seems always beneficial to start the process with some text preprocessing. Quite a lot of changes need to happen even if some seem like little details. We will separate this preprocessing into a wiki specific part, lexical analysis and the last part is devoted to lexical analysis

on standalone noun phrases.

When we eventually proceed to entity matching, there is a wide spectrum of complexity we might aim for. We are not trying to create a strong sophisticated tool for entity recognition and linking. We will describe some of those complexity tears and choose the right method for our use case.

Wikipedia parsing

Wikipedia parsing starts with an article in Wikitext and produces human-readable plain text - **clean text**. We should keep track of positions of Wikilinks from the Wikitext in the clean text.

Wikipedia is written in Wikitext (Wiki markup, Wikicode). This markup provides all usual functionalities such as determining the layout or fonts and enables commonly-used concepts like lists, links, media file insertion, or tables, and some more wiki specific concepts like infoboxes.

We plan on using one or even a combination of existing Wikitext parsers since each of them provides different functions.⁴ Therefore the parsing itself is not too troublesome.

One problem that needs to be addressed is what should we consider to be a valid text. For example, it is not clear how to work with tables. From one point of view, if we convert a table into an unstructured text, the text will not be a regular text in terms of sentence structure. From a different point of view, an unstructured text obtained by converting a table still contains information, that human readers will likely decode. Moreover, tables and other structured data tend to contain a lot of information. This will likely cause problems because we want to concentrate on sentence-like data, not just for example tuples of data that for example a table of athletes might provide (tuples of persons and countries). This kind of data might significantly damage the quality of CERED.

The elimination of all Wikipedia content, that is too structured or generally, not enough sentence-like, but at the same keeping as much as possible will be addressed later in Wikitext Parsing. That way we can see the consequences of the eliminated and kept content.

Lexical Analysis

For our purpose keeping the text in long sequences of characters (such as an article) is not the best format. We need to parse those sequences into smaller tokens, such as words and sentences. A tool that addresses such tasks is usually called a **tokenizer**. Tokenization is a process that aims to split text (sequence of characters) into separate tokens. Token is a term that generalizes the term word and often words in text are the same as tokens. In english, “aren’t”, in Czech “mohu-li” are likely considered one word but two tokens. On the contrary “M*A*S*H” or “email@email.com” might not be considered a word, but should be considered exactly one token each. Naive tokenizer might just simply split on non-alphanumeric characters. But if the tokenizer is supposed to actually perform well and recognize sentences, a more sophisticated tool is needed. Since we wish to work with the language even further, we might want to be able to tell

⁴https://www.mediawiki.org/wiki/Alternative_parsers

Lord Voldemort (Tom Rojvol Raddle, Voldemort, Pán zla), jeho jména + obrázky jak by se matchovalo přímo stringem a pak lexikálně, ale založeno už na wikidatech (co není ve wikidatech se nenajde..) jsou ve výrobě

Figure 3.3: First paragraph from Czech Wikipedia page about Lord Voldemort Q176132, also known as Tom Rojvol Raddle, Voldemort, Pán zla (Dark Lord)

the non-inflexed form of a token, moreover we might want to assign some unique identifier to such form in case it has some homographs (words that are spelled the same way). Such identifier is called a **lemma**. The collection of tokens that share the same lemma is called a **lexeme**. We will outsource handling text to a Czech tokenizer called MorphoDiTa, that achieves state-of-the-art results for the Czech language. Using such a tool, we can convert clean text into sentences made up of tokens and we even obtain the lemma and lexeme of each token.

Lexical Analysis on Names

Tokenizers (and lemmatizers) are usually trained to perform well on sentences and might be inaccurate on noun phrases when they stand alone, as entity names do. If we were determined to tokenize them, a simple trick like constructing a sentence with the name in it and tokenizing this sentence can partially solve this problem. Such a sentence that would be grammatically correct and not semantically terrible could be something like “This is /name/”, but realistically, this sentence was quite likely not at all common in the training process of MorphoDiTa. Some foreign words can be erroneous as well and keeping their original form might be the only easy way around it.

Matching methods

Entity matching can be done with various degrees of sophistication. We provide a short overview of those degrees (1 – 4). 1 and 2 do not require any knowledge of the language they work with, 3 and 4 are language dependant. For simplicity, we will assume that we are only looking for mentions in one sentence at a time unless written otherwise. We also include an example of such matchings to demonstrate how successful we are likely to be.

1 String equality. Definitely the easiest method of entity matching. This method is based on a simple substring check which is later extended with additional functionality. In more detail, for each entity, we have multiple name variants and for each of those names, we check whether the name is a substring of the sentence.

We still need to work with letter cases. Named entities should have fixed letter cases and no additional processing is needed in most cases. In other cases, an established name for a named entity might be written with the lowercased first letter (Weasley family (Q716534) has Czech names 'Weasleyovi' /the Weasleys/ but also rodina Weasleyových /Weasley family/, but there are examples where those different names are completely different (Elizabeth II and Queen of England would be translated to Alžběta II. and královna Anglie). If we consider entities, that are commonly written with lower case (), the sentence now needs to be preprocessed so that for example the first letter is not capital. Moreover, there is

značka pro překlad v caption

Formátování nadpisů

český

příklady jako škola, tužka

no guarantee that common names will be lowercased in Wikidata. To conclude, nearly nothing can be assumed about the case of letters, and therefore one of the following solutions needs to be implemented: everything can be converted to one chosen letter case or some more sophisticated attempts at predicting, which words can have more version in terms of letter cases.

Another problem that we may encounter is how to properly handle spaces. We will list some troublesome examples and accept the fact that not everything can be done perfectly. J. K. Rowling has J.K.Rownling as one of Wikidata names, confirming that both versions might appear in written text, but not all entities with similar name type have all space-variants listed in names. We assume that spacing around the “-” character might vary.

It is also not clear if word order in entity names is fixed (or at least almost always fixed). Even the simple reversion in name, that is sometimes used, will affect the performance of this method (J. K. Rowling and Rowling, J. K.). Even cases where the name is divided by for example apposition into two separate parts might exist.

The greatest weakness of this method is its inability to recognize entities if their name is inclined. To emphasize how many words are not in the same form as their lemma in Czech text, we colored them in the sample text. We elaborated on Czech language in section 1.3, but just for simplicity - in English the verb “to be” has many different forms (am, are, were, was, would and so on), all nouns and verbs in Czech behave like this, quite often with many more forms.

vybarvit
a dát
link

2 String similarity (approximate string matching). String similarity is still based on simple string manipulation, no vocabulary or other language knowledge is necessary. The goal is to find entity mention, even if its name is a little altered in the sentence. This alternation can include all of the issues listed for the previous method - cases of letters, spacing, word order, and word forms, but even better, it might help in cases, that we did not anticipate.

There are many metrics describing string similarity. Some could cope better with word order issues, some with word forms, some with spacing. We will not test all of them for our usecase, but still find it useful to mention them since in other than the Czech language, some might work well.

Edit distance. First category of string similarity metric is based on edit distance:

Levenshtein distance is the minimum number of edits (additions, deletions, and substitutions of a character) to get from one string to the other. As a metric the ratio of Levenshtein distance and of the sum of the lengths of the strings can be used. This metric, unsurprisingly, deals well with mentions that are close to the names when it comes to the amount of edits needed, so mentions differentiating in a word form, different spacing, or letter casing will could be considered a match.

Damerau–Levenshtein distance is very similar to the previous, but a transposition of two adjacent characters is considered an edit. We might argue that some Czech words tend to transpose the last characters in different word forms and thus this metric could work better for those forms, but there might be a higher risk of false positives.

One more thing to mention about edit distances is that they count the distance of two string, in our case of a name and of a substring of a sentence, because we do not expect an entire sentence to be entity mention. Therefore we need to

decide on a logic for choosing substrings of a sentence to count the distance on, without diving too deep into this, the time complexity (even though both the sentence and the name length is relatively small) can be high if we consider the amount of data we need to process.

Token based. Another category is based on tokens. For those metrics, we can either use the tokenized sentences (by an actual tokenizer) and try to tokenize names, so that the format matches, or we can use naive tokenization like removing non-alphanumeric character and splitting on spaces.

The tokenization converts both the sentence string and the name string into a set of tokens (S, N respectively), so metrics that work with sets can be utilized: Intersection over union is computed as $|S \cap N|/|S \cup N|$. Any other set similarity measure can be used.

Token based metrics - due to their set nature - ignore the order of tokens and therefore could solve issues with mentions in which the word order is not the same as in the name. On the other side, an increase in false positives is to be expected and some additional postprocessing is needed to determine which token in the sentence should be considered a mention if the token was used in the sentence multiple times. We also feel obligated to mention that once again, we are not trying to find the similarity of a sentence and a name, but of a substring of a sentence and a name. This leads to the idea of removing all tokens, that are in a sentence and would worsen the metric and in result modifying the formula for intersection over union to $|S \cap N|/|N|$.

3 Sequence based. Just to be a bit more comprehensive we include another type of metrics - sequence based, even though we doubt that it is the best approach for entity matching. They ignore words as wholes and we do not see any advantage of those metrics for our use case.

Ratcliff-Obershelp similarity finds the longest common substring that is longer than some limit and recursively does so for the non-common parts of strings. The result is based on the ratio of (double the) length of common parts and overall length.

Bigram (or n-gram) intersection over union which converts both strings into a set of n-grams (n adjecant characters) and performs intersection over union. This time reducing a sentence into a substring is not that straight forward and would require additional attention.

Morphological analysis. Moving on from methods that are mostly unaware of the language they work with, we will finally use the morphological analysis we mentioned earlier.

With lemmas of both the sentence and the name, we can use any metric from the previous subsection on string similarity (joining the lemmas on the space char if the metric expect only two input strings).

If we decide to keep the names in their original form (tokenization on them can be error prone, as we already explained), we can try to use the correct form of the tokens in the sentence. For each lemma, we get a set of all its possible forms - a lexeme, now we can modify the previous metrics to work with lexemes instead of tokens.

4 Advanced concepts.

A proper entity matching (either in named entity recognition or entity linking)

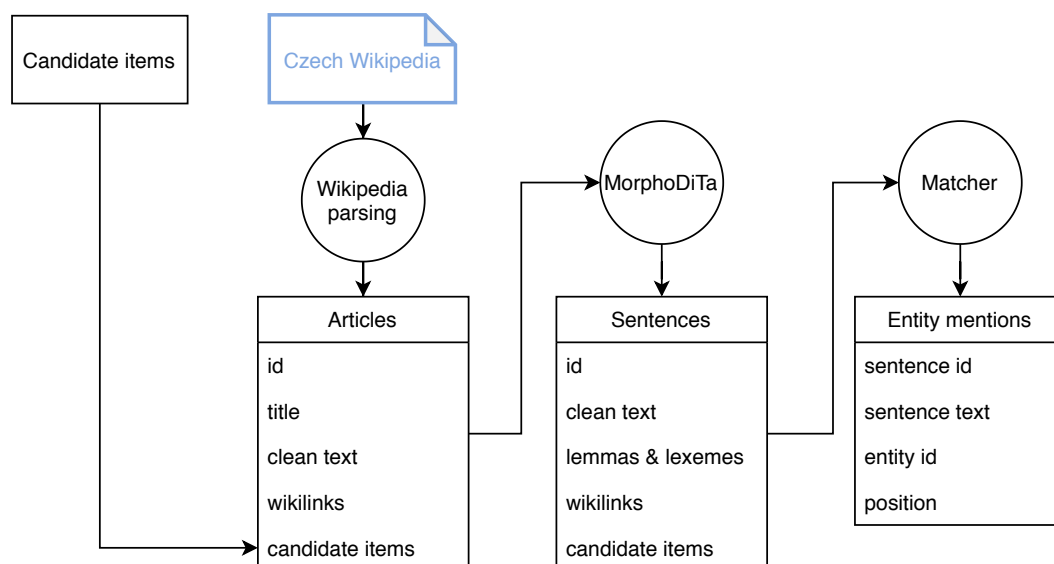


Figure 3.4: Matching with morphological analysis

might be expected to recognize entity mention even if the entity is not mentioned explicitly by its name. Pronouns should be assigned an entity they represent (if they do) and other nouns as well. In languages like Czech where the subject of a sentence is often omitted the entity mention is even less obvious but still present. Since the topic of this thesis is not entity matching, we will not debate techniques to achieve this level of matching neither will we implement them.

(vyřešit nadpisy)

Conclusion

After looking at the results, we used a simple metric but we still find it useful to keep this summarization of different string metrics as part of this thesis. A detailed description of the matching method CERED was generated with is written in subsection 3.5.3.

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

3.3.3 Wikilink Mentions

As we already mentioned, from our point of view Wikilinks are entity mentions created by Wikipedia editors. The text part of the link can in theory be anything providing us with some more advanced examples of entity linking, that our matching methods cannot perform.

V pátém díle McGonagallová říká, že předmět vyučuje již 39 let

We want to enable the users of CERED to distinguish relation mentions that were created based on two Wikilinks from all others. Since this data is not fully supervised and the word supervised is overused (semi-supervised, distant-supervised, etc.) we decided to call it **silver**, because they are not of the optimal quality that is usually labeled gold, but they are the best we can get.

3.3.4 Relation Matching

If a sentence contains two entity mentions that are related, chances are that the sentence in fact does express their relationship and thus is a relation mention.

	NYT	Wikipedia
NATIONALITY	38%	20%
PLACE OF BIRTH	35%	20%
CONTAINS	20%	10%

Table 3.1: Percentage of times a related pair of entities is mentioned in the same sentence, but where the sentence does not express the corresponding relation. Taken from [Riedel et al., 2010].

This concept is called **distant supervision assumption** and can be also formulated in the following way: If two entities participate in a relation, all sentences that mention these two entities express that relation. This assumption is commonly used, even though it is clearly not correct, because it is easy to use it. To tell how often this assumption is violated is labor-intensive, luckily research has been done on this topic. In Riedel et al. [2010], the distant supervision assumption is compared to an **express-at-least-once assumption** which states that if two entities participate in a relation, at least one sentence that mentions these two entities might express that relation.

They sampled 600 relation mentions from two corpora, both created by distant supervision on Freebase (knowledge base commonly used before Wikidata took over) and two text corpora - Wikipedia articles and the New York Times corpus. These 600 samples represented three different relation types (nationality, place of birth, and contains) and were sampled so that there were 100 samples of each type in each corpus. We include their results in table XX. They concluded that Wikipedia is a very specific type of text corpora, because articles are centered around entities. We believe that the reasoning can be extended with the fact, that freebase contained information from Wikipedia infoboxes, and those infoboxes were created based on the textual information.

For the authors the results signalized that a more sophisticated tool is needed instead of relying on the distant supervision assumption. We acknowledge that such a tool is needed but at the same time we believe that in our case, where we create CERED based on Wikipedia and Wikidata, the precision they estimated is sufficient. We also assume that Wikidata project is more suitable for this task than was Freebase.

We want to mention that we build CERED to easily fit into the modern deep learning models and to be as simple as possible. Therefore, the main piece of text we use is a sentence, it might seem intuitive, but it has one downfall. If the relationship extractor was to be used on a real text not to determine where some relations are mentioned but to provide a summary of relations expressed by the text as a whole, some information will be lost.

3.3.5 Relation Inventory

In the Existing datasets chapter some examples of Relationship classification datasets were introduced. The creators of those datasets claim that in the creation process they first decided on the relation inventory (relation types). Creating the relation inventory seems to be the straight forward and rational approach and we wanted to create such inventory before actually implementing the CERED

table,
opsat,
do cap-
tion
citace

najít
článek
nebo
někam
přesunout?
<https://www.aclweb.org/anthology/2019.acl-lr.1074.pdf>

generator. We stumbled upon the following issues.

Issue 1

Wikidata relation inventory (properties in Wikidata terminology) is an order of magnitude larger compared to the traditional relationship extraction datasets and handpicking our inventory is overwhelming. We even considered reducing the size of this inventory by creating our own relations that would combine Wikidata relations (parent would be the combination of mother and father relations).

By restricting the Wikidata inventory to just some properties we reduce the size of CERED and if we choose the inventory without considering the number of mentions per relationship, the best-represented relations could be omitted.

Issue 2

Knowledge bases, in general, do not contain negative relations (such relations that could be easily mapped to the NO RELATION or OTHER RELATION), but for relationship extraction negative mentions are essential. If we generate mentions using all properties we can later decide which relationships will be in the inventory and the rest of them relabel to OTHER RELATION. If we were to assign all tuples of entity mentions that share a sentence and are not related as NO RELATION, we could increase the noise in CERED because not all relationships are in Wikidata and therefore some of the NO RELATION mentions could in fact be a positive mention. The ratio of negative and positive mentions in the two bigger datasets we introduced in the Existing datasets chapter were approximately 80%.

While curating the inventory, we should keep in mind that we are not just choosing the relations but also their representations and we need to attempt to fulfill the three following requirements to the best of our abilities:

- Each relation needs to be represented enough.
- The more balanced relation representation sizes the better.
- There should be enough negative mentions and their negativity should be assured.

reformulovat

Result Evaluation

The most challenging aspect of working with Czech Wikipedia and Wikidata is their size and diversity. To the best of our knowledge there is no strictly followed guideline when it comes to editing either the articles or item information. Just converting Wikipedia dump to clean text will be challenging due to user defined templates and other constructs we might be unaware of. On the other side names in Wikidata can be too general (like someone's first name) and create false entity mentions.

We prepared several metrics and methods to measure the quality of the implemented generator and we will list them in the table 3.5. Unfortunately going through the article and seeing what did and did not get recognized as a mention is still the most powerful method, therefore we decided to develop a simple app that can prettify the output of the generator and makes the process of looking at the results less painful. tabulka není hotová a nedává smysl, přepsat.

přiliš neformální?

- FA The overall count of found mentions - to some extent the greater the better - at least in the beginning with naive matching methods. With more sophisticated matching, this number should increase. On the other side, this measure will decrease with more higher precision.
- FA If the distribution of the amount of found mentions (both the entity mentions before relation matching and the relation mentions) over some domain is peculiar (contains abnormalities such unexpected peaks) the quality of not only those mentions might be lower. The domain can be anything from the following or even multiple of them at the same time: sentence, article, relation, entity, sentence number (the order of the sentence in an article), entity pair.
- NA Checking the anomalies detected by the previous methods.
- NA Checking an article and realizing what did and did not get matched.
- FA Some articles could be labeled by hand and test could be created.

Figure 3.5: Quality evaluation methods and the potential for automatization (Fully Automatic, Not-Automatic)

Viewer

Začít tím, co je to viewer, že je třeba, aby prohlížení dat nebylo tolik bolestivé, zmíit, co všechno do něj implementujeme a že sice není použitelný pro ostatní, ale doporučilibychom každému ať si něco tkaového udělá. The purpose of this viewer is to present the results in a more graphical way to fasten and pleasant the process of checking them, which will be done often and will take a significant amount of time.

We want the viewer to be able to show the statistics about the generated data, that were collected during the generation, the configuration that was used to generate the dataset, and mainly to show an article with the found mentions (entity and relation).

Streamlit made implementing such viewer quite easy and we used spaCy's visualizations.

3.4 Used Technologies

We chose Python to be our main programming language. To be able to work faster with a bigger volume of data, we wanted to use a CPU cluster, which leads to Spark. To top it, we will use MorphoDiTa to work with the Czech language. We implemented a simple Streamlit app we used to comfortably view the results of our Spark queries.

In this section we will briefly introduce these technologies.

co je viewer

footnote

přestěhovat
někam
dopřít
tento
odstavec

screen
view-
ertu k
imple-
mentaci

3.4.1 Python

Python is probably the most popular programming language in the ML community. It is a high-level language with a wide range of libraries. Libraries as NumPy, Pandas and Spark enable fast and accesible computation. Tensorflow, scikit and PyTorch allow users to focus mostly on data and ideas in machine learning. Less known libraries help us with Wikipedia parsing (wikitextparser, mwparserfromhell) or easy-to-create web apps (streamlit).

3.4.2 Spark

Apache Spark framework provides query like API and runs on clusters. This way parallel computation can be implemented without actually implementing any parallelism. Therefore, Spark can boost the speed of computation as well as the available memory for the computation.

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 the Czech language. MorphoDiTa has a python package via which we can do all standard operation such as tokenization (splitting text into sentences and words) and lemmatization (disambiguation of the inflected form of a word).

3.4.4 Streamlit

Streamlit is a framework for creating simple web apps. It has minimal and practical API designed for users from the machine learning community. We use this library to create a more pleseant way of viewing the generated dataset.

3.5 Implementation

In this section we build on the analysis. In the ?? we described different approaches we could choose to solve different aspects of the generation of the dataset. Now, we will detail on the approaches we chose.

3.5.1 Wikidata Preprocessing

This module aims to load the wikidata dump file and output three lists. The first list contains QPQ triples, triples of ids representing a relationship (Q173998 P3373 Q187923 for example). The second list provides us with names for the ids mentioned in the first list (Q187923 Ginny Weasley, P3373 sibling). The third one provides the mapping of entity id to the title of the corresponding article if such exists (příklad). Those three lists will serve as the source of structured data.

The first step of our preprocessing is to filter wikidata to remove entities, that we will not use. We require each entity to have at least one Czech name (alias or label), otherwise, there would be no good way to find mentions of that entity

later in the entity matching step. CERED creation does not require the Czech names of relations so we will keep all of them.

Apart from the filtering, we should consider whether we might benefit from keeping more information about a relationship (meaning the instantiation of a relation) than just the ids. Wikidata relationships often contain additional information specific for the relation type. For example, some relationships are valid only for some period of time. The minister of magic changes every few years ("position held" relationship between Cornelius Fudge and the Ministry of Magic has following additional information attached to it: start time 1990, end time 1996, replaces Millicent Bagnold, replaced by Rufus Scrimgeour). In our use case, as we plan to limit text analysis to the syntactic level, such information is not beneficial.

The second step addresses removing duplicate relationships that differ only in the additional information tied to the relationship (CBS received many Peabody Awards for example). Such duplicates where the entire QPQ is the same would not be useful in the future and might require special attention in the relation matching step. We will remove even relationships that differ only in the "P" part. If we kept them, in the relation matching we would either create multiple relation mentions (sentence with two tagged entities and the label of their relationship) that only differ in the relationship. CERED is supposed to be a dataset on which it is possible to train a model in a single label classification task. Such mentions would result in the same inputs (sentence with two tagged entities) and different outputs (the relationship label). We could change the format of our dataset to have a set of relations for each mention, but such data is confusing, non-traditional and overall not what we aim for in this thesis.

After these steps, the first list contains approximately 2 million QPQ triples.

3.5.2 Wikitext Parsing

This module aims to parse wikitext (Wikipedia markup language) from the Czech Wikipedia dump into a clean text with attached information about wikilinks in the original markup.

As we already explained in XX, wikitext contains a lot more than fully unstructured data. Different kinds of infoboxes, tables or lists are contained within the sentence-like text. Some of these elements are implemented using the so-called template syntax. Therefore, it would be tempting to simply remove all the text that is contained in a template. The problem is that not all templates are bad. For example, we may use a template to divide the text into two columns containing valid sentences. Therefore, discarding all such data seems unnecessarily harsh.

When developing the methodology for wikitext parsing, there was not a "gold" to compare it with. The only means of evaluation we had at the time was repeatedly going through a small set of articles and trying to discard unnecessary data. We tailored the rules for wikitext parsing to these articles in such a way that only sentence-like parts remained.

Once we implemented the whole CERED generator and were able to see the relation mentions, we realized that the previous method of evaluation was not good enough. Therefore, we adopted a new one, as described in XX (nedopsaná

sekce v analýza - result evaluation). We looked at the different histograms and investigated the abnormalities. For example, a lot of sports articles report results of a match (tournament, event?) and these are often stored in custom tables that were not filtered by the rules from the previous paragraph. Moreover, these tables oftentimes contain information about the nationality of the players, resulting in a huge amount of matched entities and relations.

do
závorky
dát
jméno
metodiky

Based on the analysis of all the available data, we decided not to include the following content in the clean text:

- HTML tags within wikitext
- headings
- tables
- lists
- templates matching the following patterns: `obsazení*`, `sloupce*`, `seznam*`, `příbuzenstvo*`, `*předkové*`, `*box*`, `*locmap*`, `*tabulka*`. (In English: `cast*`, `columns*`, `lists*`, `relatives*`, `*ancestors*`, `*box*`, `*locmap*`, `*table*`)
- wikilinks to categories and files

One more technical issue we encountered was correctly assigning spans to wikilinks, i.e. where the link starts and ends in the text. We can demonstrate the problem on the following sentence: “The main [[story arc]] concerns Harry’s struggle against [[Lord Voldemort]], a dark wizard who intends to become immortal, overthrow the wizard governing body known as the [[Ministry of Magic]] and subjugate all wizards and [[Muggle]]s (non-magical people).” The correct span for Muggles should contain the trailing s even though it is not part of the wikilink itself. In Czech such trailing characters are common. The set of chars that seem to end wikilinks written in such forms are “`,.\n`”.

typografie
by
mohla
být
lepší

One more thing we mentioned in the analysis about wikitext is the possible boost of performance if the text-part of a wikilink was added to the set of names for the given entity. We exported such names, kept only those that were not already added to wikidata, and read through many of them. This process is time-consuming because one often has to actually look up the entity to know whether a given name is sensible. Even though we do not have any data about the proportion of good and bad names, the overall impression was clearly leaning towards not using such data. The two main reasons were that commonly the name was actually a class name, not instance name (like school linking to Hogwarts). Cases, where the name was just a pronoun, were less frequent but would likely cause much trouble later on.

vynechat?
nevynechat?

3.5.3 Entity Matching

We discussed in great detail the pros and cons of different entity matching methods, implying that the more complex the matching method, the better. We work with a single language and tools for lexical analysis are available and reliable. Therefore implementing language-independent matching methods (string similarity for example) is not beneficial.

We load the entity names in a slightly transformed form. We lower the case and add spaces around every dot character. We used lexical analyzer to split text to sentences and to obtain features from sentences (tokens, lemmas and lexemes). An entity name (sequence of k strings - tokens - separated by a space) is matched in a sentence if (i) the entity is a candidate entity for the sentence and (ii) there is a sequence of k consecutive tokens in the sentence such that each token in the name is a member of the lexeme of the corresponding token in the sentence.

We intended to allow a less strict word order, but we were unable to justify such a choice. After reading several articles we did not find any entity mentions that would be newly matched. This might imply that even though word order is relatively free in Czech, noun phrases tend to keep their word order. The other explanation is based more on the fact that a human reader is more likely to recognize entity mention if it is in the standard word order. Looking for entity mention is again more time consuming than it seems, because we do not know, what is a wikidata entity, therefore a lot of looking up is involved and we believe that often we would not have thought that something is a wikidata entity even though it is. Such “improvement” could still raise the number of entities matched but probably would lower the precision.

We considered allowing one special case. Most articles are based on one entity, therefore we expect many sentences to mention this entity. Often the entity is mentioned either by a pronoun (pronouns that express the subject are typically omitted in Czech), or by part of its full name. Přijali ho do školy čar a kouzel v Bradavicích. Když se Harry vydal do Bradavic, našel nového kamaráda Rona Weasleyho. .. Pak už jel “domů” a musel přežít další dva měsíce u Dursleyových. We already stated that we will not attempt to deal with pronouns. We tried to propose rules for choosing the correct substring of the entity name. The diversity of wikidata makes such a task extremely difficult. Together with the risk that we would decrease the precision of entity matching we decided to stick with full names only.

In the wikitext parsing section, we prepared spans and ids for wikilinks. We merge these with the ones matched by this module and post-process them. We discard each mention whose span is within a span of different mention of the same entity. This removes duplicates and keeps the one that is more specific.

na obrázku s volemtrem ukázat, jak to nakonec funguje

3.5.4 Relation Matching

So far we obtained sentences with tagged entity mentions. For each tuple of entity mentions within the same sentence, we checked if a relationship of those two entities was present in wikidata (using the prepared QPQ list). Given the filtering in wikidata preprocessing we are guaranteed that there is at most one such relationship.

At this stage, we need to address likely incorrectly matched entities that make the dataset bloat. One example of such bloating that we encountered was in an article about kindergarten⁵, in the sentence “*Jsou závazná pro předškolní vzdělávání v **mateřských školách**, v **mateřských školách** zřízených podle § 16 odst. 9 školského zákona, v lesních **mateřských školách** a v přípravných třídách*

⁵<https://cs.wikipedia.org/w/index.php?oldid=18388144>

základních škol.” thousands of relation mentions were found. Many kindergartens are named *Mateřská škola* (kindergarten), all of them are an instance of the abstract kindergarten entity and therefore candidate entities. If a sentence contains the term “*mateřská škola*” (or its form), all these entities will be matched. And the relationship “*Mateřská škola* is a *mateřská škola*” will be assigned many times as well. typografie

After investigating many other unusual cases, we decided to discard any sentence with at least 10 entity mentions in it. We also tried to experiment with different limits, but the results were unconvincing. For example, increasing the constant to 50 keeps an additional 13% of relationship mentions but extends the set of sentences only by 1%.

zašuměné histogramy

3.5.5 Characteristics of the Generated Dataset

The full dataset, which was obtained by the process we described in the previous sections, contains almost one and a half million relation mentions. In the next few paragraphs, we will talk about some more detailed statistic of this dataset - CERED0.

We found at least one mention in 293591 articles. In these articles, the average number of found mentions is slightly less than 5 and the median is nearly 7. The article with most mentions is Spojené království⁶ (United Kingdom). The distribution of mentions in articles is shown vložit histogramy

There are 490501 different sentences that are used in the mentions. We set the limit on entity mentions per sentence to 10. On average there were approximately 3 relation mentions in a sentence (that had at least one mention) and the maximum of 72 mentions per sentence was reached 34 times. The length of sentences ranges from 2 to 401 tokens, the very short ones usually came from templates that were not removed, on the other hand, the very long ones are often caused by incorrectly written articles⁷. We tried to remove all templates to see if the range (and distribution) of the number of words improves, but we did not find a significant improvement.

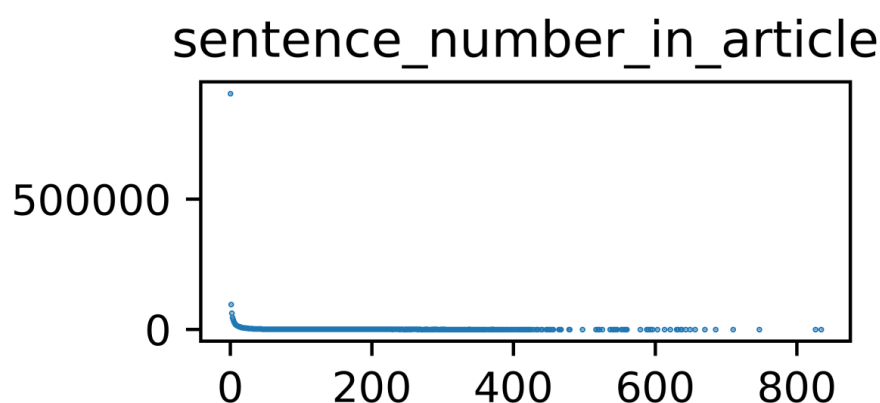


Figure 3.6: todo

⁶<https://cs.wikipedia.org/w/index.php?oldid=18752891>

⁷Example of an incorrectly written article <https://cs.wikipedia.org/w/index.php?oldid=18723498>

Another possibility is to observe how the position of a sentence in an article influences the number of relation mentions. We expected that the first sentences in an article will contain the highest number of relation mentions. The first sentences tend to contain wikilinks and the use of pronouns or shorter names is limited because each entity has to first be introduced by its full name. As we can see in Figure 3.6, our hypothesis seems to be correct, and partially applicable to a couple of sentences, not just the first one. 904803 mentions come from sentences that are first in their respective articles. This constitutes over 60% of all mentions.

3.6 CERED Versions

In this section we describe how and why we decided to postprocess the generated datasets.

The full CERED is already a valid relationship classification dataset. It has nearly one and a half million mentions, but as we discussed in the previous section, some of them might be of poorer quality than others. In this section, we will describe different versions of CERED with CERED0 being the biggest (least filtered) and CERED4 the smallest.

Each version is split up into three disjunct sets: train set, dev set and test set. Ideally, the test set would operate on a different set of entities (so that models learn to predict relationships based on sentences, not on the knowledge of entities). We believe that such a restriction is unnecessarily strong. Some entities are mentioned in many articles, which does, in fact, make them part of common knowledge (connected to the language we train on). Instead, we decided to relax the restriction to distinct articles. We sampled two sets of 10000 articles, one for test and one for dev set. In each version of CERED, the same articles are used for dev, test and train.

3.6.1 CERED0

CERED0 is the raw dataset just as we described it in the previous section. We do not artificially change the relation type, so no negative relation is present. We do not recommend this version for direct training, we mostly keep it to preserve the full information obtainable from our generator.

3.6.2 CERED1

CERED1 is close in size to CERED0. We removed long (over 100 words) and short (under 5 words) sentences. We also removed overly represented relations and changed labels of underrepresented relations to OTHER, which will be the negative relation in CERED1. We will properly define under-represented and over-represented relations in the next paragraph where we describe the CERED2 dataset.

Table 3.2: CEREDs statistics.

Dataset	# mentions	Inventory size	Negative %	Test dataset
CERED0	1 462 546	692	0	25 066
CERED1	999 287	64	14.1	17 214
CERED2	374 914	64	18.8	6 530
CERED3	157 972	64	22.6	2 787
CERED4	10 766	59	22.1	188

3.6.3 CERED2

CERED1 has 3 potential flaws that we have not addressed, yet. First, one sentence could be included multiple times in a dataset. Second, there were some relations with only a few relation mentions. Third, we did not try to handicap overly represented relations.

We try to address these issues in CERED2. We start with CERED1 and for each sentence, if there are multiple relation mentions, we keep only the mention that was the least represented in CERED1, and discard the others. Next, we count the sizes of representations of all mentions and say that a relation is under-represented if it has less than 1000 mentions. We take all the mentions whose relation is considered under-represented and relabel them to OTHER. This change is also applied to CERED1. After such relabeling, the OTHER relation still was not the most represented. We decided to discard mentions of the two most represented relations: *INSTANCE OF P31* and *COUNTRY P17*.

3.6.4 CERED3

Even after the restriction on the uniqueness of sentences, half of the CERED2 sentences are the first sentences within the articles they originated from. Such sentences have a rather unique structure that is not as common outside of Wikipedia. This inspired CERED3 which is the remainder of CERED2 after removing all mentions in “first” sentences.

CERED4

There are two stages in the CERED generation process that might be relatively unprecise. The entity matching stage and the relation matching stage. CERED4 keeps only the relation mentions from CERED3, in which both entities were manually-labelled directly in wikitext in the form of wikilinks. The CERED4 dataset is, unfortunately, small, but has the highest potential to be precise.

Other Considered Variations

When curating the CEREDs we considered many different criteria. For example, we chose the constant 1000 for underrepresented relations after several experiments. The good thing about this value is that only two relations were more dominant than the new OTHER relation and at the same time, we would keep 63 positive relations. We wanted our dataset to be at least somewhat comparable to other datasets in the proportion of data and inventory size. S10T8 contains 10717

mentions and 10 relations, TACRED contains around 100 000 mentions and 41 relations. The size of Riedel NYT depends, just like its inventory, on what data is used (and whether duplicates are removed and so on). We worked with nearly 700 000 mentions and 58 relations. Since the sizes of CERED1-3 range from approximately a million to approximately 160 000 mentions, we wanted the size to be around 50 or 60. The distributions of relations in different datasets is captured in ??.

Named entities. One more labour intensive variation we wanted to create was supposed to be focused on named entities. We tried to curate a set of named entity types (such as person, place, etc.) based on wikidata, in order to later use them as additional information during training.

We used the QPQ triples described in the ?? section to create an oriented graph with two types of edges. The first edge type - instance edges - connect wikidata items Q_1 and Q_2 if there is a triple Q_1P31Q_2 , where P31 is the wikidata property *instance of P31*, the second type - subclass edges - are added for the *subclass of P279* property. The goal of the graph is to capture the transitive nature of the two properties, i.e. if Harry Potter is an instance of a fictional human and fictional human is a subclass of a human, we can assume that Harry Potter is a human. We consider a wikidata item to be a named entity if it is an instance of some other item. If we are interested in all the named entities belonging to some entity type (node in the graph), we can simply find all the named entities for which there is a path between the type and the entity.

We hoped that with such representation, we will be able to curate reasonable set of types. For example, we might want to have a “human” entity type and we might intuitively expect *Harry Potter Q3244512* (fictional human) to be human. However, as he is fictional, he is not considered to be human in the wikidata ontology. Such an issue can be easily solved by having a “person” entity type instead of “human”. The problem is that there are just too many similar problems and it is out of scope for this thesis to properly analyze the whole wikidata ontology.

Part II

Training

4. Previous Work on Relationship Extraction

In this chapter, we first introduce the most popular pre-trained models that are currently used in natural language processing (NLP). Then we discuss different metrics that are used on Relationship Extraction. Lastly, we describe previous work done on the topic of training models for the relationship extraction task.

4.1 Deep NLP models

Lately, NLP tasks have been dominated by solutions using pre-trained deep neural models with the Transformer architecture. This section aims to roughly describe such a model to a reader familiar with the basics of neural networks. We introduce the two big concepts used in such models: the Transformer architecture and pre-training. Then we are able to explain how an actual modern NLP model looks and is trained. We choose the BERT model for this purpose since it is likely the best-known, and we also use it later in this thesis.

4.1.1 The Transformer Architecture

Suppose that we aim to perform some NLP task, like translation, using deep neural networks. We want the network to read a sentence in language A and output a translation of the sentence in language B. If the model was just a sequence of densely connected layers (so-called feedforward neural network), it would be nearly naive to expect the model to function. There is no structure that would help with gathering information about the entire sentence, and there is also no intuitive place where we would expect the model to switch from one language to another. There is nothing ensuring that the translated sentence is good. Those arguments are vague but might provide some intuition.

The recurrent neural networks (RNN) were used in the past to help the network capture the entire sentence. The RNNs introduce a cell that is more complicated than a typical neuron (perceptron). This cell usually has internal memory and the ability to both add and forget information. In the network, this cell is then presented with a sequence of inputs. The cell then goes through the sequence one by one and edits its internal state. If we feed it with words of a sentence, we might expect the internal memory to contain an abstract representation of the sentence.

There are two main issues with RNNs. Working with RNNs is slow. We input the sentence “word by word”, so instead of one step per layer, we need n steps for a sentence of length n . Moreover, especially if the sentence is long, it is difficult for the cells to remember the whole sentence and not just the later part. Both of these issues arise from the fact that input is fed into RNN cells sequentially.

In Vaswani et al. [2017], the Transformer architecture is introduced. It builds on a so-called encoder-decoder architecture. In the translation task, we can roughly say, that the encoder would process the sentence in language A, the

decoder an already predicted part of the translated sentence, and the network would predict the next word for the translated sentence.

The issue of the accumulation of the right information at the right place is addressed with an attention mechanism. Attention was used first in the RNN architecture. Later, it was discovered, that the attention itself is powerful enough to make the slow RNN cells unnecessary. The traditional attention mechanism makes it possible to focus only on the important parts of the input sentence. For example, if we want to translate a specific word in a sentence, the network can learn that some parts of the input sentence are more important than others.

An additional attention mechanism - self-attention - was introduced in the Transformer architecture. Unlike the traditional attention, which is used to look at a different layer (for example, it can transfer information from encoder to decoder), the self-attention usually works with inputs of the same layer as where it is applied. Self-attention attempts to add context information to the current embedding from the current embedding of the input sequence. For each input vector, we first compute three vectors: key, query and value. When computing self-attention for the i -th input, we compute the dot product of the i -th query vector with each key vector, we call the resulting numbers scores. The j -th score represents the distance of the i -th query and the j -th key vector. We then convert the scores to non-negative numbers that sum up to 1 and use those values to weight the value vectors (simply by multiplication). Lastly, we sum such weighted vectors to create the i -th output. We described the operations on vectors, but they can be rewritten for whole matrices, which makes them faster to compute. Precisely, let Q , K and V be matrices, where the i -th row is the i -th query, key and value vector. Then the attention can be computed as $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V$, where d is the dimension of the key vectors.

In the Transformer architecture, both the decoder and the encoder contain a sequence of blocks, where each block is a self-attention layer followed by a feed-forward network.

So far, we ignored the fact that words cannot be directly fed to the network as an input. Usually, words are mapped to a vector in an embedding space. Such mapping is reusable in multiple tasks and brings us to the next section, where we will explain them a bit more.

4.1.2 Pre-training

The training of deep neural networks can require a great deal of both computational resources and annotated data. The term pre-trained model indicates a (part of) neural network, that can be reusable for multiple tasks or settings. Moreover, the pre-training is often based on unsupervised data, which lowers the requirements for annotated data later in the fine-tuning. Such reusability fastens deployment of complicated models and makes them accessible for a broader spectrum of people and companies. Pre-training is time-efficient, economical and consequently ecological.

Word embeddings can be seen as an example of pre-training. Word embedding is a vector that represents a word in a vector space (as, for models, vector spaces are easier to work with than words). The distance of embeddings of two words

should reflect their semantic relation. An embedding itself is usually inferred from a pre-trained neural network and is used as an input for the task-specific model. The model that generates embeddings is usually trained by some variation of predicting a missing (masked) word from its context. The popularity of word embeddings rose in the previous years, most popular embeddings were likely the Word2vec Mikolov et al. [2013].

Nowadays, a new type of embeddings is starting to get popular. In the classical word embeddings, a word always has the same embedding. The word “book” has the same embedding in “*She wrote a book on car maintenance.*” and “*We were advised to book early if we wanted to get a room.*” which is intuitively wrong. Contextualized embeddings take into account the entire (or at least preceding part of) sentence. One such embedding can be created via the BERT model.

4.1.3 BERT

BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. [2019]) is a family of models that achieves state-of-the-art results on various NLP tasks. BERT model takes a sequence of words as an input and computes a representation of each word while taking into account both the left and right-side contexts of the word. BERT models are pre-trained on large corpora of unlabeled text and then fine-tuned for a specific task.

The model was pretrained on two NLP tasks - Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

MLM is an adaptation of the task of prediction of a word from its context. The model takes a sentence, where some percentage of the words is replaced with a special token. The goal is to predict the original words in the sentence. The masking is essential; without it, the task would be trivial; at the same time, some context has to stay unmasked; otherwise, the task would be impossible. If we always mask the percentage of the words with the special token, a significant mismatch between real sentences and the one used in training would be created. The masking is, therefore, more sophisticated. We choose 15% of words to be masked. If a word is to be masked, with 80% change, it is replaced with the special token, with 10% it is replaced with a random word, and in the last 10% the word is kept (which helps to bias the model towards the actual word).

In NSP, we try to guess whether two sentences were consequent in the original text. The input is structured into two sentences via a separator token and segment embeddings. This task helps the model to learn to work on the sentence level, not just on a token level.

To fine-tune the model, we usually add a shallow feed-forward network on top of the BERT, which classifies chosen features from the BERT into the result space. Then we train the model on the task-specific dataset as we would train any other model.

In the original paper (Devlin et al. [2019]) two BERT models were published: BERT_{BASE} and BERT_{LARGE}, they differ in size, but both trained on the same English corpora (the BookCorpus and English Wikipedia). Later multilingual BERT was released¹. It was trained on the top 100 languages with the largest Wikipedias.

¹<https://github.com/google-research/bert>

4.2 Metrics

This section focuses on metrics since the quality of a model in the relationship extraction task is often hard to determine. We first define some standard metrics and discuss the pros and cons of each of them. Then we overview metrics used for the relationship extraction task on different datasets.

4.2.1 Binary classification

We will start with metrics for binary classification. In binary classification, we are presented with an input vector, and the goal is to determine whether the vector is of class A or class B. Each prediction then falls into one of the following categories: correctly classified A input, correctly classified B input, wrongly classified A input as B, and a B input wrongly classified as A (Figure 4.1).

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

		Predicted class	
		A	B
Actual class	A	True A	False B
	B	False A	True B

Figure 4.1: Confusion matrix for binary classification, normally count of occurrences would be written in each square.

Another way to look at the same situation is just to predict whether an input is of class A or not. This way the prediction is a True / False value determining whether the input is of class A. This way, we can define the previously mentioned categories without using the specific classes as **true positive** (TP), **true negative** (TN), **false negative** (FN) and **false positive** (FP). Visualization of the result of classification on a dataset is called a **confusion matrix**, we include such matrix in Figure 4.1. We will use the abbreviations to represent the number of predictions that belong to the given category.

Accuracy expresses the ratio between correct predictions and all predictions.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

Precision expresses the ratio of correctly predicted positives within all predicted positives. Therefore, precision is a good metric if we want to avoid mistakenly classified negatives as positives.

$$Prec = \frac{TP}{TP + FP} \quad (4.2)$$

Recall is the complementary metric to precision. It expresses the ratio of all positives that were correctly predicted. In other words, it should be used when we need to find the maximum of positives in the data.

$$Rec = \frac{TP}{TP + FN} \quad (4.3)$$

Often we might want a trade-off between being as precise as possible and recalling as much as possible. **F1 score** is a harmonic mean of precision and recall (scaled to range from 0 to 1):

$$F1 = 2 \frac{Prec \cdot Rec}{Prec + Rec} \quad (4.4)$$

F1 is quite widely used in competition tasks.

To emphasize that the right choice of metric is significant suppose that we have balanced data (both true and false classes are equally represented). If our classifier just predicted that every input is positive, we would obtain the following: 0.5 accuracy, 0.5 precision, and 1 recall. If we were to predict all negatives, accuracy and precision would remain 0.5 but recall suddenly drops to 0. If we were to predict the result with even chances for both classes randomly, the expected results would be 0.5 for all of those metrics. We just described three very different classifiers, and the only thing we learned from accuracy and precision was that they were equally bad, without any insight about them. Recall, in contrast, successfully gave us insight about what the predictions likely are, but evaluated a bad classifier with the highest possible score. We should also note, that the metrics will likely change if we exchange which class corresponds to True and which to False.

This whole section is in this thesis mostly to remind us that if we want to score well in a given metric, we will likely exploit the metric even if it might worsen our classifier. The choice of a metric for a task determines what gets optimized. Later in this section, we will debate such issues in our scenario, in the relationship extraction task.

4.2.2 Multiclass classification

We already run into issues with the asymmetry of precision and recall in binary classification (it is dependent on which class is chosen to be the positive one). We can address this by creating metrics per class. In the previous subsection, where we had a classifier for classes A and B, we would get two sets of metrics, each describing the ability of the classifier to recognize given class apart from the rest.

Now we can easily extend this per class approach to multiclass classification. The formulas will remain the same; only the way we obtain the TP, FP, TN, and FN values is a little different. In a sense, nothing changed - if we imagine that the classifier is still binary, then the situation is the same. However, if we compute those values out of confusion matrix (for class B) then TP is the value on position [B, B], FP is the sum of all in column B without TP, FN the sum of row B without TP and the sum of cells outside of the Bth row and column are the TN. (Figure 4.2)

		Predicted class		
		A	B	N
Actual class	A	True A	False B	False N
	B	False A	True B	False N
	N	False A	False B	True N

Figure 4.2: Confusion matrix for N-class classification

As a robust way of examining the quality of the classifier, one could simply look at the confusion matrix and all the per-class metrics. Although this would be insightful, it is not the most practical in terms of a clear comparison of two classifiers. Ideally, we aim for a metric or metrics that are as descriptive and comprehensive as possible and also define an ordering of the classifiers.

Ideally, we would compute one number that fully describes the quality of the classifier. The common practice is to assign each metric a weight and combine them. Before we do so, we should acknowledge that the dataset we evaluate the classifier on needs to be taken into consideration.

An ideal dataset would be perfectly balanced. In real life, we encounter two types of imbalance in datasets:

- class representation distribution (CRD) is not uniform - classes are not equally represented
- class representation distribution is different in the test and the train part of the dataset

The second imbalance is tricky. Often, when optimizing the classifier, we do not know the CRD of the test dataset. We will, therefore, mostly focus on the first one.

Macro-averaged metrics

An easy way to combine a metric over multiple classes into a single value is to compute the arithmetic mean. In most libraries and papers, the term macro-[metric] (macro-recall, macro-F1, etc.) is used. Macro averaged metrics tend to be the easy option that is used without much thought. So much so, that even though two different definitions of macro-F1s exist, the exact formula is often not included in papers. (The more common formula is the arithmetic mean of

classes F1s, but the less commonly used formula where the F1 is computed from macro-recall and macro-precision is also used as Opitz and Burst [2019] reports.)

Micro-averaged metrics

The second most used approach of computing a single metric over multiple classes is to compute a weighted average with the weight corresponding to the support of individual classes. In most libraries and papers, the term micro-[metric] (micro-recall, micro-F1, etc.) is used.

The main difference between micro and macro metrics is that macro metrics treat all classes equally, i.e. if there is a massively under-represented class, the metric prefers TODO. On the other hand, we should use a micro metric if we want to optimize the total accuracy of the model.

4.2.3 Relationship Extraction Metrics

In Existing datasets we described three relationship extraction datasets. In this subsection, we will focus on the metrics they use and propose the recommended metric for CEREDs.

S10T8 metric

For the S10T8 datasets, authors (Hendrickx et al. [2010]) define the official metric directly in the paper: “Our official scoring metric is macro-averaged F1-Score for (9+1)-way classification, taking directionality into account.” This means that the model classifies into all $2 * 9 + 1$ classes, but before they compute the macro averaging, they merge the result into only the $9 + 1$ direction-less classes.

TACRED

On TACRED’s git page², we can find the official implementation of an evaluator which computes a slightly modified version of micro metrics. Micro metrics are computed as the average of metric per class weighted by the support of that class. The authors modified the metric such that the weight of the NO RELATION class is zero. The micro-F1 is computed from the micro-precision and micro-recall. The primary metric for comparing results seems to be the micro-F1.

The decision to exclude the NO RELATION class is understandable. The class makes up 80% of the dataset, therefore the micro metrics would be mostly influenced by the performance on this class. Such behaviour is unwanted considering the fact, that in usual use cases we care more about the model to correctly classify positive relations rather than the negative one. We would like to note that it is not possible to cheat the metric by forbidding the classifier to predict the negative relation. If we do so, it would negatively influence the false positives of all classes and, therefore, decrease the overall performance.

²<https://github.com/yuhaozhang/tacred-relation>

Riedel NYT

For the Riedel NYT dataset, there was no official metric determined by the authors. Papers report different metrics and therefore the results are hard to compare.

CERED

The Riedel NYT dataset shows, that we should decide, what the primary metric for CEREDs will be. We decided to use micro-F1. The negative relation in our datasets is not as dominant as in TACRED, so we do not think it is necessary to exclude it. On the other side, the datasets is not as nicely balanced as S10T8 and any macro-metric would be incline the models to overly focus on smaller classes.

4.3 Relationship Extraction Models

matching the blanks .. psát? i když se neimplementuje? U výsledků je dvakrát zmiňuji, že jsou nejlepší nebo tuhle sekci prostě odstranit?

5. Relationship Extraction on CERED

The goal of this part of our thesis is to train a relationship extraction model for the Czech language. To the best of our knowledge, there is no baseline for such model, since there is no Czech relationship extraction datasets. In the first part of this thesis, we generated CEREDs, a family of relationship extraction models, that we can use.

If we just trained the model on our dataset, we would be unable to determine, if the model is well designed. We therefore adapt the model to be trainable both on Czech and on English. We assume that if the English version is competitive on popular English relationship extraction dataset, the Czech version is reasonably good.

In this chapter, we first describe the model, then we evaluate it on English datasets and lastly we report the results on CEREDs.

5.1 Model

We based our model on the BERT model described in the BERT section. The model architecture is quite straight forward. The goal of the whole model is to predict the relation based on a given relation mention. We modify the input by adding some special tokens as we show in Figure 5.1. Then we use the tokenizer, that is attached to the specific BERT, and we feed the output of the tokenizer to the model. The model computes abstract representations of all input tokens. We concatenate the representations that correspond to some of the special tokens ([E1], [E2] and [CLS]). On top we apply a dropout layer and finally a dense layer which predicts the class.

For Czech, we fine-tune the multilingual BERT, for English we fine-tune BERT_{BASE}. We use the transformers library (Wolf et al. [2019]) for BERT manipulation.

Rose Weasleyová je nejstarším dítětem **Rona Weasleyho** a Hermiony Weasleyové.
[CLS] [E2] Rose Weasleyová [/E2] je nejstarším dítětem [E1] Rona Weasleyho [/E1] a Hermiony Weasleyové.

Figure 5.1: Special tokens added for each mention. The sentence is a mention of FATHER(Ron Weasley, Rose Weasley). We add tokens to mark where each entity starts and end. The [CLS] token is part of BERT input and was pre-trained to capture the whole input sequence.

5.2 Results

We trained the model described in the previous section on the three datasets introduced in Existing datasets and on CEREDs. We hoped for our model to be competitive with other models, that build on BERTs. At the same time, we will not use any additional data for the training, therefore we do not expect to score

as well as state-of-the-art results. We referred to results provided by nlpprogress¹ when researching the scores achieved on each dataset.

We did not try to fine-tune the meta parameters (such as the learning rate or dropout rate).

5.2.1 S10T8

On S10T8 the model stopped training after 10 epochs (i.e. in less than 20 minutes on a single GPU) and achieved 86.54% in the official metric. The state-of-the-art result was achieved by Baldini Soares et al. [2019] with 89.5%.

5.2.2 TACRED

Training on TACRED did not take much longer (4 epochs and under two hours). The state-of-the-art on this dataset was also achieved by Baldini Soares et al. [2019] with 71.5% F1 (micro-averaged over instances with positive relationships). We scored 65.65.

5.2.3 Riedel NYT

We are not able to properly compare our result with other, since the Riedel NYT dataset (introduced in [Riedel et al., 2010]), is extremely underdocumented. We did not find any official information about which data should be used for training. We also are not sure, if duplicities test datasets should be removed as part of preprocessing, or if they should be kept. Moreover, there is no official metric on this dataset.

Apart from that, our model learns to always predict the negative relation (about 80% of the dataset is the negative relation). So no matter the metric, the model does not work. The aim of evaluating our model on more datasets was to show how it performs, not to change the architecture, so that it performs well.

5.2.4 CEREDs

We trained our model on all the CERED versions. The performance of respective models was mostly as expected. We will pinpoint few things we observed in the results.

We assume, that the highest macro-F1 was achieved on CERED2 because we choose the relation inventory for CERED1-4 based on CERED2. CERED0 has a huge relation inventory and therefore it is not surprising, that when weighting over classes, the metric is low.

The fact, that model trained on CERED4 performs relatively well, shows, how remarkable BERTs (or pre-trained models) are for tasks with small datasets. The low macro-F1 on CERED4 can be explained quite easily. When we put constraints on which mentions will be kept in CERED4, we did not ensure, that the dataset will be balanced. In figure Figure 5.2 we see, how imbalanced the dataset really is.

¹http://nlpprogress.com/english/relationship_extraction.html

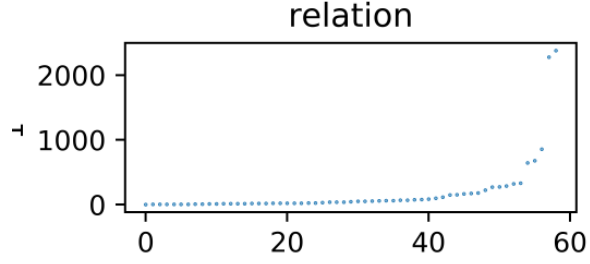


Figure 5.2: Relation representation in CERED4.

Table 5.1: Results on CEREDs, we measured micro and macro F1 on the corresponding test sets for each version. Moreover, we measured the performance of models trained on CERED1-4 on CERED2 test set, to see the direct comparison.

Dataset	#	micro-F1	macro-F1	micro-F1 on CERED2 test
CERED0		84.49	22	-
CERED1		83.94	77	81.72
CERED2		84.78	80	84.78
CERED3		77.29	74	82.33
CERED4		77.66	32	79.79

Conclusion

In this thesis, we proposed a methodology for generating relationship extraction datasets using Wikipedia and Wikidata. We analysed different pitfalls that we faced when we implemented the generator and generated CEREDs (Czech Relationship Extraction Datasets).

We designed a neural network architecture for the relationship extraction task. In the network, we employ BERT, which increases the quality of the network. We demonstrated that the proposed architecture preforms not much worse than the state-of-the-art on multiple English relationship extraction datasets. We reported the performance of the network on CEREDs.

5.3 Future work

5.3.1 Other Languages

This thesis focused on relationship extraction in the Czech Language. We believe that our work could be extended into different languages. In the CERED generation process, there are three language-dependent steps:

- Wikidata preprocessing - we only worked with items, that had at least one Czech name. The filtering can be easily changed to arbitrary language. For languages with small Wikipedia, it might be reasonable to consider relaxing the condition.
- Wikitext parsing - names of templates are in the language of the Wikipedia. We can remove the content of all templates. It will negatively impact the size of the dataset, but it too drastically.
- Entity matching - this step is heavily dependant on the language. We believe that for some languages it might be satisfactory to change the language in MorphoDiTa (for example MorphoDiTa supports the Slovak language). We might consider exchanging MorphoDiTa with a more mainstream tool (for example, spaCy) that supports more languages.

If we implemented the changes proposed above and automatised the whole process, we might be able to create a dataset (and a model) for each language, that the multilingual BERT supports (coincidentally BERT was trained on the 100 biggest Wikipedias).

5.3.2 Wikidata ontology

In section Other Considered Variations, we mentioned that additional information for training could be obtained from wikidata. Such information is often available in other relationship extraction dataset. In the future, we would like to use Wikidata for extracting such information and add it to CEREDs.

Bibliography

- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1279. URL <https://www.aclweb.org/anthology/P19-1279>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://www.aclweb.org/anthology/N19-1423>.
- 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>.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.
- Juri Opitz and Sebastian Burst. Macro f1 and macro f1. *ArXiv*, abs/1911.03347, 2019.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. pages 148–163, 09 2010. doi: 10.1007/978-3-642-15939-8_10.
- 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>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.
- 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].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

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