BIOGRAPHICAL INFORMATION EXTRACTION: A LANGUAGE-AGNOSTIC METHODOLOGY FOR DATASETS AND MODELS

ALISTAIR PLUM

A thesis submitted in partial fulfilment of the requirements of the University of Wolverhampton for the degree of Doctor of Philosophy

2022

Signature: . . . . . .

Date: . . . . . .

ABSTRACT

Information extraction (IE) refers to the task of detecting and linking information contained in written texts. While it includes various subtasks, relation extraction (RE) is used to link two entities in a text via a common relation. RE can therefore be used to build linked databases of knowledge across a wide area of topics. Today, the task of RE is treated as a supervised machine learning (ML) task, where a model is trained using a specific architecture and a specific annotated dataset. These specific datasets typically aim to represent common patterns that the model is to learn, albeit at the cost of manual annotation, which can be costly and time-consuming. In addition, due to the nature of the training process, the models can be sensitive to a specific genre or topic, and are generally monolingual. It therefore stands to reason, that certain genres and topics have better models, as they are treated with a higher priority due to financial interests for instance. This in turn leads to RE models not being available to every area of research, leaving incomplete linked databases of knowledge. For instance, if the birthplace of a person is not correctly extracted, the place and the person can not be linked correctly, therefore not leaving linked databases incomplete.

To address this problem, this thesis explores aspects of RE that could be adapted in ways which require little human effort, therefore making RE models more widely available. The first aspect is the annotated data. During the course

of this thesis, Wikipedia and its subsidiaries are used as sources to automatically annotate sentences for RE. The dataset, which is aimed towards digital humanities (DH) and historical research, is automatically compiled by aligning sentences from Wikipedia articles with matching structured data from sources including Pantheon and Wikidata. By exploiting the structure of Wikipedia articles and robust named entity recognition (NER), information is matched with relatively high precision in order to compile annotated relation pairs for ten different relations that are important in the DH domain: birthdate, birthplace, deathdate, deathplace, occupation, parent, educated, child, sibling and other (all other relations). Furthermore, the effectiveness of the dataset is demonstrated by training a state-of-the-art neural model to classify relation pairs. For its evaluation, a manually annotated gold standard set is used. An investigation of the necessary adaptations to recreate the automatic process in a multilingual setting is also undertaken, looking specifically at English and German, for which similar neural models are trained and evaluated on a gold standard dataset. While the process is aimed here at training neural models for RE within the domain of digital humanities and history, it may be transferable to other domains.

ACKNOWLEDGEMENTS

Diese Arbeit ist gewidmet an / This work is dedicated to:

Heinrich & Käthe Plum

Leslie & Doreen Williamson

I would like to thank my director of studies Prof Ruslan Mitkov, for granting me the opportunity to study at RGCL, as well as my supervisors Prof Constantin Orăsan and Dr Richard Evans, for their supervision and guidance. Thank you also to Dr Tharindu Ranasinghe, Dr Maria Kunilovskaya, and Rocío Caro Quintana for not only being the best colleagues, but also valued friends.

I will be forever grateful to Mum & Dad, Charlotte, and Sonja, for their encouragement and support at all times.
I would quite surely not have made it this far without them.

CONTENTS

List of Tables

List of Figures

List of Examples

List of Abbreviations

AAS Austrian Academy of Sciences

ABD Austrian Biographical Dictionary

ACE Automatic Content Extraction

APIS A Prosopographical Information System

CNN Convolutional Neural Network

CoNLL Computational Natural Language Learning

CRF Conditional Random Field

CSV Comma Separated Values

CURID Wikipedia Identifier

DH Digital Humanities

DUC Document Understanding Conference

F1 F1-score

GDS Guided Distant Supervision

GND German National Dictionary

HMM Hidden Markov Model

HTML HyperText Markup Language

IE Information Extraction

IOB Inside Outside Beginning

JSON JavaScript Object Notation

KB Knowledge Base

KBP Knowledge Base Population

LSTM Long Short-Term Memory

ML Machine Learning

MUC Message Understanding Conference

NE Named Entity

NEL Named Entity Linking

NER Named Entity Recognition

NLP Natural Language Processing

NLU Natural Language Understanding

NN Neural Network

OIE Open Information Extraction

P Precision

POS Part-of-Speech

R Recall

RE Relation Extraction

RNN Recurrent Neural Network

QID Wikidata Identifier

SemEval Semantic Evaluation Conference

SOTA State-of-the-Art

SQL Structured Query Language

SVM Support Vector Machine

TAC Text Analysis Conference

XML Extendible Markup Language

List of Publications

Preliminary versions of the research presented in this thesis appeared in the following peer-reviewed papers:

Plum, Alistair, Tharindu Ranasinghe, Spencer Jones, Constantin Orasan, and Ruslan Mitkov (2022). "Biographical Semi-Supervised Relation Extraction Dataset". In: *45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22, pp. 3121–3130.

Plum, Alistair, Marcos Zampieri, Constantin Orasan, Eveline Wandl-Vogt, and Ruslan Mitkov (2019). "Large-Scale Data Harvesting for Biographical Data". In: *Biographical Data in a Digital World 2019*. Vol. 3152. CEUR Workshop Proceedings, pp. 66–72.

CHAPTER 1

INTRODUCTION

The importance of *biographical information* as an integral part of ancient to modern history should need no further explanation, and as such, the challenge of making such information available to all humans warrants appropriate research. Obtaining biographical information has, up until the wider-spread availability of online information resources, involved consulting books or other kinds of written texts. Extracting the information in a structured way remains a challenge either way, as it requires reading and understanding these texts. Even with online information resources and digitised texts, the challenge of extracting the information in a structured way still remains, and therefore requires automated computational methods.

Modern methods of computing, that allow making biographical information available in a structured database for researchers and members of the general public alike, have become a topic of great interest. While there are well-known projects, such as the online encyclopedia Wikipedia and its Wikimedia derivatives, large biographical databases are increasingly becoming available, not only in English, but across a variety of languages (Reinert and Ebneth, 2017). This is evidenced in particular by projects that have been or are being digitised, such as

the Slovenian Biography (Erjavec et al., 2018), the Deutsche Biographie (Reinert et al., 2015), and the Austrian Biographical Dictionary (ABD) (Wissenschaften, 2013). Whether digitising existing resources or finding new biographical information, there is a clear trend towards larger, easily accessible resources. This trend highlights a clear need for methods of detecting and extracting all kinds of biographical information.

Information extraction (IE) is a process in *natural language processing* (NLP) that detects, extracts and structures information contained in written texts (Jurafsky and Martin, 2018). A vital task for extracting information from unstructured texts, IE is essential for any natural language understanding (NLU) task. In general, IE combines the output of specific NLP tasks, which usually include, but are not limited to, *named entity recognition* (NER), *named entity linking* (NEL) and *relation extraction* (RE) (Jurafsky and Martin, 2018). These tasks each solve various NLP problems and are usually addressed separately before being combined. A generic approach may therefore involve the detection of persons, locations and companies, the linking of these entities to an existing database for verification and finally relating these entities to actions, events and other entities.

In the past these tasks have often been solved in application-specific ways, although nowadays the solutions usually involve *machine learning* (ML). This means that a ML model is trained to predict whether a span of text is a named entity, to link a named entity to an existing database or to predict the relation between two constituents. Training these models can be carried out in a supervised

or unsupervised setting. While the former requires annotated textual data to learn how to make predictions, the latter does not, but does require other solutions in place of textual data, such as defining certain patterns to generalise the structure of a certain piece of textual data. Supervised models, especially in the context of neural architectures, often require a large amount of annotated textual training data. The requirement for large amounts of annotated textual training data can often be prohibitive for certain target applications, including certain domains and multilingual approaches.

This thesis will present research concerning IE and the steps necessary to make it useable in the context of biographical information, as well as adaptations for multilingual use. More specifically, the thesis will address the requirement of annotated textual data, by suggesting a new approach to textual data compilation for machine learning which is related to distant supervision (Mintz et al., 2009). As such, the approach makes use of Wikipedia and Wikidata to automatically compile annotated textual data, for use as training data for neural machine learning models. Furthermore, the usability of this training data will be explored by training models for biographical relation extraction, followed by an examination of potential multilingual adaptations of the previously proposed approach. The objective of the thesis is to assess whether the new approach to training neural models is suitable for biographical relation extraction, and whether such an approach can be adapted for multilingual use.

1.1 Research Questions

The research presented in this thesis addresses the following research questions:

RQ-1 *Is Wikipedia a suitable data source to facilitate the extraction of biographical information?*

RQ-1a *Which processing steps are required to use Wikipedia?*

RQ-1b *Could Wikipedia and Wikidata be used in an automatic annotation approach for training data?*

Research question **RQ-1** will be addressed in Chapter 3, where two pilot studies are described that utilised Wikipedia as a data source for biographical information extraction. The research presented includes a description of the necessary processing steps to use Wikipedia, therefore answering **RQ-1a**. Moreover, Chapters 4 and 5 will also contribute to answering **RQ-1**, as well as answering **RQ-1b**, by means of a more in-depth analysis of a methodology revolving around the use of Wikipedia to train a neural network for biographical relation extraction.

RQ-2 *Are semi-supervised datasets effective for training a biographical relation extraction model?*

RQ-2a *How do certain processing steps affect the model performance?*

RQ-2b *Does pretraining improve perfomance over fine-tuned models?*

Research question **RQ-2** will be addressed mainly in Chapter 4, which presents and analyses a semi-supervised methodology for compiling a biographical relation

extraction dataset. Furthermore, the chapter presents an evaluation of various neural models trained with this dataset in various preprocessing and training settings. The different settings also serve the purpose of answering sub-questions **RQ-2a** and **RQ-2b**. In addition, Chapter 5 also contributes to **RQ-2**, by validating the approach on another language.

RQ-3 *Is it possible to adapt guided distant supervision to another language?*

RQ-3a *Can the data compilation approach be adapted to German?*

RQ-3b *Do cross-lingual and monolingual models perform comparably?*

RQ-3c *Is machine translation a more effective alternative?*

To answer these questions, Chapter 5 demonstrates various adaptations to the approach established in Chapter 4 for language-independent use. The adapted approach directly addresses research question **RQ-3**. The research described uses German as an example language, tests various neural models under varying learning settings and evaluates the results against a baseline system using machine translation. These characteristics of the methodology are aimed to address the sub-questions **RQ-3a**, **RQ-3b** and **RQ-3c**, respectively.

1.2 Contributions

The research presented in this thesis makes a number of contributions to the field:

- A detailed approach to process Wikipedia and Wikidata for information extraction purposes.

- The introduction of guided distant supervision, a methodology to automatically compile annotated datasets to train information extraction models.
- An overview of the steps involved to adapt the guided distant supervision approach to another language.
- All models trained with the datasets presented during the course of this thesis. Includes monolingual, multilingual and cross-lingual models.
- All datasets compiled using the methods presented in this thesis (Table 1.1)

A number of datasets were compiled using methods presented throughout this thesis. An overview of these datasets is presented in Table 1.1. Because of the lower quality of the *WikiSents* dataset (see Section 3.3 for further details), this dataset will not be made publicly available at the time of submitting this thesis. It is important to note that the sizes for *Biographical* and *Biographical DE* are the total across all processing methods (see Chapters 4 and 5 for further details).

Name	**Annot. Level**	**Labels**	**Lang.**	**Size (Sentences)**	**Chpt.**
WikiSents	Sentence	2 Binary 4 Multi	EN	400,000 Train 4,984 Test	3.3
Biographical	Entity	10 Relations	EN	902,103 Train 3,000 Test	4; 5
Biographical DE	Entity	10 Relations	DE	124,562 Train 2,000 Test	5

Table 1.1: Overview of dataset contributions. Information includes the name of the dataset, the level of annotation, the number of labels, the language, the size counted in sentences, and which chapter it is used in.

1.3 Thesis Structure

The thesis is structured as follows. Chapter 2 presents an overview of background information related to the research described in this thesis. Chapter 3 presents methods for processing Wikipedia for biographical IE in a specific a research project, while also introducing a method for automatically annotating Wikipedia texts for IE purposes. The chapter also highlights the challenges of using such a large data source for NLP tasks. Following this, Chapter 4 introduces *guided distant supervision*, and details how parts of the previously described processing methods can be used to compile a relation extraction dataset for English. The chapter also evaluates a number of neural models trained on the dataset, discussing both pretraining and fine-tuning the models. Finally, Chapter 5 demonstrates how to adapt the guided distant supervision methodology to another language, as well as evaluating other potential alternatives to multilingual adaptation, machine translation and cross-lingual learning. Chapter 6 concludes the thesis by summarising its findings and referring back to the main research questions.

CHAPTER 2

BACKGROUND INFORMATION

This chapter presents background information for contextualising the research presented in the following chapters. Section 2.1 addresses the field of NER, which is a vital component of every information extraction system. Included with this is NEL (Section 2.1.3), which provides important functionality in order to resolve and correctly assign named entities to knowledge bases, a task which the methodology described in this thesis could precede. Section 2.2 describes tasks directly related to the extraction of information, looking in-depth at how information is defined, and the multitude of extraction methods that have been developed over the years. Section 2.3 presents resources related to biographical information, and as well as describing some methods framed in the context of Digital Humanities. The chapter is concluded by Section 2.4, which highlights areas where the literature presented throughout this chapter presents opportunities for further research.

2.1 Named Entity Recognition

NER is a fundamental part of natural language processing and an important part of IE. The NER process detects persons, locations, organisations, dates and more,

which can then be linked to other information in a text at a later stage. In more specific terms, the task of NER is to detect and label each type of proper name or *named entity* (NE) in a given text, where the interpretation of NE is related to the domain of the given text (Jurafsky and Martin, 2018). Whereas persons, organisations and locations usually represent the most common types of NE, there are many others. The NE types are often related to the domain of the texts being processed, for instance names of virus strands could be more relevant in the biomedical domain, and programming method names could be more relevant in the software engineering domain.

(1) [PER **Mike**] works for [ORG **British Telecom**] in [LOC **Dudley**].

(2) [PROT **ZntR**] is an autoregulatory protein and negatively regulates the chromosomal zinc resistance [PROT-GRP **operon znt**] of Staphylococcus.

(3) [RT **Cross-site scripting (XSS)**] vulnerability in [SV **Apple**] [SP **Safari**] allows remote hackers to [RT **inject arbitrary web script**].

Example 2.1: Examples of different named entities according to various domains.

Example 2.1 shows the expected results for some input sentences[1]. The first example can be considered as standard output for any general sentence, where *PER* stands for *person*, *ORG* stands for *organisation* and *LOC* stands for *location*. Standard tags are used for annotation, which are common for most NER systems and have been used since the first larger scale evaluation series of NER systems (Sundheim, 1995). The second example is from the biomedical domain (Kim et

[1] Some of the tags have been simplifed for illustration purposes.

al., 2004) and demonstrates the use of more specialised tags. Here, *PROT* stands for *protein* and *PROT-GRP* stands for *protein group*. The final example is from the domain of cyber security. It features tags that identify *relevant terms* (RT), *software vendors* (SV) and *software products* (SP) in the given example. These examples demonstrate that while there are tags that are somewhat standardised that capture more generic named entities, systems can be adapted to various domains and use appropriate tag-sets for each domain.

The development of NER systems was closely tied to shared tasks and their respective datasets. Most notably, the Message Understanding Conferences (MUCs) focused on IE, with NER regarded as a big component. The datasets (MUC-6 and MUC-7 in particular) accompanying this series are still used to this day, and feature the four basic entity types: *person*, *location*, *organisation* and *miscellaneous* (Grishman and Sundheim, 1996). The series of MUC conferences was followed by the Automatic Content Extraction (ACE) program (Doddington et al., 2004), as well the Conference on Computational Natural Language Learning (CoNLL) series. The datasets accompanying the CoNLL series focused on multilingual tasks, some of which NER tasks, and were not only released for English, but also Spanish, German and Dutch (Sang, 2002). These datasets also introduced the CoNLL format, which is still widely used today. Figure 2.1 shows an example of the CoNLL format[2], where a sentence is split into one word per line. Each line denotes further information about the current word: the part-of-speech tag, the noun chunk and finally the entity tag. The chunk and entity tags

[2] https://clips.uantwerpen.be/conll2003/ner/

are prefixed with a letter according to the IOB[3] scheme, which indicates the span of the respective tag. It should be pointed out, that the datasets discussed here can be said to be more general in terms of topic. It is often quite common for datasets to be highly specialised - either in terms of language, domain or both. This means that a large number of datasets exist for many different languages and domains, most notably including biomedical (Tanabe et al., 2005; Crichton et al., 2017) and social media (Basaldella et al., 2020; Derczynski et al., 2016; Nie et al., 2020).

```
U.N.        NNP  I-NP  I-ORG
official    NN   I-NP  O
Ekeus       NNP  I-NP  I-PER
heads       VBZ  I-VP  O
for         IN   I-PP  O
Baghdad     NNP  I-NP  I-LOC
.           .    O     O
```

Figure 2.1: Example of the CoNLL annotation scheme.

In the context of the research presented throughout this thesis, NER always plays an important role, since it detects and tags the named entities to be treated. Understanding how well NER algorithms can perform is important to contextualise the results obtained from the evaluation of the information extraction methods presented in later parts of this thesis. Furthermore, Section 2.1.3 shows how related methods can be used to connect named entities with databases, so that extracted relations can be grouped.

[3] **I**nside, **O**utside and **B**eginning.

2.1.1 Approaches

While there are many different approaches to NER, the majority of methods can be assigned to one of four main groups. Yadav and Bethard (2019) categorise these as follows. *Knowledge-based* systems rely on human knowledge in the form of lexical resources, such as lists and dictionaries, as well as knowledge that is specific to the target domain. Next, there are *unsupervised* systems, which infer knowledge from very small seed amounts of data, taking into account specific textual features. They stand in contrast to *supervised* systems that require large amounts of annotated textual data and specifically engineered machine learning models. It should be noted that there is an extra category of approaches that can be considered *hybrid*, usually combining knowledge-based systems with machine learning-based systems. Finally, there are *neural network* systems that learn from large amounts of vectorised data at different linguistic levels, such as character, word or sentence levels. Although neural networks are supervised, they are listed separately here because the method of implementation differs greatly.

Knowledge-based Approaches In general, knowledge-based systems use common NLP components, such as a tokeniser, sentence splitter, POS tagger and semantic taggers. A pipeline of these components is then used in conjunction with other resources, such as gazetteers, databases or sets of domain-specific rules (Maynard, 2003). Gazetteers are lists containing proper names, organisation names, location names, as well as indicative words like salutations, professions, geographical words, company name indicators, and so on. Databases are similar

to gazetteers in this sense, although much larger and usually compiled for different purposes. Sets of rules define certain patterns that allow systems to detect NEs. These systems can also detect entities by matching strings against a gazetteer or database, applying rules that could pertain to the shape or certain part-of-speech sequences (also referred to as feature-engineering). Yadav and Bethard (2019) highlight that the advantage of these approaches is high precision due to their specific dictionaries, which in turn causes low recall because these dictionaries are generally incomplete and language-specific. In addition, it can be very costly in terms of human effort to create rules and lexicons for a specific domain and human errors can lead to incomplete resources.

Unsupervised Approaches In the case of unsupervised systems, rules are used to extract a small set of examples in order to learn to classify unseen entities. The amount of textual data required here is minimal, commonly collected using a handful of simple rules, allowing the system to automatically extend as it trawls collections of unlabelled textual data for further examples, a process also referred to as *bootstrapping*. For instance, Collins and Singer (1999) describe an early approach that utilises this technique, with algorithms adapted from supervised learning approaches. On an evaluation set compiled by the authors, the system achieves reported accuracies of up to 91%, depending on the learning approach. Other approaches have involved pattern extractors (Etzioni et al., 2005) as well as automatic gazetteer compilation (Nadeau et al., 2006). The latter approach achieves an average F1-score of 69% on the MUC-7 dataset. A more recent

unsupervised approach is presented by Zhang and Elhadad (2013) who used inverse document frequency with shallow syntactic parsing to produce labels, with reported F1-scores of 53 % on biological, and 69.5% on medical data.

Supervised Approaches The main idea behind supervised systems is the classification of data by learning from large amounts of labelled data. The data is usually annotated manually by humans, or automatically in semi-supervised approaches. Once a system has been trained on annotated data, the system can then classify unseen and unlabelled data. Common machine learning architectures used for these approaches include Conditional Random Fields (CRF), Support Vector Machines (SVM), Hidden Markov Models (HMM) and decision trees.

McCallum and Li (2003) used CRFs and report F1-scores on the CoNLL-2003 dataset of 84% and 68% for English and German NER, respectively. These results were about average for the task, although not far off the winning system, which used an ensemble of classifiers, scoring 88% and 72%, respectively (Florian et al., 2003)). Furthermore, Kazama and Torisawa (2007) describe how the structure of Wikipedia can be exploited to extract entities and their respective tags. This allows for specialised dictionaries to be compiled, which can then be used to create a new feature for the classifier. In comparison to McCallum and Li (2003), the authors report their best F1-score for English to be 89%. Zhou and Su (2002) report F1-scores of 96% and 94% on the MUC-6 and MUC-7 datasets using HMMs. Li et al. (2005) report an F1-score of 88% on the CoNLL-2003 dataset using an SVM-based system.

Systems combining supervised and unsupervised methods have also been proposed. For example, Neelakantan and Collins (2014) present their approach to automatically creating dictionaries for NER in the biomedical domain. The authors show that using dictionaries as a feature in ML-based taggers can result in performance enhancements, although this is usually at the cost of human effort. Therefore, the authors automatically compile dictionaries from a specialised biomedical corpus to enhance a CRF-based tagger. The authors report F1-scores of 62% and 48% for two separate biomedical corpora.

Neural Approaches Although technically these systems could be classed as *supervised*, neural systems differ in terms of implementation and the amount of data used for training. With these approaches, vast amounts of data are used to automatically train a system that learns to classify by analysing patterns and adjusting weights. A crucial part is played by the system architecture, which determines how certain inputs are calculated. In general, these approaches make use of embeddings, which vectorize textual data, in place of other features for classification. Embeddings have been obtained at various levels, including the character, word and context-levels.

Semi-supervised approaches using word embeddings (Mikolov et al., 2013) have achieved up to 91% on the English CoNLL dataset (Agerri and Rigau, 2016), which was considered SOTA at the time. One successful neural architecture has been the Long Short-Term Memory (LSTM), which is able to control the degree to which past inputs are utilised for the current input. Both Huang et al. (2015) and

Lample et al. (2016) introduced different variants of this architecture for sequence tagging, with reported F1-scores of 84% and 90%, respectively, on the CoNLL-2003 dataset. More recently, a lot of success has been achieved with the T5 (Wang et al., 2022) and Transformers architectures (Devlin et al., 2019). The latter has achieved the best result for CoNLL-2003 at the time of writing at 94% (Wang et al., 2020), and is also used for the experiments in later chapters of this thesis.

2.1.2 Multilingual Named Entity Recognition

The need to process texts in languages other than English has long brought about systems that not only focus on a specific language, but rather adopt a multilingual approach. Although research has recently started shifting from rule-based and feature-engineered systems to machine learning based systems, this does not mean that the former type of systems have been completely abandoned. Nevertheless, the question as to how easily rule-based and feature-engineered systems are portable to other languages still remains, highlighted by the following case:

Maynard (2003) presents a multilingual and multi-source system that relies on various NLP components and a set of rules. In addition, it uses a so-called switching controller that detects certain features of the document that is being processed and determines the components that should be used. Furthermore, it is demonstrated that the system is easily adaptable to other languages, with only language-dependent components needing to be replaced. The languages are Arabic, Bulgarian, Cebuano, Chinese, French, German, Hindi, Romanian, and Russian. This is said to be the advantage over machine learning based systems, as

only smaller parts of the system would have to be changed.

The approach described above is a good example of a clean rule-based system, that still works today. This kind of simplicity can not be replicated with machine learning based approaches. However, it is not clear whether the system can actually be ported to other languages so easily. Especially when thinking of languages with other writing systems (e.g. Hindi, Russian) many of the modules would have to be replaced, including the resources, which could prove difficult to do. Moreover, languages with writing systems that differ a lot (e.g. Japanese, Chinese) would require even more effort to implement. This raises the question whether it will take longer to adapt the system or to collect the annotated data necessary for a machine learning system.

Addressing the necessity of annotated data for NER, Nothman et al. (2013) make use of the multilingual resource Wikipedia in order to automatically create an NER corpus for multiple languages. The approach makes use of links between different Wikipedia articles, in order to gather and label entities. The approach produces silver-standard data, which is not as precise as gold-standard data[4]. However, due to the much larger size than gold-standard datasets, the authors claim that higher F1-scores can still be achieved. The approach was used for nine different languages: Dutch, English, French, German, Italian, Polish, Portuguese, Russian, and Spanish. For these languages, respectively, the authors report the following coarse-grained[5] F1-scores on a gold-standard dataset

[4]Gold-standard data is the best available annotated data available, usually annotated manually by humans. Silver-standard data is not as precise, because is is usually compiled semi-automatically, i.e. with less human oversight.

[5]Coarse-grained evaluation allows for less strict matches, usually in cases where partial

compiled from Wikipedia pages: 92.9%, 94.6%, 93.8%, 94%, 93.8%, 93.8%, 93.1%, 93.3%, 93.9%. The system is said to be ideally suited for resource-scarce languages. Moreover, the authors also analyse other corpora used for NER training and compare the results to their own corpus. This approach proves how much can be learned from the linked data within Wikipedia. Most importantly, the approach is shown to be useful when taking many different languages into account. This approach highlights that Wikipedia is an extremely useful resource for automatically creating datasets and addressing multilingual problems, with low-resource languages benefiting the most from this.

It is clear that multilingual NER systems rely on large amounts of language-specific resources, be this in the form of annotated data or precise pre-processing systems. Addressing this problem, Rahimi et al. (2019) propose a new method of large-scale cross-lingual transfer learning. The approach, which aims to transfer knowledge from high-resource to low-resource languages, is based on inferring information from multiple unreliable sources or outputs. Two methods are proposed to make use of tiny annotated datasets, the first an estimation of reliability and the second selection and fine-tuning of an explicit model. While the authors conclude that cross-lingual transfer is not a simple out-of-the-box solution, they do present two multilingual transfer models that are based on small amounts of annotated data and outperform state-of-the-art unsupervised, single-source transfer models. Some of the components are trained on Wikipedia data, such as the named entity tagger and linker (Pan et al., 2017). Moreover, by

matches of a longer match would still be correct.

transposing the information gained from Wikipedia markup, the authors claim that the system is able to aggregate information not only from English, but also Russian and Ukrainian texts. Importantly, the authors highlight the fact that they did not use machine translation to cover different languages, because of the low-quality for the addressed language pairs.

With the rise in use of the aforementioned neural models for many NLP applications, much recent success has been achieved with cross-lingual transfer learning models (Baumann, 2019; Xie et al., 2018). These models are able to map words of multiple languages into similar semantic spaces, and can therefore transfer what they learn from one language to another. Being able to learn in one language which may have many resources available, such as English, and then to apply the model to a lower resource language alleviates the need for costly human annotation (Ji et al., 2020; Huang et al., 2021).

2.1.3 Named Entity Linking

Once named entities in a text have been detected these can be resolved and linked to entries in knowledge bases, in order to establish which words and phrases refer to the same entity, e.g. person, organisation, location and so on (Hachey et al., 2013; Jurafsky and Martin, 2018). The relationship between entities and mentions may be one-to-many and many-to-one. As this makes it potentially difficult to track entities, NEL can be useful in these types of situation. According to Hachey et al. (2013), NEL is said to allow computing "with direct reference to people, places and organisations, rather than potentially ambiguous or redundant character

strings". Figure 2.2 shows a visualisation of how entities detected in a sentence or phrase would be linked to knowledge base entries.

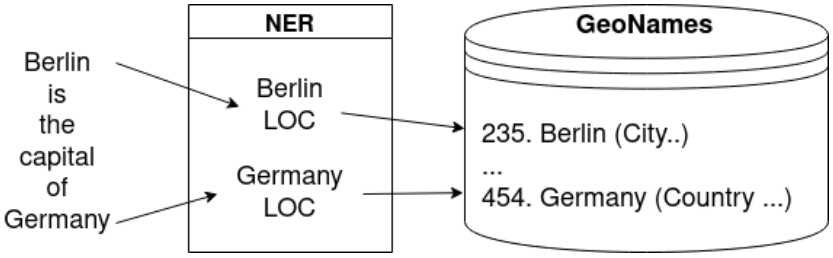

Figure 2.2: Example of entity linking: Locations to GeoNames database.

Two tasks that are very similar to NEL are *word sense disambiguation* (WSD) and *wikification*, which differ in terms of approach and scope. WSD is the task of matching words to their correct meaning or dictionary entries. Wikification on the other hand is the task of linking mentions in text to corresponding Wikipedia articles, therefore establishing whether something exists or not, and allowing use of further Wikipedia metadata, as well as the links between Wikipedia pages. NEL is often applied with databases or knowledge-bases (KB) that are incomplete and where expansion of the KB is the goal (Hachey et al., 2013).

Toponym resolution (TR), a variant of NEL, involves disambiguating and resolving locations named in texts. TR is said to be non-trivial and closely related to NER (Piskorski and Yangarber, 2013) and also NEL, to some extent. While the process of NER would detect and tag proper names as locations with some success, there are further challenges to face when working with locations. Firstly, not every location mention in a text refers to an actual location. This becomes more clear when looking at the example *Birmingham City Council*, where the

location *Birmingham* forms part of the name of a local authority in the UK and does not directly refer to a location (although it does denote a general area the authority is responsible for). Although this is an example of a named entity, and should be tagged as such by means of NER, cases like this can be overlooked by imprecise NER programs. Second, it is very common that location names are ambiguous, i.e. referring to possibly more than one location. For example, *Birmingham* can refer to a city in the Midlands region of the United Kingdom or a city in the US state of Alabama, and possibly others. Finally, having overcome the previously mentioned obstacles, a location should be resolved to a specific location, usually described in the form of coordinates.

An overview of various NEL systems is presented by Hachey et al. (2013), who also test using Wikipedia as a KB, and evaluate three NEL systems. The authors claim that Wikipedia was chosen due to many applications and shared tasks of the time being focused on it. The authors particularly emphasise that no thorough analysis on the multiple components of increasingly complex NEL systems had been carried out until that time. Therefore, three different systems were implemented and evaluated, with particular emphasis on the different components of NEL systems, including the *extractor*, *searcher* and *disambiguator*. While the extractor detects and processes named entity mentions, the searcher generates candidate entities from a given KB and the disambiguator selects the best suited entity link. Overall, Hachey et al. (2013) find that while a lot of attention is often paid to the disambiguator, due to its algorithmic nature, the performance of a system is largely dependent on the searcher component. It

is also mentioned that newer approaches at the time were starting to incorporate more machine learning based ranking methods for the searcher component.

The article by Hachey et al. (2013) describes the differences between NEL and other tasks extremely well, as well as providing a good overview of methods at the time. What is particularly useful is the in-depth analysis of the various components of NEL systems, which allows for good insights into how NEL systems work and can be improved. It is interesting that Wikipedia was considered a suitable KB at the time, although it has much looser structure than a resource description framework (RDF) database such as Wikidata or DBpedia. Nowadays, it makes more sense to use one of these latter resources as opposed to Wikipedia.

2.2 Information Extraction

IE is described as the task of detecting unstructured information relating to entities in texts and then converting it into a structured form. The extracted information should represent key information about the entities from the target text, including relations between entities (Azzam et al., 1999). As the extracted information is structured, it can be stored in databases for further use, including mining and linking data and summarisation in natural language (Azzam et al., 1999). While NER plays an important part in this thesis, IE is the central topic, and provides the methodologies that are central to the research presented throughout this thesis.

There are many parts that work together when extracting information, including event extraction and template filling (Jurafsky and Martin, 2018), however, the most important is *relation extraction* (RE). Relation extraction refers

to the the task of detecting certain semantic relations between previously detected entities (see Section 2.1) in the text. These include but are not limited to *part-whole*, *child-of*, *employed-by* and so on (Jurafsky and Martin, 2018). The relations are commonly expressed by certain verbs and prepositions, although they could also be expressed indirectly by adjectives and nouns, as well as other features of text. Example 2.2 lists a selection of the described relations. While the first and second items show a simple and a double-anchored relation, respectively, the last item is an example that is typical amongst lexicon type text (such as is common on Wikipedia) and is not indicated by a verb.

(1) Gordon Freeman works at the Black Mesa Research Facility.
→ (Gordon Freeman, work, Black Mesa Research Facility)

(2) Having lived in Lisbon for many years, Anne is now moving to Egypt.
→ (Anne, move, Lisbon to Egypt)

(3) Michael Scott (born on March 15, 1964 in Scranton, Pennsylvania) is a manager at Dunder Mifflin.
→ (Michael Scott, born, March 15 1964)

Example 2.2: Examples of different entity-relation pairs.

A defining aspect of IE has been the conferences specifically aimed at this task. The conferences, and their corresponding series evaluations, have not only been useful for gaining a better overview of the field, but also for the supply of annotated data (Grishman, 2019). According to Grishman (2019) the most notable of these conference series were the Message Understanding Conferences (MUC) (Sundheim, 1995), the Automatic Content Extraction (ACE)

conferences (Doddington et al., 2004) and the Knowledge Base Population (KBP) evaluation series (Ji and Grishman, 2011), with each providing a unique dataset and evaluation aspect. The MUCs are said to have established IE in the NLP field and provided a first dataset and evaluation, as well as highlighting the need for coreference resolution. While the IE performed here was very much focused on a specific domain and template-filling, the ACE conferences first brought about a more domain independent task. The KBP series increased the scale of the data to be processed, as well as providing a unified database of relations and weakly annotated data, in turn promoting the development of semi-supervised methods.

2.2.1 Approaches

As with NER, approaches to IE can be broadly split into four categories, namely *knowledge-based*, *unsupervised learning-based*, *supervised learning-based* and *deep learning-based*. Furthermore, the category of *semi-supervised* approaches can be added, which directly influenced the new methods proposed in this thesis. It is also important to note that each of these approach categories has contributed in some way to the approach described later.

Beyond these categories, approaches are often differentiated in terms of the type or structure of the information that is extracted. The term IE can either refer to the broad topic of extracting information, encompassing a multitude of methods and approaches, including the ones previously mentioned in this chapter. On the other hand, it can also refer to a more traditional approach of extracting information, where the information that is to be extracted is defined beforehand.

In contrast to this stands *open information extraction* (OIE), which is aimed at being more domain-independent than traditional IE, therefore making it suitable for extraction tasks from large and varied collections of text. Where an approach is focused on OIE, there is usually no limit on the type and number of different relations that are extracted, whereas other approaches may be limited in this sense.

Knowledge-based Approaches Similar to NER, knowledge-based systems require human effort in order to define rules or patterns to use for extracting information. For instance, Hearst (1992) used five lexico-syntactic patterns to extract hyponyms from different text types. This approach, which essentially extracts noun phrases that are connected by words that suggest a hypernym-hyponym[6] relation between the nouns, was low-cost even for systems at the time. However, even though both approaches were very successful at the time, Hearst (1992) demonstrates that these patterns do not always work as simply as this by example of the meronymy relation. Similar approaches have been presented by Coates-Stephens (1991), Appelt et al. (1993), and Piskorski et al. (2004). Knowledge-based approaches are still used today as well, such as Falke et al. (2016) and Del Corro and Gemulla (2013).

Unsupervised Approaches Systems that are unsupervised may use a small amount of annotated data for validation, but generally rely on patterns that can

[6]According to the Oxford dictionary, a hypernym is "a word with a general meaning that includes the meanings of other particular words, for example fruit is the hypernym of apple, orange, etc." and a hyponym is "a word with a particular meaning that is included in the meaning of a more general word, for example dog and cat are hyponyms of animal"

be applied by algorithms to extract information without learning from annotated data. Among the first of these approaches was Brin (1998). A later approach is described by Bunescu and Mooney (2005), who exploit the structure of dependency parse graphs. By making the assumption that the shortest path between two entities usually indicates the relation between them, the authors are able to use a shortest path kernel in order to extract information. Evaluated on the ACE dataset, the authors report precision scores between 63% to 71%, at recall scores of 26% to 43%. Etzioni et al. (2005) on the other hand, present the KnowItAll system that combines three different ways to learn extractions. The combination of pattern learning, sub-class extraction and list extraction is said to achieve a precision of 90% on a custom dataset compiled from Google.

Banko et al. (2007) introduce the concept of OIE, which as described above, is particularly suited to unsupervised approaches. As such, the authors present TextRunner, a system that labels its own training data and then trains a Naïve Bayes classifier. It is difficult to compare to traditional IE methods, due to the difference in end goal. Nevertheless, the authors compare it to KnowItAll (Etzioni et al., 2005), a SOTA system of the time, by means of selecting a limited set of relations and comparing the system outputs. Overall, the authors report an error rate of 12% for their system, with 18% for the other.

Supervised Approaches Instead of using human knowledge to create patterns, rules or gazetteers, supervised systems are designed to use machine learning algorithms that learn from text examples. These approaches often also require

a great deal of human effort, since they rely on a large amount of annotated data from which a specified algorithm learns how to extract or classify information, as well as feature-engineering. An early example of a supervised approach is presented by Freitag (1998), who use a so-called relational learning algorithm to classify relevant information at the token-level. The top-down approach induces rules to fill templates with the tokens labelled according to the templates. The authors also set up their approach to work on HTML, making it possible to extract directly from websites. Newer supervised approaches use machine learning classifiers, such as SVMs, dependency trees and kernel approaches. Various features have been examined for these tasks, including words, overlap, dependencies, parse tree and entity type relating to the information pairs (Kambhatla, 2004; Jiang and Zhai, 2007; Zhao and Grishman, 2005).

Neural Network Approaches Again, although neural network-based systems are technically supervised, they stand apart from other approaches due to their architecture, which is a distinguishing factor for neural networks. One of the first successful architectures in relation extraction was the *Convolutional Neural Network* (CNN). Zeng et al. (2014) made use of the CNN architecture, with the aim of exploiting the fact that neural networks generally do not require complicated pre-processing of input data. The system assigns each word a vector, which represents lexical features, as well as other sentence level features. The authors evaluate their system using the SemEval-2010 Task 8 dataset, which is freely available and contains 10,717 annotated examples across 9

relationships, and an undirected *Other* class. The relationships include *Cause-Effect*, *Component-Whole* and *Entity-Origin*. The reported F1-score of 82.7% compares well to other systems compared in the paper.

Cui et al. (2018) describe a state-of-the art neural approach to Open IE that is based on an encoder-decoder framework. The system, implemented using the open source OpenNMT framework, learns from the output of a rule-based Open IE system. Using only output triples with a confidence score of above 90%, the proposed neural system is able to learn broader patterns, mainly due to the more flexible context embeddings when compared to the more rigid patterns of other Open IE systems that rely on the output of other NLP systems, such as POS-tagging and dependency parsing. The evaluation of the system in terms of performance and computational cost ranks it similarly to other OIE systems, such as OLLIE (Schmitz et al., 2012), TextRunner (Yates et al., 2007) and ClausIE (Del Corro and Gemulla, 2013), on a benchmark dataset compiled from Wikipedia by the authors. Han and Wang (2021) evaluate a transformer-based system with other OIE systems (ClausIE, OLLIE, Stanford) and show that transformers outperforms the other systems, with high precision scores. Jin et al. (2021) have successfully used neural models for their CogIE system, aimed at being an all-in-one solution for IE and KBP.

IE is often used for entity linking, as described in Section 2.1.3. Of note is an approach described by Bordes et al. (2013), who treat IE as a translation problem[7] and therefore developed a model that treats relationships between

[7] A translation from a detected relation to a row in the IE database

entities as translations. Using this model, the authors claim to model relations in a KB with the aim of being able to add links automatically with low effort. A similar approach described by Xu and Barbosa (2019), which combines the representations of language and knowledge for relation extraction in a novel framework. The approach combines the representations of a sentence and the relation in a knowledge base. By then training this novel framework by means of distant supervision, Xu and Barbosa (2019) are able to significantly improve the state-of-the-art in relation extraction. This is also demonstrated in a short case study, where it is shown that the novel representations are able to convey implicit relations that were previously not extractable by conventional methods. The approach is similar to BERT context embeddings, in that it combines knowledge (i.e. context) with language representations.

Semi-Supervised Approaches Approaches that attempt to combine the best aspects of both supervised and unsupervised learning are often referred to as semi-supervised approaches, with variants of this referred to as *distant supervised*, *weakly supervised* or *hybrid systems*. This often includes attempts to automatically annotate data or to avoid the use of manually annotated data but still validating the approach by defining some base assumptions that allow the inference of correct and incorrect classifications. Some approaches in this area are referred to as *weakly-supervised* approaches, such as those described by Yangarber (2003) and Stevenson and Greenwood (2006). These approaches require a small set of seed patterns, which are then automatically extended by an

algorithm. This is achieved by using a set of weakly labelled data, usually in the form of relevant and non-relevant documents, from which information is extracted using the seed patterns. New patterns are then inferred from information that is similarly structured.

Mintz et al. (2009) describe their approach based on *distant supervision* which also aims to accomplish the combination of supervised and unsupervised approaches. By using a small set of seed relation triples, the authors then use the entities in these to gather sentences where both occur. From these, relations can be extracted connecting the two entities and the approach is repeated over and over, resulting in a very large dataset of (weakly) annotated data. In this approach, Freebase[8] is used as a semantic database, and text is extracted from tokenised Wikipedia text. The classifier itself is a logistic regression classifier that takes lexical features, syntactic features and a combination of both. Mintz et al. (2009) carry out experiments to determine the most useful features, which they claim to be a combination of both lexical and syntactic features. Overall, the authors extract 10,000 instances of 102 relations at a precision of 67.6%.

A problem that arises in distant supervision is incorrect labelling of training data. This is due to the underlying assumption that two entities with a relation in a database always have the same relation in any sentence that they appear in. Although this is said to be a non-trivial problem, many attempts at alleviating it have been made. Lin et al. (2016) propose a neural-based approach that uses

[8]Freebase was a knowledge base that was moved to Wikidata in 2014. More information about Wikidata is presented in Section 2.3

so-called selective attention to weight certain examples in the training data over others. Essentially, the network chooses which sets in the training data are better examples of a relation and which are weaker examples of a relation. On the other hand, Wu and He (2019) propose a solution that uses two extra layers in the neural extraction algorithm in order to identify noisy sentences and remove them from the training data. In essence, the approach attempts to learn from labels that better represent the data, rather than all labels.

It is important to take note of the approaches within Open IE, which are becoming more popular, especially with the wide-spread use of neural architectures. However, a drawback of Open IE is the fact that the extraction of information is not limited. These systems are inherently targeted at any kind of relation, although there could potentially be some kind of filtering mechanism added on afterwards. Nevertheless, systems that can be limited in terms of relation types would be better suited when specific relations are targeted. In the case of biographical information, this is exactly the case, as it should only deal with information about human beings. In addition, if the end goal is to compile biographies, more limitations on the exact relations to be extracted might be placed. As a further consideration, the evaluation of these kinds of system can be challenging, since any conceivable relation might be detected, which may not be a feasible annotation task for manual annotation. Therefore, it is important to show that Open IE methods do exist and can be trained easily, however, they are not always suitable for every application.

Other approaches have also been suggested for improving distant supervision

(see Sections 4.1 and 5.1). Moreover, this thesis will present the concept of guided distant supervision, involving a mechanism to ensure relations are labelled more precisely. Chapters 4 and 5 will provide an evaluation of both monolingual and multilingual approaches to this concept.

2.2.2 Multilingual Information Extraction

Multilingual information extraction is a key area of research for this thesis. As such, it is important to understand that multilingual IE has been a topic of interest since at least 1999 (see following paragraph) and this section will also highlight the many approaches and obstacles with multilingual IE.

Azzam et al. (1999) describe a system to achieve multilingual information extraction, while also investigating problematic aspects of the task. Highlighting the growth of the World Wide Web and electronic texts, the authors explain that extracting information from these texts is limited by the number of languages that a user understands, therefore underlining the need for multilingual information extraction systems. It is argued that a general IE system with a language-independent domain model used to produce a discourse model of a text would allow for information to be extracted in the form of concepts. These could in turn produce templates with English texts, while any further languages would require syntactic and semantic analysers as well as mappings to be constructed between a given domain model and a lexicon. Azzam et al. (1999) continue with this approach, as opposed to other approaches using machine translation systems, as it would be easier to integrate new languages due to the lack of

having to consider language pairs. In order to achieve this, the approach uses a semantic representation called QLF (quasi-logical form), and which the authors claim to be uniform across languages although not language-independent itself. This representation then needs to be mapped to each language, which is also where the biggest problems associated with this approach are said to arise. The authors cite problems such as lexical gaps where one word concepts may be realised in more than one in other languages, word sense disambiguation and unknown words. Overall, Azzam et al. (1999) report the results for 20 English-French parallel texts, and claim the system to be comparable to MUC-6 systems of the time period. Although the authors also added Spanish, their main conclusion is that this kind of approach will always rely on a robust domain model, which is not always available. Nevertheless, the language-independent approach has influenced the methodology discussed Chapter 5 of this thesis.

Annotated Multilingual Data The approach outlined above has been echoed more recently by Embley et al. (2011), who describe an information extraction and semantic search system that relies on multilingual ontologies to work across multiple languages. In addition to machine translation and some human input, ontologies that can be mapped to different languages are used for a given area of interest. The assumption is that as structures of an ontology would vary only slightly across languages, words and phrases could be converted using lexicons. Other conversions, such as measurements, currencies etc. can be done using web services or conversion metrics. The authors also describe how a pay-as-you-go

scheme could improve the localisation of mappings, using longer comments that could help a person creating very specific ontologies. The approach is evaluated on various cross-lingual combinations with English, Japanese and Chinese using a custom dataset compiled from restaurant recommendation web services and the authors report the lowest F1-scores of 73% (English to Japanese source), with the highest at 94% (Japanese to English).

Gamallo and Garcia (2015) describe a multilingual approach to OIE that does not require training data. The approach relies on dependency parsed input and a set of rules to detect argument structures. By means of detecting these structures, the system is able to output the subject and the object, as well as the attribute and preposition that relate the two to each other. Following this step, a further set of rules that act like filters is applied in order to detect trivial information. Since the system does not require any training data, it is multilingual in principle, as long as a robust dependency parser is available for a given language. This is indicated by the results of the experiments reported by the authors, although they are not given explicitly. While English performs at state-of-the art levels for OIE, Spanish and Portuguese suffer from not having such robust dependency parsers. In addition, it is mentioned that the rules to detect the argument structure are based on the structure of Romance languages, meaning that the approach may not be truly multilingual, but in fact limited to languages that share the same structure.

Machine Translation Stressing that state-of-the-art relation extraction systems are only available for English, Faruqui and Kumar (2015) present a pipeline to

develop open relation extraction systems for any given source language. The described methodology relies on machine translation of texts in a source language to English. This step is followed by word alignment of the sentences, which can then be used for RE. The authors use the English extraction system OLLIE (Schmitz et al., 2012), and Google Translate for translation. Once extracted, the relations are then mapped back to the source language sentence. A significant finding of the paper relates to the use of the machine translation automatic evaluation metric BLEU, which the authors use to indicate confidence in the relation projection. It was found that higher BLEU scores indicated more extracted relations, although not in all cases. Russian is said to suffer from low extraction rates even with higher BLEU scores, with the opposite being true for French. Overall, the indication is that high quality MT is needed for relation extraction, which was more difficult with statistical machine translation (used here) and will be better with current neural machine translation, with the drawback that this method may not work for certain language pairs. Additionally, the problem of low-resource language processing is not solved.

Cross- and Multilingual Embeddings Many current state-of-the-art NLP systems use text embeddings to achieve their respective task. These embeddings, whether character, word or context based, are usually created for one language and are therefore monolingual. There are two main approaches for including multiple languages in embeddings: multilingual embeddings and cross-lingual embeddings. The former approach combines monolingual embeddings of various

languages into one, usually using bilingual dictionaries to aid the process. The latter approach on the other hand focuses on building the word embeddings of multiple languages in the same vector space. This in turn results in words of different languages being mapped in similar vector spaces, showing semantic overlap (Karthikeyan et al., 2020).

Cross-lingual models were first introduced by Conneau et al. (2018), with the XLM-RoBERTA model. These models can be used in so called zero-shot transfer learning approaches. This approach can be trained on large English datasets, and then used with other languages. The embeddings place words from two or more languages with similar meanings into similar vector spaces, allowing for direct transfer of knowledge. This means that a system using cross-lingual embeddings can theoretically be trained in one resource rich language for a task, and then be used for all languages that have available cross-lingual embeddings. In combination with the recently highly successful transformers architecture for neural networks, very competitive evaluation scores can be achieved. Making use of such an approach, Ni et al. (2020) describe their multilingual relation extraction system using zero-shot transfer learning. With the XLM-RoBERTA embeddings, the authors achieve 49% and 65% for Arabic and Chinese, respectively, on the ACE dataset. Overall, the reported scores are up to 10% higher than other systems on the established ACE (Walker et al., 2006) and KLUE (Park et al., 2021) datasets. The system is tested on English to Arabic, Chinese, German, Spanish, Italian, Japanese and Portuguese.

Similarly, Subburathinam et al. (2020) demonstrate that structural

representations of relation and event arguments can be transferred to other languages without training. The system described by the authors uses graph-convolutional networks, and is tested on English, Chinese and Arabic. An interesting experiment that is carried out is the training of the approach on not just English, but also the other languages and language combinations. The system achieves between 43% and 63% F1-scores depending on the paired languages.

This section has demonstrated that there is a long-running need for multilingual methods in IE, and how different methods have been applied to this problem. It seems clear that neural architectures using cross-lingual embeddings achieve good results. However, with the increasing quality of machine translation methods, it is also still questionable whether this could serve as a way to use accurate approaches for English in other languages. Both cross-lingual and machine translation approaches will be investigated in Chapter 5.

2.3 Biographical Information Resources

While the previous sections have dealt with processing information, this section presents some widely used sources for biographical information, as well as some projects from the field of *Digital Humanities* (DH) that deal with digitalising biographical information. Three sources of information are central to the methods presented in this thesis, while the DH projects have validated and inspired them.

Wikipedia Wikipedia[9] is a free online encyclopedia that contains vast amounts of information about almost any given topic (Ayers et al., 2008). Information is grouped in articles, which describe a given topic. While articles are available in English, many are available across multiple languages (although they are not parallel). Due to its size and availability, Wikipedia has become an extremely popular online source of information for biographies (Biadsy et al., 2008). As such, it also serves as a source of textual data for the research presented in Chapters 3 to 5. It should be pointed out that the information contained within Wikipedia can sometimes be incorrect, as it is a collaborative resource, written by many online users. As this requires verification by other humans, incorrect information can be overlooked for long periods of time. However, for the purposes of creating annotated datasets (Chapters 4 and 5) the aim is to find patterns in language that express certain relations (or similar), rather than extracting factually correct information (this is the task of Wikipedia and its derivatives itself).

Wikidata Wikidata[10] is described as a *"free, collaborative, multilingual, secondary database" that "provides support for Wikipedia [...]"* (Vrandei and Krötzsch, 2014). This effectively means that Wikidata can be considered to contain the information in Wikipedia, but in a structured and queryable way. Processing Wikipedia articles is time-consuming and can be imprecise if targeting specific information, therefore, the structured information available in Wikidata makes targeting information in Wikipedia much easier.

[9] https://wikipedia.org
[10] https://wikidata.org/wiki/Wikidata:Main_Page

Pantheon According to its creators, *"Pantheon [is] focused on biographies with a presence in 15 different languages in Wikipedia"* and consists of roughly 85,000 entries (Yu et al., 2016). While it was initially created mostly by hand, its later iterations have used a classifier to determine and extract further entries. One particular characteristic of this dataset is that each article has to contain unambiguous links to the corresponding Wikipedia and Wikidata pages. This allows identification of the corresponding articles in Wikipedia, which can then be extracted. While this could be done just using Wikidata, Pantheon has been (at least partly) manually verified. Because Pantheon only includes persons whose articles are available in 15 different languages, this ensures that a person is somewhat well-known, in turn making a longer Wikipedia article more likely.

Digitalising Existing Resources The field of DH is generally concerned with the computational treatment of humanities subjects, such as history or languages, often including methods for digitising texts and generating appropriate metrics. These tasks involve NLP technology to some extent, for instance *optical character recognition* (OCR) to digitise texts. It could be said, however, that the use of more advanced NLP methods is not as widespread, for example making use of machine translation to translate older texts, or information extraction systems to automatically compile databases of historical information. Therefore, it stands to reason that interdisciplinary research in this area could potentially be highly productive, which is demonstrated by some existing research that involves these more advanced NLP technologies. This is particularly the case when looking at

some of the work done with regard to biographical information.

Using IE methods for biographical information has been suggested by Brin (1998), where the focus was on describing a person with extracted information. Although this kind of focus is typical of IE, it has also been used in the context of summarisation, such as the Document Understanding Conferences (DUC) taking place from 2001 to 2007, later Text Analysis Conferences (TAC) from 2008 (Dang et al., 2007; Dang, Owczarzak, et al., 2008). Zhou et al. (2004) describe a multi-document summarisation system for biographies that was evaluated at the DUC2004 shared task. To develop their system, the authors analysed a set of 130 different biographies about 12 people. Nine common elements[11] among the biographies were found. The texts were then annotated at clause level according to the nine common elements previously found, as well as a none class. The authors test three different classifiers on 10-class (the previously described common elements) and two class (binary biographical/non-biographical) tasks using different feature combinations. Similarly, Biadsy et al. (2008) present an unsupervised multi-document summariser for biographical information that is similar to Zhou et al. (2004). As input the system assumes a selection of documents relating to a certain person, where named entities have been tagged and coreference resolved. Sentences within each document are then filtered using a sentence classifier that detects biographical sentences. The classifier is trained on class-based and lexical features, using an automatically compiled corpus of

[11]The elements are: bio (info on birth and death), fame factor, personality, personal, social, education, nationality, scandal, and work.

biographies from Wikipedia for biographical sentences and the general news corpus TDT4 for non-biographical sentences.

In the last decade, large biographical databases have become available in a number of languages (Reinert and Ebneth, 2017). This is the case with many purely online data sources and sources which have been digitised such as the Slovenian Biography (Erjavec et al., 2018), the Deutsche Biographie (Reinert et al., 2015), and the Austrian Biographical Dictionary (ABD) (Wissenschaften, 2013) (also referred to as the ÖBL). As a result of the large amounts of data available, researchers have been exploring ways to process this data using computational methods. In addition to Wikipedia, DBpedia[12] and Wikidata provide structured information based on Wikipedia and have been used to generate biography summaries using natural language generation methods (Moussallem et al., 2018; Chisholm et al., 2017). A number of projects such as the APIS project at the Austrian Academy of Sciences (AAS) (Schlögl and Lejtovicz, 2018) and the Dutch BiographyNet project (Fokkens et al., 2014) have addressed the problem of retrieval of information from biographical encyclopedias and dictionaries.

Fokkens et al. (2014) present the BiographyNet project, which is aimed at enhancing the existing Biography Portal of the Netherlands by using NLP methods and joint work with historians. Firstly, the authors present the Biography Portal of the Netherlands, which consists of around 125,000 entries about 76,000 people. It is added that BiographyNet is aimed at making the portal more accessible by allowing more detailed queries. In order to achieve this, the authors

[12] https://wiki.dbpedia.org

explain that NLP methods such as entity linking, information extraction and other tasks should be used. However, as an interdisciplinary project much importance is placed on the documentation, explanation and transparency of the methods used. It is argued that historians base their decisions and information on a multitude of parts of information from various sources, so-called "building blocks" for each entry. Therefore, information extracted by (automatic) NLP methods should mirror this structure. In order to do so, raising awareness with historians and computational linguists alike would be crucial, as both sides have to understand how data obtained by heuristics or machine learning methods can affect extracted information, as well as introducing bias in the latter case. Finally, Fokkens et al. (2014) present a basic information extraction system developed in the scope of the project. The system, which is said to be inspired by systems such as Kylin (Wu and Weld, 2010), TextRunner (Yates et al., 2007) and Open IE (Etzioni et al., 2011), was trained to identify sentences containing relevant information, and in a next step to extract this information. To this end, the authors automatically created a labelled training corpus. Overall, the authors report lower accuracy results than expected, but highlight that the concept of introducing NLP methods to aid historians in a clear and understandable way can be extremely useful.

Further research that aims to combine work with biographical information and NLP is presented by Russo et al. (2015). Here the authors propose an NLP framework in order to extract temporally anchored events from biographical texts. Having extracted the events, the tool is able to visualise these in a graph, allowing for further analysis by historians. However, it is pointed out that while the tool is

able to extract events, these are often ambiguous, uncertain or simply not events. This would require an addition to the tool to allow for the user to discard these events, which is planned for the future. The authors also plan to add a sentiment analysis component, allowing for events to be classified as positive or negative.

Many efforts have been targeted at making biographical information available, highly interactive and extendable. One such approach is described by Petram et al. (2018), who present the Huygens research infrastructure and biographical data policy, aimed at bringing all biographical information available to the institute online, and making it easily accessible. Furthermore, the aim of the project is to make it semi-automatically connectable with outside sources of data. This is said to make it possible to work with a large number of much smaller collections of biographies, which may otherwise not be possible.

A similar effort has been introduced by Schlögl and Lejtovicz (2018), who present *A Prosopographical Information System* (APIS), a project similarly aimed at making biographical data easily accessible. However, in the case of APIS, the focus is to make this data available and easy to work with for non-experts in digital technology. The APIS system therefore aims to provide a highly-customisable research environment, including NLP technology. The aim of APIS is the establishment of new methods for re-using qualitative biographical research products (i.e. encyclopedias) for quantitative research and facilitating a digital transformation process against the background of Humanities (Gruber and Wandl-Vogt, 2017). To achieve this, the project has developed a web-based, customisable virtual research environment that allows researchers to work with

programs especially designed for processing biographical texts[13]. Another goal of the APIS project is to reveal information encoded in texts such as names of persons, institutions, places, and to detect relationships between them and the person depicted in a biography. Primarily, this is in relation to the ABD project, by which APIS is third-party funded. Since 2004, the ABD has been transitioning to a digitised version and a database has been established to support the manifold editorial processes. A rich network to neighbouring analogue endeavours and close personal relationships exist, for instance the European Biography-Portal[14] (Gruber and Wandl-Vogt, 2017). In this context, the aim is to collect and make visible the lives and careers of persons who had some kind of impact in the territory of the former Austrian-Hungarian monarchy, as well as the first Republic of Austria. In order to find relevant candidates, the ABD is primarily aimed at looking beyond the "usual suspects", in order to find lesser known and less easy to find knowledge carriers, influencers and impact holders. Currently, about 18,500 biographies are available via the ABD, with more being added.

2.4 Conclusion

This chapter has presented an overview of the areas of NER, IE, and DH in relation to the NER and IE. Common methods of the subjects have been established, while presenting some generally relevant work within the areas. The related work has highlighted some aspects of all the methods mentioned that warrant

[13] https://apis.acdh.oeaw.ac.at
[14] https://biographie-portal.eu

further investigation and give rise to the research questions answered throughout the course of this thesis.

With regard to IE, it should be clear that approaches have changed considerably, most notably with the shift from being heavily reliant on multi-part pipelines and a great deal of case-specific engineering, to machine learning methods that rely on standard NLP packages and require much less engineering. The sections on NER and IE have both established that neural methods represent the SOTA today, and as such should be used for any research in the area of IE. However, it is also evident that knowledge-based systems can perform well, at the cost of much (expert) human effort. It could be argued that with the recent focus on neural methods, there is a danger that the human role shifts to the development of annotated data, therefore making these systems reliant on outside help. While this has invited approaches to reduce the need for human effort, the automatic compilation and annotation of data still warrants further research.

With regard to DH and Biographical IE, the BiographyNet project clearly highlights that NLP and DH work extremely well together, if enough attention is paid to the ideas and needs of the historians on the one hand, and the capabilities of the system and computational linguists on the other hand. In order for this process to be of use, good communication and explanations are needed, so that both sides understand how a given system works and what its potential capabilities are. In addition, it has shown how useful IE could be for enriching and even creating biographical databases. It is also clear, however, that there is much room for improvement, which includes but is not limited to the automatic creation of

a large training corpus, as well as a robust machine learning component, which would ideally be semi-supervised or unsupervised.

Looking ahead, it is clear that there is enough NLP methodology in place to facilitate the needs of DH. Using these techniques together could yield a new and low-cost approach for use in the field of DH. It just requires the right pairing of methods, such as distant supervision and neural models.

Chapter 3

Wikipedia for Biographical Information Extraction Methods

As one of the largest collaborative sources of information available at present, Wikipedia is useful as a source of textual data for research purposes. Using Wikipedia to extract text for NLP is by no means a novel idea, as is evident from the many approaches that report making use of it in some form. This indicates that it would also lend itself well in the context of gathering biographical information, as evidenced by past and current related work (Zhou et al., 2004; Toral and Mu, 2006; Kazama and Torisawa, 2007; Biadsy et al., 2008; Balasuriya et al., 2009; Hachey et al., 2011; Cheng and Roth, 2013; Zaila and Montesi, 2015; Aprosio and Tonelli, 2015; Tsai and Roth, 2016; Chisholm et al., 2017). In addition, there are a number of existing datasets that are based on extracted text from Wikipedia and subsequent annotation in some shape or form (Mintz et al., 2009; Yu et al., 2016). Based upon this evidence, it is clear that Wikipedia is a good candidate for automated text and information extraction.

The research described in this chapter addresses research question **RQ1** of the thesis, which examines aspects of biographical information extraction within the context of the A Prosopographical Information System (APIS) digital humanities project at the Austrian Academy of Sciences (AAS), which deals with the Austrian

Biographical Dictionary (ABD). The contributions[1] of this chapter are:

(1) Establishment of a method to process Wikipedia and Wikidata for the purposes of Biographical IE (Section 3.2).

(2) Provision of a dataset that is labelled in terms of biographical information per sentence (Section 3.3).

The rest of the chapter is structured as follows. First, work related to information extraction using Wikipedia is summarised in Section 3.1. This is followed by Section 3.2, which presents research carried out in the context of the APIS project, which deals with using Wikipedia for biographical information extraction purposes. Section 3.3 presents a pilot project inspired by the work carried out on the APIS project, and details an approach to automatically label sentences in Wikipedia articles for the final goal of training a neural sentence classifier. Both sections build the foundations of the approach outlined in Chapter 4.

3.1 Related Work

A recent example of a knowledge-based approach to IE is presented by de Araujo et al. (2013), who make use of domain-specific ontologies, as well as deeper levels of text processing in conjunction with different sets of annotations. The approach is split into a linguistic and a computational phase, with the former encompassing an extensive linguistic analysis of documents from the target domain. By analysing, linguistic patterns or rules are extracted that accurately depict certain

[1] Parts of these contributions have been discussed in Plum et al. (2019)

types of information, in this case events in legal documents. The result is a list of rules and a domain ontology, written in three different formal languages. In the computational phase, the three previously created ontologies are used on dependency parsed texts in order to extract a knowledge base of events that occurred in the corpus of 200 legal documents collected for evaluation. Overall, the authors report excellent results, with 96% to 100% precision depending on the type of event. However, this was achieved on a limited set of documents with considerable human annotation effort.

Other IE approaches have specifically targeted Wikipedia as a source of information. Russo et al. (2015) explore methods to extract biographical information from Wikipedia and DBpedia. Relevant information about a given person that is extracted includes the name, birthplace and birth date. The system presented by Chisholm et al. (2017) generates one sentence summaries from structured biographical information, using a dataset that is composed partly of structured data from Wikidata and partly of Wikipedia articles for text. Gotti and Langlais (2017) have made use of the multilingual nature of Wikipedia, training IE systems for different languages. Biographical information extraction has also been applied to texts other than Wikipedia articles, such as the rule-based approach described by Bonch-Osmolovskaya and Kolbasov (2015). The approach extracts biographical information from works by Tolstoy and commentaries on these. However, the authors describe challenges in the development of rule-based approaches. These include how to create strict classification rules (mainly text span boundaries denoting a relation), ambiguities arising from matching words to

correct meanings, and NLP problems such as complex noun phrase detection.

Turning to IE methods that are more machine learning-based, Yu et al. (2015) demonstrate an information extraction method for biographical data based on so-called triggers. Triggers are key words that denote style of relation and are defined before extraction. For instance, the word *sergeant* might trigger the *job* label, or the word *university* the *education* label. The authors present a rule-based and an SVM-based system to classify the scope of each trigger word. The scope is defined as the accompanying information of each trigger word (such as names dates or locations, as well as other word types such as prepositions). Overall the approach is reported to be highly successful. The authors present an evaluation performed using the KBP 2013[2] dataset, with an overall F1-score of 71%. This method is applied in Section 3.3, however, and its usefulness proved to be mixed.

Aprosio and Tonelli (2015) suggest a simple approach to classifying biographical sections in Wikipedia texts. Following an automated approach for collecting data, the authors select the Pantheon (Yu et al., 2016) dataset and assign labels to sections in articles based on keywords in their individual titles. In addition, a small subset is manually annotated, with a reported Inter-Annotator Agreement (IAA) of 0.88 (using Cohen's Kappa). The supervised learning task is carried out on a CRF classifier, which is compared to a non-sequential SVM-based approach as well as a baseline based on word frequencies. For the machine learning classifiers, six features are selected including section titles in

[2]The Knowledge Base Population datasets are released for conferences of the same name, and feature annotated sentences with entity relation pairs.

various configurations (such as lower-cased, upper-cased, first word only, last word only, etc.), bigrams and section position. The authors report results for two different evaluation settings, one strict (exact matches) and one that accounts for overlap. In terms of F1-score, the baseline, SVM and CRF approaches score as follows: 56% and 84%, 61% and 89%, 67% and 89%, respectively (strict and loose). The authors also claim that increasing the training set does not result in a significant increase in performance. The approach confirms that regularities within Wikipedia can be exploited to automatically obtain training data.

More recently, neural networks have been used in place of traditional ML models for information extraction tasks (Wu and He, 2019; Han and Wang, 2021; Jin et al., 2021). However, these approaches, although showing better performance than previous methods, generally rely on either robust patterns or large amounts of training data. The second half of this chapter, Section 3.3, presents an attempt to easily generate data for these kinds of models.

In terms of ranking biographical entities in terms of relevance, research is much more limited and specific to certain applications. Nevertheless, it is worth mentioning Hogue et al. (2014), who present an approach which uses page traffic to determine sentences of importance in Wikipedia articles. In this novel approach, page traffic is measured on Wikipedia by looking at when an event happens and then looking at sentences being edited in the articles corresponding to the events. Using a manually annotated corpus of 10 Wikipedia articles, the system is evaluated against a baseline that simply returns the first sentence of a Wikipedia article (the assumption being that the first sentence of a Wikipedia

article always contains the most important information). Overall, the system does not perform better than the baseline, even when combined with the first sentence baseline approach. The authors list a number of reasons for the results, including alignment problems of traffic to the articles, annotation difficulties with a high level of disagreement between annotators, and the size of the dataset.

The related work presented here shows that Wikipedia can indeed be useful for information extraction. While Wikipedia will be used throughout this thesis as a source of data, the following sections will test its use in two slightly different settings, before Chapters 4 and 5 apply the findings at a larger scale. Section 3.2 will look at how information can be extracted from Wikipedia itself, and whether its structured derivative Wikidata contains the same information. Section 3.3 presents a pilot study using Wikipedia as a source for training data.

3.2 Biography Extraction from Wikipedia

In the context of facilitating the digitisation of the ABD, this section presents a processing pipeline which can be used to identify relevant biographical candidates in Wikipedia, and to extract information about these candidates. The work presented was undertaken by the author of this thesis, although it became an integral part of work on the APIS project, and is aimed at enriching the ABD by means of automatically suggesting candidates for inclusion in the newer editions of the ABD. Biography candidates include people that died between 1951 and 2000 and who had an impact in the present Republic of Austria. The work focuses on processing English texts (but can potentially be adapted to other languages) and

selecting relevant candidates for inclusion in the ABD. However, since previous work on adding biographies to the ABD was carried out by a team of historians, with decisions sometimes being based on factors which are difficult to model, evaluation of the pipeline is not straight-forward.

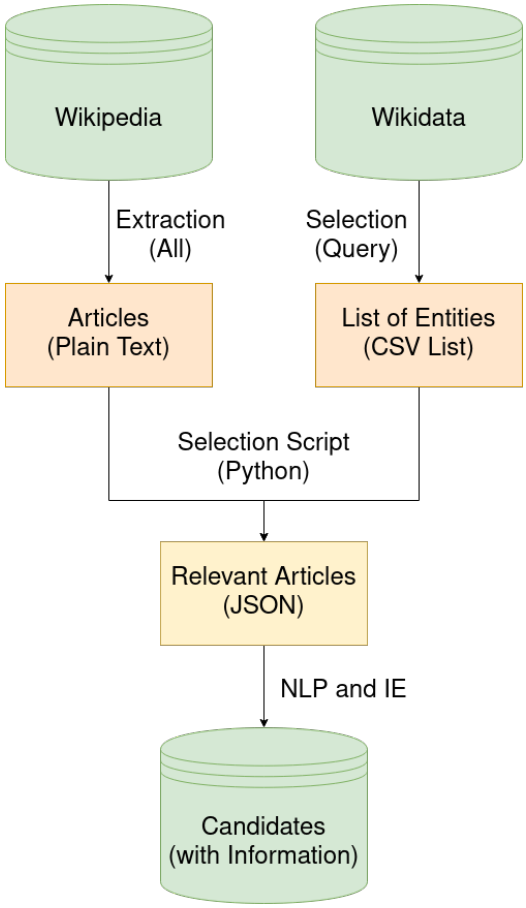

Figure 3.1: Sequence diagram of the rule-based candidate extraction pipeline.

The pipeline consists of three steps, and is depicted in Figure 3.1. An important aspect of the approach is data preprocessing, presented in Section 3.2.1. This includes entity selection from Wikidata, using parameters in accordance with

the project scope, as well as article extraction from Wikipedia. Section 3.2.2 presents a basic rule-based approach for extracting potential candidates and relevant information about them. Section 3.2.3 discusses an attempt at ranking the potential candidates, and Section 3.2.4 concludes the study with a discussion about the evaluation of various aspects of the pipeline.

3.2.1 Data Preprocessing

When using Wikipedia as a source of data for biographical IE, selecting relevant data is an essential task. Whether data is relevant depends very much on the task or project at hand. The present thesis deals with biographical information, which can be widely defined as any information given about a human being, but in particular facts like *date of birth*, *place of birth*, *occupation*, *parents*, *siblings* and similar relations. It has been shown that certain sections within Wikipedia articles are labelled according to the kind of information that they present. According to Aprosio and Tonelli (2015), however, this is only true in about 20% of pages about persons, where the remaining 80% of sections are not labelled correctly. For approaches that are aimed at large scale processing of data or that require large amounts of data for training neural networks it may be desirable to cast a wider net. Presently, it may also be the goal to process whole articles about a certain subject instead of just sections. Additionally, there may be finer-grained search criteria that have been defined, such as the case with the ABD project, one of the motivating influences behind this thesis.

In order to facilitate the data selection, the search parameters must first be

defined. Following along the example of the ABD project, it was established that *biographical* information about the following types of entities be detected:

(1) Entities are Human.

(2) Entities died between 1951 and 2000.

(3) Entities had an impact in Austrian history.

Entity Selection from Wikidata Extracting articles from Wikipedia that belong to a certain category or fulfil parameters such as the ones defined above is inhibited by a number of factors, as was observed during this research. Due to the collaborative nature of Wikipedia there can sometimes be a lack of standardisation. Finding articles that belong to a certain category (i.e. articles about humans, locations, companies and so on) is limited by the fact that these have to be explicitly grouped together, however, this is often only done for popular subgroups of these (i.e. *famous* humans). Concerning more specific search parameters, here the problem of precision presents itself. Wikipedia articles mainly consist of text, with structured information varying from article to article. The available query endpoint is mainly designed to search the whole text of Wikipedia articles. Since the parameters for this thesis are quite precise, it would require a large amount of post-processing in the form of filtering out texts that are irrelevant. Second, the query endpoint is online and therefore not suitable for large queries, since these tend to be limited in some form, resulting in the API blocking access from the computer making the query. Therefore, extracting articles according to a set of parameters requires additional tools.

Selecting data could be done by extracting articles from a Wikipedia dump or scraping directly from the web, however, this would require more effort in terms of preprocessing and filtering in order to obtain specific content. This method could also lead to extracting text which may not be very precise. One solution to this problem is to analyse Wikipedia infoboxes to obtain further information about the relative text (Chisholm et al., 2017). There are problems with the infobox approach though, as they are not uniform and do not always exist for every article. In addition, they are said to be difficult to process due to their non-standardised format (Attardi, 2015). Therefore, another solution that has proven to be useful is using a structured derivative of Wikipedia, such as DBpedia or Wikidata (formerly freebase). These databases offer access to most (if not all) of the information available in Wikipedia in a structured and easily searchable way. Since DBpedia only offers an older database (compiled from a 2016 snapshot of Wikipedia), the best alternative appears to be Wikidata, part of the Wikicommons group and offering the most up-to-date structured information from Wikipedia.

Wikidata offers labelled structured data about all entities in Wikipedia. Information is generally gathered from Wikipedia and formatted according to the RDF[3] standard. This makes Wikidata very suitable, because querying structured data is more reliable than searching for certain strings in text, where the context cannot be taken into account. Wikidata has also been used as a reliable data source for a similar approach with biographical data (Chisholm et al., 2017).

[3]The **R**esource **D**escription **F**ramework groups facts as statements, where two entities are linked by a relation.

Wikidata provides an online query endpoint, which allows for queries to be made. These are written in SPARQL[4] which is derived from SQL[5]. The endpoint is suitable for quick searches in order to gauge the amount of available information for a given set of parameters. A sample query for the parameters set for this project is shown in Figure 3.2. The query sets a number of named variables with different selection characteristics from lines 1 to 9. Lines 10 to 27 set filters, including attributes such as *human*, *date of death* and whether or not an entity has a *job*. Filters are also set for target languages of the entities, which are English, German, Czech and Hungarian. The final two lines define the ordering of the returned results. Figure 3.3 shows part of the response to the described query, with the columns showing each previously defined variable, such as *born* (*date of birth*), *loc born* (*birthplace*) and *GND ID* (*German National Dictionary ID*).

As with the online query endpoint for Wikipedia, however, the Wikidata query endpoint is not suitable for large scale tasks. This is mainly due to the limitations placed on the depth and number of queries allowed from the endpoint. If, for instance, the number of parameters increases for each search, the return time of the query also increases, timing out if the number of parameters is too high (a limit that varies from subject to subject and complexity of the parameter itself). In addition, querying the database for a list of entities that fulfil the parameters will eventually time out if the number of returnable entities becomes too large. During initial testing of the approach, it was found that for the target entities as described

[4]SPARQL is recurisvely defined as SPARQL Query Language.
[5]**S**tandard **Q**uery **L**anguage is used to query databases that follow the SQL standard.

```
SELECT ?h
(SAMPLE(?gnd) AS ?GND_ID)
(MIN(?name) AS ?name)
(MIN(?dob) AS ?born)
(MIN(?date) AS ?died)
(MIN(?pobLabel) AS ?loc_born)
(MIN(?podLabel) AS ?loc_death)
(MIN(?desc) AS ?description)
(GROUP_CONCAT(DISTINCT ?jobLabel; SEPARATOR=" / ") AS ?jobLabels)
WHERE {
  ?h wdt:P31 wd:Q5. # human
  ?h wdt:P569 ?dob. # born
  ?h wdt:P570 ?date. # died w/ filter
  ?h wdt:P19 ?pob. # born in
  ?h wdt:P20 ?pod. # died in
  OPTIONAL { ?h wdt:P106 ?job. } # has job
  OPTIONAL { ?h wdt:P227 ?gnd. } # show/filter GND ID
  FILTER(?date > "1950-12-31T00:00:00Z"^^xsd:dateTime)
  FILTER(?date < "1952-01-01T00:00:00Z"^^xsd:dateTime)
  #FILTER(?date < "2019-01-01T00:00:00Z"^^xsd:dateTime)
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en,de,cs,hu" .
                           ?h rdfs:label ?name.
                           ?h schema:description ?desc.
                           ?job rdfs:label ?jobLabel.
                           ?pob rdfs:label ?pobLabel.
                           ?pod rdfs:label ?podLabel. }
}
GROUP BY ?h
ORDER BY ?died
```

Figure 3.2: Wikidata query written in SPARQL to search for specific entities from Wikidata.

wikidata_link	GND_ID	name	born	died	loc_born
http://www.wikidata.org/entity/Q29917973		Wilhelm, Károly	1886-01-01	1951-01-01	Österreich-Ungarn
http://www.wikidata.org/entity/Q8019667	117258741	William Valentine Schevill	1864-03-02	1951-01-01	Cincinnati
http://www.wikidata.org/entity/Q51540730		Charles Frederic Ramsey	1875-01-01	1951-01-01	Pont-Aven
http://www.wikidata.org/entity/Q1270329		Heinrich Reese	1879-02-19	1951-01-01	Basel
http://www.wikidata.org/entity/Q17471830		Fazıl Doğan	1892-01-01	1951-01-01	Mytilini
http://www.wikidata.org/entity/Q1358977	141494964	Ernst Klein	1876-04-15	1951-01-01	Wien
http://www.wikidata.org/entity/Q5276568	122114973	Dikran Kelekian	1868-01-01	1951-01-01	Kayseri
http://www.wikidata.org/entity/Q21010063		Leon Karp	1903-01-01	1951-01-01	New York City
http://www.wikidata.org/entity/Q2383242		Sievert Allen Rohwer	1887-01-01	1951-01-01	Telluride
http://www.wikidata.org/entity/Q35226291	124032192	Arthur David Gayer	1903-03-19	1951-01-01	Pune

Figure 3.3: Example of results returned from the Wikidata query.

above, the limit to the number of results before the query times out is roughly 10,000. Launching many queries that time out additionally leads to queries being blocked from the IP address of origin.

The solution to the query problem presents itself in the form of a so-called *dump file* of the whole of Wikidata (or simply a *Wikidata dump*). These files are

available to download freely, and contain the entirety of the Wikidata database at a given point in time. As is to be expected, these dumps are extremely large in size, with a compressed file size of about 1TB (decompressing was not possible on the hardware available). Querying this offline version of Wikidata can be solved in two ways. The first is to create a server running a local version of this file, utilising database software and implementing a SPARQL query endpoint. This approach should theoretically not time-out, assuming both a capable server and local network. However, this requires some knowledge of networks and servers, and requires suitable hardware.

The other approach is to use a programming script to decompress the dump file in small parts at a time, and to then search the text in each part line by line. This can be accomplished by using Python and some extended libraries. The approach for this project was to unzip the dump file one line at a time, using the bz2file library. Due to the fact that the dump file is in JavaScript Object Notation (JSON) format, each line can then be loaded and queried according to the available properties in a normal SPARQL query of Wikidata. Matching entities can then be written to a separate file. The speed of this approach is governed by the machine specifications of the computer, but was found to be 2 to 3 hours on the machine available. At the time of processing this returned around 401,695 entities (equivalent to lexicon pages), and of these 172,131 had corresponding articles in English. The total number returned indicates that multilingual methods could extract from articles in languages other than English.

It is also worth mentioning at this point that some discrepancies exist between

Wikidata and Wikipedia information, leading to some retrieved entities not being used (see Section 3.2.4 for further information). This approach is only limited by the speed of the computer and is therefore usually more suitable for large-scale queries such as the one described for this project.

Once a query has been carried out, the output is a file of matching entities. Figure 3.4 is an example of one entity as it appears in the output file. The first block from line 1 to 7 provides the internal ID, the Wikidata ID, the description, the name in English and the name in German of the entity. The following blocks, denoted by the letter P and a number, each provide information about a certain attribute of the entity and the assignment of these identifiers is done via Wikidata. For example, the first block denoted "P21" is the attribute *Gender*, which is in this case *male* (in German *männlich*). In some cases these values are also assigned a unique Wikidata ID. All the subsequent blocks follow this pattern.

The resulting output file can then be used as a starting point for the following steps. It serves as a database of entities matching certain criteria within Wikipedia, offers more information about each entity (if available) and also provides the Wikidata entity identification number (referred to henceforth as *QID*), which will be needed at a later stage.

Wikipedia Article Extraction Extracting text from Wikipedia articles and annotating it requires a considerable amount of preprocessing. While Wikipedia is freely accessible online, extracting a large collection of articles from Wikipedia is not a straight-forward task, as was pointed out in Section 3.2.1. A number of

```
{
    "_id" : ObjectId("5ed92a39bb725d32e22ceff3"),
    "qid" : "Q2496",
    "description" : "German politician",
    "label" : "Ludwig Erhard",
    "label_de" : "Ludwig Erhard",
    "P21" : {
        "cid" : "Q6581097",
        "en" : "male",
        "de" : "männlich"
    },
    "P27" : {
        "cid" : "Q183",
        "en" : "Germany",
        "de" : "Deutschland"
    },
    "P735" : {
        "cid" : "Q14159020",
        "en" : "Ludwig",
        "de" : "Ludwig"
    },
    "P734" : {
        "cid" : "Q36912046",
        "en" : "Erhard",
        "de" : "Erhard"
    },
```


Figure 3.4: Wikidata Entity extract in JSON format.

steps are involved in order to prepare the text so that it can be further utilised. The first step is to obtain the text, which is done by means of a Wikidump. Next, articles have to be extracted from the dump file to a plain text format, since the XML representation includes a large amount of markup that is not relevant for the type of work carried out in this research. An overview of the extraction process is shown in Figure 3.5.

Wikidumps are created on a monthly basis and are freely available to download online[6]. Wikidumps are an image of all Wikipedia articles available in a selected language at a given time. These images are in XML format, and therefore not only include the text but also the structure of an online Wikipedia article.

[6] https://dumps.wikimedia.org/enwiki/

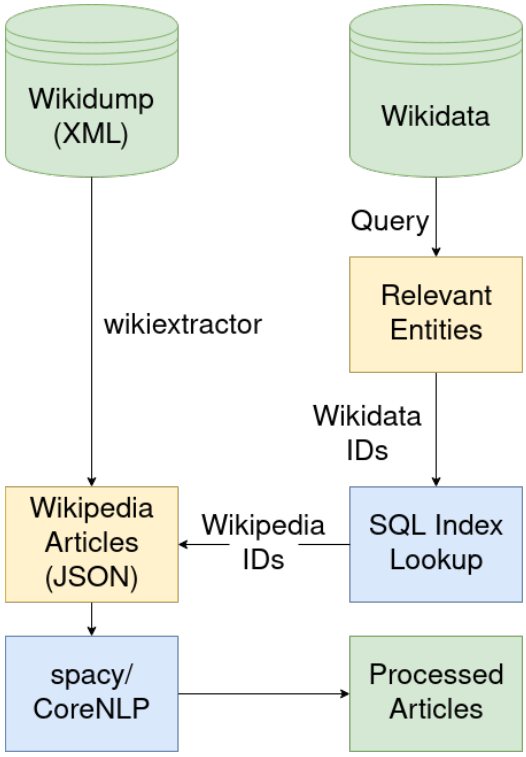

Figure 3.5: Overview of the Wikipedia article extraction process.

Article structures can be quite complex, including sectioning and any number of infoboxes, while not following a uniform guide (Attardi, 2015). Wikidumps are archived and have to be unzipped, requiring a fair amount of disk space since they can expand a great deal (around 70GB for English at the time of writing). The result is a single large XML file, which is difficult to process at once.

In the next step, the article text has to be retrieved from the XML file. This involves navigating the complex XML tree structure of the file, as well as dealing with the limitations that are imposed when working on such a large file. The solution is to use libraries that process the large file in parts rather than loading the complete file into memory. Since extracting text from the complex tree structure

is quite difficult by itself, it is recommended to use a thoroughly researched implementation. The python program *wikiextractor* (Attardi, 2015) fulfils these specifications, taking as input an XML dump and outputting many smaller files containing all extracted articles. To ease processing at a later stage, the output should be set to JSON format, in order to easily identify text and article IDs. The resulting file is formatted as follows: Each curly bracket pair contains one article, including the Wikipedia ID, the original URL to the article, the title of the article and the content of the article itself. Figure 3.6 is an extract of the JSON file grouping all relevant articles together, and shows four different articles (although the text has been shortened for presentation reasons).


```
{"id": "28663075", "url": "https://en.wikipedia.org/wiki?curid=28663075", "title": "Kang Ji-young",
"text": "Kang Ji-young\n\nKang Ji-young (born January 18, 1994), also known as Jiyoung or JY, is a South
Korean [...]"}
{"id": "28330763", "url": "https://en.wikipedia.org/wiki?curid=28330763", "title": "Bailey Wright",
"text": "Bailey Wright\n\nBailey Colin Wright (born 28 July 1992) is an Australian professional
footballer who [...]"}
{"id": "28332094", "url": "https://en.wikipedia.org/wiki?curid=28332094", "title": "Mason Ryan", "text":
"Mason Ryan\n\nBarri Griffiths (born 13 January 1982) is a Welsh professional wrestler and actor [...]"}
{"id": "28332613", "url": "https://en.wikipedia.org/wiki?curid=28332613", "title": "Gy\u00f6rgy
Bessenyei", "text": "Gy\u00f6rgy Bessenyei\n\nGy\u00f6rgy Bessenyei (1747\u20131811) was a Hungarian
playwright and poet.\n\n"}
```


Figure 3.6: Extract from JSON file of Wikipedia articles.

For some parts of the methodology, it is crucial to link the structured information in Wikidata to the Wikipedia articles. This needs to be done to unambiguously identify and extract relevant articles in order to avoid processing large amounts of text very often. It can also be to identify articles containing certain information that is useful for automatic annotation. However, the major difficulty here is the fact that Wikipedia articles are identified by a unique identification number provided by Wikipedia (referred to henceforth as *CURID*). As the selection of

relevant articles is carried out using Wikidata, this requires mapping the QIDs to the CURIDs. Unfortunately, although Wikidumps should contain a *sitelinks* entry for each entity that links to the corresponding Wikipedia article (Chisholm et al., 2017), this was found to not always be the case. Wikidata should link corresponding Wikipedia articles to each entity online, this field was not present in various[7] Wikidata dumps accessed for the purposes of this thesis. There are tools available to map Wikipedia articles to Wikidata entities, such as Wikimapper[8], although this tool uses article names instead of CURIDs, which can be more difficult to match.

The problem can be solved by retrieving an SQL index file, which is usually available for each Wikidump. These files contain an SQL database of articles in a given Wikidump and provide the QID link to Wikidata. Mapping the IDs can be carried out by using SQL database software, or by compiling a list of CURIDs and looking these up in the SQL file using a regular expression that concatenates the CURIDs together. Once an indicative list of the IDs is complete, the articles can be filtered from the collection of all Wikipedia articles by matching the CURIDs. The result is a list of relevant articles only.

The final step is to apply general NLP methods to each article in a uniform way. Processing such a large number of texts is in itself a time-consuming task. Finding fast and efficient ways of doing this can be quite time-consuming. Initially, the approach described here used the StanfordCoreNLP (Manning et al.,

[7]Dumps from various dates in English and German were tested
[8]https://github.com/jcklie/wikimapper

2014) framework written in Java. The reasoning behind this choice was that this programming language is generally considered to be fast due to it being a lower level language. However, it was found to be quite slow when applying a general NLP pipeline, including the somewhat power-intensive dependency parsing, and would take roughly 24 hours for 10,000 texts. A more detailed description of the challenges of processing large amounts of text using Stanford is presented in Plum (2018). Since then, the spaCy[9] framework has been used instead, providing much faster and more reliable processing, since it is trained partly on Wikipedia articles, therefore offering robust results. The processing time was cut to two hours for 200,000 texts.

Applying this approach to the entities selected with Wikidata in the previous step, the result was 170,517 articles, down 1,614 from the previous number. The reduction in the number of articles is due to mismatches with the IDs (CURID versus QID), as well as extraction errors from XML, where articles could not be extracted properly. In addition, some articles may not be processed due to conversion artefacts, such as XML tags making the text unreadable for the selected NLP tool. If the approach used here was intended to be exhaustive, these problems could be addressed. However, since the number of articles lost here is minimal in this case (around 1%), these problems were not addressed as part of this research.

[9] https://spacy.io

3.2.2 Rule-Based Candidate Extraction

The information extraction step follows a shallow rule-based approach in order to take advantage of the basic information from Wikidata. Exploiting the structure of Wikipedia articles, it is possible to extract various information, including the name and date of birth, place of birth, etc. By extracting the tokens of the heading of each article, the name can be extracted from the article. Any information that is contained in brackets is removed, such as *(Jobtitle)*. These additions are sometimes present to disambiguate certain persons with the same name. Furthermore, a simple rule based on observations and preliminary work carried out by Plum (2018) is used: the second full date that is mentioned in the first sentence of each article is usually the date of death. As each article has been tagged in terms of named entities, including expressions of time, this part is simply extracted from the annotations by iterating over the time expressions. Each date is converted to *YYYY-MM-DD* format, and all full dates formatted in this way are accepted. Partial matches where at least the year is present are accepted in case there is not other match. Should no dates be detected, the information is left blank. The extraction is carried out using a Python script, which compares the extracted information with that contained in the meta information.

Figure 3.7 shows an extract of the beginning of a Wikipedia article. Of note should be the first sentences, which is of a typical structure for biographical articles in Wikipedia: the name followed by the dates of birth and death in brackets. Furthermore, that sentence contains a short identifying part about the

person, such as the job or main occupation. In the case of this example, the first sentence is followed by a brief timeline of the most important events of that persons life. Since the person was a famous chess player, it summarises notable chess tournament results.

Hans Müller (1 December 1896, Vienna – 28 February 1971, Vienna) was an Austrian chess player, theoretician and author of books.

In 1921, he played in Vienna; tied for 9-10th (Friedrich Sämisch won), tied for 1st-2nd with Gruber, and took 6th (Vladimir Vuković won). In 1922, he tied for 4-5th in Innsbruck (Ernst Grünfeld and Rudolf Spielmann won). In 1923, he tied for 4-6th in Budapest (Endre Steiner won). In 1924, he tied for 8-9th in Győr. In 1925, he tied for 1st-2nd in Debrecen. In 1925, he tied for 5-6th in Vienna. In 1926, he took 7th in Bardejov (Bardiov). The event was won by (Hermanis Matisons and Savielly Tartakower. In 1926, he tied for 7-9th in Trenčianske Teplice (Trentschin-Teplitz, Trencsénteplic). The event was won by Karl Gilg and Borislav Kostić. In 1926, he tied for 5-6th in Hyères (Abraham Baratz won). In 1926, he tied for 8-9th in Vienna (10th Trebitsch-Turnier). The event was won by Spielmann. In 1927, he took 9th in Kecskemét (Alexander Alekhine won). In 1927, he tied for 1st with Albert Becker in Vienna. In 1928, he took 8th in Vienna (Richard Réti won). In 1929/30, he took 3rd in Vienna (13th Trebitsch). The event was won by Hans Kmoc for 7-8th in Vienna (14th Trebitsch; Becker won). In 1932, he tied for 5-7t Becker won).

Figure 3.7: Extract from the Wikipedia article about Hans Müller.

3.2.3 Candidate Ranking

As the articles have been annotated in terms of named entities, locations can be extracted by simply searching for any *LOC* or *GPE* tags. This approach is used to compile a list of locations mentioned in each article. Next, the country of

each location is determined. The first approach was to use the GeoCoder api[10] connected to GeoNames, to retrieve the country of each location. Unfortunately, the GeoNames api is restricted to 1,000 requests per hour for locations. With such a large dataset at hand, this would not be a viable approach. Other than buying requests as part of a premium service, the other option is to download the CSV file of GeoNames locations. These lists are available in various sizes, containing locations based on the population size. For this approach, the largest dataset was used, containing every location available in GeoNames. The file lists every location, i.e. villages, towns, cities, landmarks, as well as the country that item is in, and other information such as population, alternate names, etc. Disambiguating ambiguous location names is a complex task in itself, and was not carried out here, with the assumption being that the largest match would be correct. The MongoDB[11] database system can be used to create a local, indexed database, which allows for efficient searches. Since so many locations are queried during the extraction process, it is necessary to ensure that a location query can be carried out quickly.

For each location, a query is made about whether a location among those found in each article belongs to the Republic of Austria. If this is the case, the article is included, meaning the main entity (person) of that article is a candidate. In the first iteration, it was found that invalid locations were being matched in

[10] https://geocoder.readthedocs.io/api.html

[11] MongoDB is available via https://mongodb.com and is a highly customisable implementation of a database system. It is less strict than more traditional databases such as SQL, and is therefore easily extendable. For research purposes, this offers flexibility, which is desirable.

the documents, such as *City* or *Lake*. These location names are always part of a longer name, but can be matched as they are tagged individually. Because the unrestricted GeoNames database was used, these locations were found, although it is unclear why these kinds of locations are in the database in the first place. One explanation could be the open-source nature of GeoNames, meaning that there is a margin of locations that need to be filtered out. Interestingly, these locations always showed a population of zero, again hinting at the fact that they were mistakenly or incorrectly added. This is a side-effect of using the largest GeoNames database without population restrictions. Rather than using a restricted dataset with a small number (i.e. a population of more than 500), a parameter was added to discard locations with a population of zero.

In order to test out how a method of ranking relevant candidates could work, all locations within Austria per article were counted, and compared to the number of locations not within Austria. The main idea behind this was that these counts could put the ranking into better perspective, if for example the number of locations that are not within Austria are far higher than the number of locations within Austria, this could hint at a non-relevant candidate. An evaluation of the counts is presented in Section 3.2.4.

3.2.4 Evaluation

This section presents the evaluation results of the information extraction task, and investigates the selection of possible candidates. The evaluation of the results is challenging, due to the lack of any annotated data or indication by historians

of what is relevant. For the IE task, the extracted names and dates of birth are compared to those returned by Wikidata, but this assumes Wikidata as a gold standard. For the ranking of the candidates in terms of relevance, there is no gold standard, for reasons that will be explained. Therefore, it should be clear that this is not an evaluation of the extraction itself, but rather the process of selecting biography candidates.

Differences in Wikipedia and Wikidata As described in the previous section, the name and date of death of entities were extracted from the Wikipedia articles. Each result was compared with the information obtained from Wikidata. Of the 170,517 articles, the name did not match in 18,267 cases. Upon closer inspection, it was found that this is largely due to differences in spelling, and name variations. Table 3.1 shows a selection of the most common errors: The first two rows are examples of differences in shorter names. Rows 3 and 4 show different levels of preciseness in naming, e.g. abbreviations. The last row shows an example where Wikidata returned the name in German, whereas the system extracted the name in English, indicating that the English name may not be present in Wikidata.

Wikidata	**Extracted**
Robert Joshua	*Bob* Joshua
Francisco Javier Vidarte	*Paco* Vidarte
Joe C. Davis**,** Jr.	Joe C. Davis Jr.
Vincent Graber	Vincent **J.** Graber **Sr.**
Karl Aloys von und zu Liechtenstein	Prince Karl Aloys of Liechtenstein

Table 3.1: Various examples of naming differences in Wikipedia vs. Wikidata.

In terms of the date of death, there were 30,153 cases where the date of death did not match the Wikidata records. Using a Python script the different errors that occurred were counted and assigned to one of three main types of error: *no date* errors, *minor difference* errors and *birthday* errors. In the first case, the extraction rule returned 0000-00-00, indicating no date was extracted. This error was caused by the NER algorithm not detecting a date, or it being missing in the article itself. The second most common error was the *minor difference error*. In this case, the difference between the Wikidata date and the extracted date was minor, i.e. only between one to five days difference. This suggests that Wikidata also gathers information from other sources, or that it could be caused by time zone differences. The last error to occur was the *birthday* error. Here, the date that was extracted did not match the date of death extracted from Wikidata, but rather the corresponding date of birth. This is caused by the fact that a date is extracted from all sentences even if only one date is found by the NER algorithm.

The frequency distribution of the different types of error is shown in Table 3.2. The percentage is taken from the total number of articles. Therefore, one can expect a date error in 17.68% of all articles. More specifically, this would be a *no date* error in 11.67% of articles, a *minor difference* error in 3.9% of articles and *birthday* errors in 2.11% of articles.

Taking these results into account, it is interesting to see where differences in information lie. Using Wikidata and Wikipedia as complementary components has some benefits. Firstly, Wikidata excels as a tool to select data according to specific criteria. As processing Wikipedia articles is extremely time-consuming,

Error	Count	% of Total Articles
No date	19,895	11.67%
Minor difference	6,654	3.90%
Birthday	3,604	2.11%
Total date errors	30,153	17.68%

Table 3.2: Frequency distribution of date error types.

this reduces the time dramatically because irrelevant articles can be omitted. Secondly, Wikidata can serve to some extent as a kind of gold standard against which to compare the results of any extraction carried out on Wikipedia articles. Of course, this is only to a limited extent, as not all relations are available in Wikidata, as is the case with the extraction of locations to determine candidates.

Going beyond the use as a gold standard, the two data sources can also be used to extract information in a more complementary way, i.e. using Wikidata for basic information, and building on these known relations to extract further information from Wikipedia. However, whether used as a gold-standard or not, this analysis has shown that there are differences in the two resources. Even though both Wikipedia and Wikidata are thought of as gold-standard, the differences in names and dates show that this may not always be the case. In relation to the relatively small amount of target articles extracted here, the number of name and date errors is quite high. Should an approach such as the one suggested here, or similar, be scaled up to larger numbers of articles, it is feasible that these errors would also increase. This would cause disruption to the process, putting into question which source to trust more over the other, which can be difficult to answer at that scale. Nevertheless, a deeper investigation into these matters is beyond the scope of this

thesis.

Location	Count
Hall	4,191
Vienna	3,529
Point	997
Salzburg	359
Bergen	339
Nassau	197
Innsbruck	192
Graz	180
Königsberg	173
Inn	120

Table 3.3: Top 10 locations that led to candidate selection.

Extracting Biographical Dictionary Candidates Using the location matching script (as described in Section 3.2.3), 13,521 possible candidates were determined. An investigation of candidates picked at random shows that the method is most probably not precise enough. For each candidate, the location that caused its inclusion in the candidates list was analysed. The ten most common locations are listed in Table 3.3. While this list includes many valid locations, it is clear that many articles are chosen as candidates due to matches because they contain the words *Hall* and *Point*. Further examples include articles selected because they contain the words *Sand*, *Fall*, *Gray* and further similar words. These match proper locations in Austria, however, they also match English nouns and adjectives, and are most probably part of longer location names. Another common problem in this regard was the matching of names which are ambiguous, as they also match

locations in Austria, and therefore contributed to being considered as candidates.

In terms of the ranking of the candidates using the location counts, this is just as hard to evaluate as to measure. A manual analysis indicates that the ranking mechanism at this point is not precise enough. Quite often candidates rank very highly, even though there is no relevance to Austria whatsoever. This is mainly due to the false matches, described previously. Some irrelevant candidates rank very highly, whereas some very relevant candidates rank very low. Candidates that should probably be considered with priority are ranked very low, due to only a few or one locations being matched. However, this is mostly due to extremely short Wikipedia articles, which do not hold much information. This could be solved by normalising the values, but would require some feedback from historians.

Further evaluation of the locations extracted against a gold standard is not possible. Wikidata queries rely on a relation, such as *born in*, to be present in order to extract the corresponding location. In this case, the aim was to go beyond these relations and find any kind of mention of locations that are relevant. Ultimately, the candidates derived have to be evaluated in an iterative process by historians in order to say how well the method performs. Even if the precision seems low, an approach such as this may have a much higher recall than a manual approach could, and would therefore still be worth using. Other forms of automatic evaluation do not exist at this point in time, especially considering there is no gold standard for this work, as it is mainly aimed at finding completely new candidates.

3.3 Neural Sentence Classification with Wikipedia

During the course of the research described in the previous section, the idea for exploiting certain structures within Wikipedia to automatically compile a labelled dataset was formulated. The proposed methodology involves using Wikidata to select a target group, for example persons, as in Section 3.2.1, and to then extract the corresponding articles from Wikipedia as described. Following this, a rule-based approach can be used to match certain entities in the sentence with the structured information provided by Wikidata. This would allow labels to be assigned automatically, therefore allowing an automated approach to assigning labels to large amounts of textual data. A corpus annotated in this way is appropriate for training neural classification models. The following sections describe a pilot study carried out to test the viability of such an automatic labelling approach. In order to keep the labelling and classification task more simple, the focus was at the sentence-level.

Section 3.3.1 describes how the dataset was compiled. Following this, Section 3.3.2 discusses the reasoning for the inclusion of coreference resolution and Section 3.3.3 describes the neural models selected to be trained using this data. The evaluation is discussed in Section 3.3.4.

3.3.1 Data Compilation

The dataset is compiled from a selection of 20,000 Wikipedia articles, split into two groups. One half is made up of articles that describe people (human), the other half is made up of articles that are not about people (other). The idea behind this

split is to obtain a balanced set of articles that contain biographical information and articles that do not. In the past, this has been achieved by using two different sources of data, such as with Biadsy et al. (2008), who use Wikipedia on the one hand and a news corpus on the other. However, the assumption that sentences from another corpus are always non-biographical may not be correct with news articles. By instead checking the topic of each article, the approach presented here was aimed at improving upon this problem, since it seemed like a more robust approach for grouping sentences. Furthermore, it could be argued that there would be a certain number of sentences that do not contain biographical information just from articles that describe people. Sentences from articles on different topics should ensure that the dataset is varied and in turn allows a model to have seen many different kinds of sentences.

Example 3.1 shows sample sentences from the dataset. These sentences show that a wide variety of topics are covered, therefore offering a varied selection of sentences.

(1) Walter “Gulle” Oesau (28 June 1913 11 May 1944) was a German fighter pilot during World War II.

(2) Long Apung Airport is an airport serving the city of Long Apung, located in the Malinau Regency, North Kalimantan, Indonesia.

(3) At the 2005 World Amateur Boxing Championships he once more won bronze.

Example 3.1: Sample sentences from the Wikipedia dataset.

The first sentence is an instance of biographical information, describing what a

person was known for and the dates of birth and death. The second example does not contain any biographical information, describing an airport. The final sentence does contain biographical information, however, it does not directly name the person and refers to a pronoun instead. This is a problem that was intended to be resolved by means of coreference resolution for the purposes of this project, but may conceivably be solved differently in other projects.

Selecting articles based upon their contents can be achieved by looking up each article ID in Wikidata, where the Wikidata property *P31* corresponds to *human* or not. Articles are then processed and split into sentences using the spaCy NLP toolkit. Before being processed with spaCy, each article is cleaned of line breaks. In addition the title of each article is removed, as it leads to parsing errors if left in the main body of text. Next, each sentence is assigned labels according to one of three labelling subtasks, by connecting and matching the language processing information.

The first label is for a *binary task*. First, named entities in each sentence are compared to the main article entity. If a named entity matches the main article entity, then the sentence is used. Next, partial parsing information is used to determine whether a verb (or *action*) is connected with one of the main subjects. In the case of a directly connected verb, if the named entity matches the main article entity and the article entity is a human, then the sentence is labelled as *biographical*. If the main article entity is not human, or no corresponding verb can be detected the sentence is labelled as *non-bio* (non-biographical).

The second label is part of a *multi-class task*. This can be one of four labels,

and depends on the amount of information contained in the sentence: The binary labels from the *binary task* are kept, and extended by two further labels. The label *bio-location*, which is assigned when a location (NER tag *GPE*) is detected in the sentence as well. The label *location-date* is assigned when all the previous conditions are true and an additional date is found (NER tag *DATE*).

The final label or set of labels is assigned according to a *multi-label* task. This is done by using a list of *triggers*, an approach first proposed by Yu et al. (2015). Here, each token of a sentence is matched against lists of words which trigger their respective label. If a token such as *born* is present in a sentence, then it could for instance be assigned the *date of birth* tag, since it most probably denotes this information. However, during testing, it quickly became clear that neural models are excellent at learning trigger words. This meant that any time a trigger word was in a sentence, it would trigger that relation without any regard for context. This in turn would lead to an inflexible model. This observation was one of the first factors to give rise to the refined process presented in the following chapter. In addition, the first round of annotating showed that assigning labels was more complex than the other tasks. Therefore, this labelling task was not evaluated during the manual annotation process.

The approach outlined here allows for fast, automatic annotation of Wikipedia articles for different classification tasks. The example sentences shown in Table 3.4, taken from human and non-human articles, demonstrate how the different labels would be assigned.

Sentence	Binary	Multi	Labels
Olga Alekseyevna Zaitseva (born 16 May 1978) is a former Russian biathlete.	biographical	biographical	altName,born
Swarowsky was born in Budapest, Hungary.	biographical	bio-location	altName,born
Browningia is a genus of cacti, comprising 11 accepted and 3 unresolved species.	non-bio	non-bio	-
Wintu has 28 (to 30) consonants:	non-bio	non-bio	-
He died in 1923, and was posthumously promoted to Marshal of France.	biographical	location-date	died,job

Table 3.4: Labelled examples from the biographical sentence classification dataset. The last three columns are each a sub-task of the labelling process.

3.3.2 Coreference Resolution Considerations

One important question for this approach is whether coreference resolution improves the results in any way, as highlighted by the last sentence from Table 3.4. Even though the coreference chain of each entity that is considered is matched at the time of processing, this does not rule out matching errors entirely. This is especially the case with pronouns, which are unavoidable when dealing with biographical information. The treatment of pronouns is not a simple task and can have different implications across languages. Since the approach presented here is aimed to work in multilingual settings later (see Chapter 5), it is vital that pronouns are treated properly. In English, a the assumption that the pronouns *she*, *her*, *hers*, *he*, *him*, *his*, *they*, *their* and *theirs* refer to persons could be made. However, in German for instance, every noun is assigned one of three

genders (*masculine, feminine, neuter*) regardless of its actual gender (as it might be understood in English). This means that objects can be masculine (*the table* is referred to as *der Tisch*) and persons can be neuter (*the girl* is referred to as *das Mädchen*). The result is that pronouns can not be simply matched and would possibly require coreference resolution techniques to be treated properly.

However, the use of a coreference component introduces some noise in the data. This was particularly evident in testing for the research presented in this Section, and is indeed also seen in experiments conducted in Section 4.2.2. It is also clear from relevant literature that coreference resolution development has been focused on English, far outnumbering other languages (Mitkov, 1999; Steinberger et al., 2011; Sukthanker et al., 2020). Therefore, for the approach at hand, the assumption was made that this problem had to be solved by treating only the English training data for coreferences and testing the impact of coreference resolution in different variations.

While the study presented here has no multilingual aspect, this study was aimed at preparing the research presented in Chapter 5, where multilingual processing is a factor. Therefore, to assess the impact of coreference resolution for this approach, two sets of sentences were produced where anaphora were treated in different ways. For the first set, no coreference resolution was carried out, the sentences were left in their original state and labelled as described in Section 3.3.1. For the second set, the sentences were processed using spaCy's coreference resolution functionality. The anaphora where the algorithm deemed a replacement fit were replaced by the most likely proper named entity, i.e. avoiding cases where

a pronoun would link to another pronoun, therefore avoiding coreferential chains.

3.3.3 Neural Sentence Classifier

To test the suitability of the compiled dataset, a neural sentence classifier was trained for each of the classification tasks. At a *binary* level the system decides whether a sentence contains biographical information or not. At a *multi-class* level the system makes predictions concerning the granularity of information contained in the sentences.

A number of considerations contributed to the neural architecture setup that was chosen for the sentence classifier described here. In the past, the classification of sentences and shorter bodies of text has been carried out using probabilistic and statistical methods, including Bayesian classifiers, SVMs, CRFs and logistic regression (Glazkova, 2020). More recently, text classification has moved to the use of RNNs, as well as the use of word and context embeddings such as ELMo and BERT to represent words (Glazkova, 2020). Moreover, significant performance gains have also been achieved by using transformer models for a multitude of NLP tasks in English (Devlin et al., 2019). The proven success of the latest models is a strong argument for the use of a neural network, specifically one using transformer models, in the context of the project at hand.

The final setup of the neural classifier was based on the transformers architecture, which has shown excellent results in classification tasks (see Section 4.2.4 and Section 5.2.3). BERT and XLM-R were used, although the results in Section 3.3.4 are only reported for XLM-R since it performed better overall. Since

this study was intended as a pilot, the simpletransformers[12] implementation was chosen with default training settings. These include a learning rate of $4e$-5 with the AdamW optimizer. Training was carried out on the training set of around 400,000 sentences, balanced to equal numbers for the binary task. The same set was used for the multi classification task, with unbalanced sub-classes.

3.3.4 Evaluation

In order to assess the results of the classifications, a set of sentences was manually annotated for the purposes of this thesis. A set of 5,000 sentences was selected randomly from the total set of articles extracted (see Section 3.3.1) and not processed further (for instance coreferences using spaCy). After removing some duplicates the actual number was 4,984. The annotation was carried out by a native speaker, with the only guideline of assessing each sentence individually without letting any prior knowledge change the judgement. This means that each sentence has to explicitly state what the assigned label indicates.

For the binary classification task, Table 3.5 shows the evaluation results, without coreference resolution and with coreference resolution, respectively. Interestingly, the model trained on sentences with coreference resolution applied achieves higher precision when predicting *biographical* sentences, however, this is at the cost of recall. While it is not unreasonable that adding coreference resolution might improve precision, the aim of adding coreference resolution was to improve the recall, which it does not though. The *non-bio* sentences suffer a

[12] https://simpletransformers.ai

small reduction in precision only. In terms of overall accuracy, the models achieve roughly the same results, with 84% and 83%, respectively.

	NORMAL			COREF			
Label	**P**	**R**	**F1**	**P**	**R**	**F1**	**Support**
non-bio	.88	.88	.88	.83	.95	.88	3468
biographical	.73	.74	.73	.83	.54	.66	1516
macro avg.	.81	.81	.81	.83	.75	.77	4984

Table 3.5: Evaluation results of the binary task.

For the multi-class classification task, Table 3.6 shows the evaluation results, without coreference resolution and with coreference resolution, respectively. The results largely correspond to those of the binary classification tasks, with the the *biographical* and *bio-location* sentences increasing in precision but lowering recall. It is noteworthy that the final class *location-date* does not seem to improve, but does achieve reduced recall. In terms of overall accuracy, the models achieve roughly the same results again, with 80% and 81% respectively.

	NORMAL			COREF			
Label	**P**	**R**	**F1**	**P**	**R**	**F1**	**Support**
non-bio	.88	.88	.88	.83	.95	.88	3468
biographical	.63	.62	.63	.73	.46	.56	1136
bio-location	.37	.50	.43	.44	.37	.40	131
location-date	.56	.53	.54	.55	.37	.45	249
macro avg.	.61	.63	.62	.64	.54	.57	4984

Table 3.6: Evaluation results of the multi-class task.

Overall, the evaluation of this pilot study shows that a classifier is indeed capable of learning to classify sentences in terms of biographical content. The

application of coreference resolution to the training set does not offer much of an obvious advantage, although perhaps treating the input to the classifier after training as well could increase the performance. The most important observation of this pilot study, however, was made during manual annotation and pertains more to the general approach of annotating sentences in this way. It appears that the sentence level could be too broad for the more detailed classification task at hand. Simply labelling a sentence as "containing biographical content" may not always be useful, especially if all context is lost.

Moreover, a small test was conducted on a set of sentences taken from eBooks from the Gutenberg project[13]. A sample of these sentences is shown in Example 3.2. These incorrectly labelled samples include headlines, fragments, and other non-biographical sentences that have all been predicted to be biographical. Overall, these sentences all share one characteristic: the absence of any named persons. This indicates that classifications, such as the entity level, may be more useful since this would ensure that an entity is contained.

[13] https://gutenberg.org/

(1) This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever.

(2) Where the printed book showed translations or parallel versions side by side, the e-text has consecutive blocks of text.

(3) The Introduction of the Art of Printing

(4) His reports (de bello Gall.

(5) history of the celts.

(6) The Briton coins, several of which have survived, were imitations of the Roman ones and contain only Roman characters.

Example 3.2: Example of incorrectly labelled sentences from non-Wikipedia source.

These examples highlight that the labelling would probably not be useful for any further research. The examples indicate that the structure of Wikipedia sentences has had too much of a bearing on the training of the model. This is particularly evident in the 4th and 6th sentences: In the first case, this is similar to Wikipedia sentence fragments that show the place of birth for instance. In the second case, it seems as though the model has simply selected because of two entities (Roman and Briton) being present. This may point to over-training, which could be remedied by a smaller training set. In addition, the question is raised as to whether the sentence-level is too broad for this kind of classification, and whether a lower level, such as the entity-level may be more suitable.

3.4 Conclusion

This chapter has described two information extraction pipelines using Wikipedia and Wikidata. In addition, it has established a method for processing Wikipedia to use in these contexts, and investigated how well Wikipedia and Wikidata work together. Furthermore, it has presented a preliminary investigation on the use of automatically annotated data to train a neural information extraction model. As such, the research presented in this chapter is aimed at addressing **RQ-1**:

RQ-1 *Is Wikipedia a suitable data source to facilitate the extraction of biographical information?*

RQ-1a *Which processing steps are required to use Wikipedia?*

RQ-1b *Could Wikipedia and Wikidata be used in an automatic annotation approach for training data?*

With regard to **RQ-1**, evidence has been presented in this chapter that allows for a partial answer: In the context of a limited project, Wikipedia is useable as a source of data to extract biographical information from. Section 3.2 presented a processing approach for Wikipedia and Wikidata to extract limited biographical information for a biographical dictionary. The processing approach answers **RQ-1a**, concerning the kind of processing that is needed to work with Wikipedia in the context of biographical information extraction.

In the past, knowledge-based or rule-based approaches to IE were widely used, which has been highlighted in Section 2.2. Although these approaches often

require task and text-specific engineering, they often present a good starting point for information extraction. Additionally, for the pipeline described in Section 3.2, this meant that a better understanding of the structure of Wikipedia articles was required. Therefore, a rule-based approach was appropriate.

The NLP pipeline presented has been used to identify biography candidates and to extract information about them from Wikipedia and Wikidata. It has been shown that shallow extraction methods work well for obtaining basic information about biography candidates. However, for determining possible relevant candidates more accurate metrics probably need to be used. While the simple rule-based method extracts information well, it is quite limited and would require further engineering efforts for it to be usable. Nevertheless, in the context of the research project it was carried out in, it provided the required results.

The other finding of this chapter concerns **RQ-1b** and the implementation of a classifier for biographical information sentences using a neural model. A method was presented for automatically compiling textual data for training a neural model, which showed good evaluation results, but produced a lot of noise. With an error rate of around 20% to 40%, it is questionable how useful such an approach could be to historians, as this would entail a lot of work removing irrelevant sentences. Nevertheless, most of the evidence pointed towards the fact that this approach may work if not focused on whole sentences, but rather on smaller units. This could be in the form of entity pairs linked by a relation, which will be the basis for the approach presented in Chapter 4.

CHAPTER 4

GUIDED DISTANT SUPERVISION FOR BIOGRAPHICAL RELATION EXTRACTION

As web technology continues to thrive, documents containing biographical information are continuously generated and published online in large numbers (Nasar et al., 2021). These online documents contain essential facts or information about events related to the lives of well-known and lesser-known individuals, which can be used to populate structured biographical databases (Wang et al., 2021; Smirnova and Cudré-Mauroux, 2018). These databases are capable of supporting many interesting studies in humanities, and related areas (Zhang et al., 2017). However, manually extracting information from a massive document collection is impossible, given the amount of information available online. Therefore, NLP methods, especially ones that extract information, can be used to process these documents automatically.

Although NLP methods are suitable, a major weakness of some studies relating to biographical information extraction is that they can not be used directly to populate databases or create knowledge graphs. Instead, they need to be combined with other NLP techniques to extract the structured information required for databases. A different approach, presented in this chapter, is to design biographical information extraction as a *relation extraction* (RE) task. RE, a

subtask of IE, is the task of extracting semantic relations between entities from a document, which can in turn be used to populate a database with relational facts contained in a piece of text.

Sample 1: William Shakespeare was born and raised in Warwickshire. At the age of 18, he married Anne Hathaway, with whom he had three children: Susanna Hall and twins Hamnet Shakespeare and Judith Quiney.

Sample 2: Henry Baynton (23 September 1892 in Warwickshire 2 January 1951 in London) was a British Shakespearean actor of the early 20th century.

Example 4.1: Text samples of sentences expressing semantic relations between two entities.

For the texts shown in Example 4.1, the RE model can extract triples. In the example at hand, a triple would be <*William Shakespeare*, SPOUSE, *Anne Hathaway*>. Table 4.1 shows further relationship triples that can be extracted from the above two text pieces. These triples can be represented as edges in a knowledge graph for instance. Knowledge graphs are commonly used to provide information to end-users, and for understanding relationships between various types of entities. By combining such triples, a system can produce a knowledge graph of relational facts between persons, occupations, and locations in a text. This is demonstrated by Figure 4.1, which shows a knowledge graph derived from the relationships in Table 4.1.

This chapter introduces the concept of *guided distant supervision*, a novel approach for producing RE datasets which is based on a semi-supervised approach

Entity	Relation	Entity
William Shakespeare	*Birth Place*	Warwickshire
William Shakespeare	*Spouse*	Anne Hathaway
William Shakespeare	*Child*	Susanna Hall
William Shakespeare	*Child*	Hamnet Shakespeare
William Shakespeare	*Child*	Judith Quiney
William Shakespeare	*Occupation*	Actor
William Shakespeare	*Occupation*	PlayWright
Henry Baynton	*Occupation*	Actor
Henry Baynton	*Birth Place*	Warwickshire

Table 4.1: Example of biographical relationship triples.

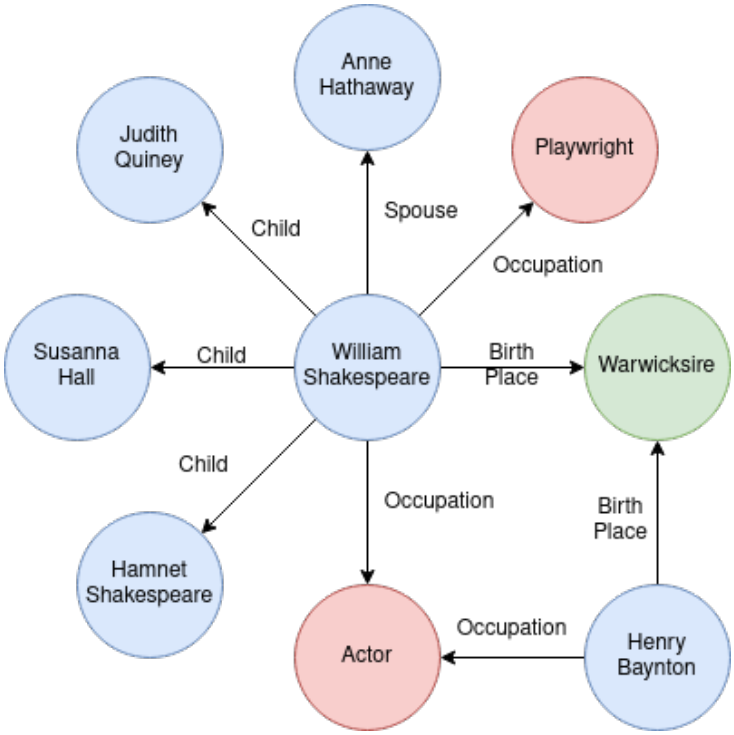

Figure 4.1: Example of a knowledge graph.

and can potentially be expanded easily to other domains and languages (see Chapter 5). At the time of writing this thesis, an approach such as this has not previously been proposed. To demonstrate how effective such an approach can

be, the first dataset of this kind has been developed and evaluated. If the approach is useful, it can significantly reduce the burden on the manual annotation process, as well as language and domain-specific expertise. The main contributions[1] of this chapter are the following:

(1) *Guided distant supervision* is proposed and used to compile *Biographical*, the first and largest dataset for biographical RE with ten relationship categories. Additionally, a manually annotated subset that can be used for evaluation is produced.

(2) Four machine learning models to perform biographical RE are evaluated, based on state-of-the-art transformer models such as BERT (Devlin et al., 2019) and using different learning approaches.

(3) Important resources are provided to the community: the dataset, the code, and the pre-trained models are made available to everyone interested in working on biographical RE using the same methodology.

In this chapter, Section 4.1 presents an overview of related work. Section 4.2 describes the data compilation process involved in this study. In Section 4.3, the various training experiments are explained, as well as an evaluation of the experiments. The evaluation section is concluded with an in-depth error analysis, and finally Section 4.4 provides conclusions.

[1] Parts of these contributions have been discussed in Plum et al. (2022).

4.1 Related Work

Previous studies have used many NLP techniques including text classification (Aprosio and Tonelli, 2015; Hogue et al., 2014), NER (Jiang, 2012) and summarisation (Zhou et al., 2004) to perform biographical information extraction. Biadsy et al. (2008) used an unsupervised sentence classification framework to extract biographies from Wikipedia articles. In more recent work, Aprosio and Tonelli (2015) have trained various machine learning classifiers to detect biographical sections in Wikipedia texts using a supervised approach.

As mentioned, there are two primary components in a traditional IE system; entity extraction and relation extraction (Finkel et al., 2005). Mintz et al. (2009) and Riedel et al. (2010) use pipeline approaches where a named entity recognition (NER) system is used to identify the entities in a sentence, followed by a classifier to determine the relation between them. However, due to the complete separation of entity extraction and relation extraction, these models miss the interaction between multiple relation tuples present in a sentence (Nayak and Ng, 2020) and can also propagate errors from one component to the other (Zheng et al., 2017). Due to this limitation, Nayak and Ng (2020) have suggested joint entity and relationship extraction for IE systems. These approaches extract triplets that contain two entities, and the relationship between them.

More recent work in biographical information extraction has modelled the task as a RE problem. Early approaches for RE were based on traditional machine learning models such as support vector machines (Liu et al., 2007), and decision

trees (Singhal et al., 2016). But with the introduction of word embeddings and the success of neural network architectures in different areas, the NLP community has used a wide range of neural network architectures for RE. Zeng et al. (2014) have used a convolutional neural network (CNN) architecture and a synonym dictionary to integrate semantic knowledge into the neural network. In a different approach, Zeng et al. (2015) use lexical features with the word embeddings (Turian et al., 2010) fed into a CNN to perform RE. Shen and Huang (2016) also used a CNN architecture. Recurrent neural networks (RNN) have also been popularly used in RE. Miwa and Bansal (2016) utilised a Tree Long Short-Term Memory (LSTM) network to perform RE. Zhou et al. (2016) used an attention-based bi-directional LSTM network on the SemEval-2010 relation classification task (Hendrickx et al., 2010) and show that it provides good results.

The current state-of-the-art in RE, also used for this research, is based on neural transformers (Baldini Soares et al., 2019). These transformer models are trained using a language modelling task such as masked language modelling or next sentence prediction. Trained transformer models have been used successfully to perform RE as a downstream NLP task. Results on recent RE datasets show that transformers outperform the previous architectures based on RNNs and CNNs (Xue et al., 2019; Baldini Soares et al., 2019).

All the ML models for RE mentioned above follow a supervised paradigm where an annotated dataset is required to train the ML model. The most common datasets used for this are NYT24 (Hoffmann et al., 2011), NYT29 (Riedel et al., 2010) and TACRED (Zhang et al., 2017). All these datasets have been created

using manual annotation. As mentioned before, since the annotation process is expensive, these datasets are limited in size. For example, TACRED, the largest RE dataset, has only 106,264 instances. This can prove not enough to train data-driven methods, especially those based on neural networks. Furthermore, the manual annotation process limits the expansion of RE research to different domains and languages.

It is clear that neural models provide excellent results when applied to RE. However, it also stands out that there is a need for labelled data for best performing models, a problem that is difficult to solve due to reasons presented in this section. To address the data problem, Section 4.2 introduces guided distant supervision, to allow for automatic compilation of annotated data, but with reliable precision for the labels. The approach is based on distant supervision first introduced by (Mintz et al., 2009), and is similar to Chisholm et al. (2017).

4.2 Guided Distant Supervision

The method for compiling the data used here is related to distant supervision, and will be referred to as *guided distant supervision*. Since the target here is biographical information, relations that provide distinctive information about a person were used. While these relations are by no means exhaustive, they present enough information about an individual to distinguish them clearly. This was assessed from feedback provided by multiple historians, during the APIS project (Section 3.2), as well as during the work on this dataset (Plum et al., 2022). The various relations and their labels are presented in Table 4.2.

Relation	Label
Date of birth	BIRTHDATE
Place of birth	BIRTHPLACE
Date of death	DEATHDATE
Place of death	DEATHPLACE
Place of education	EDUCATED
Job title	OCCUPATION
Has a parent	PARENT
Has a sibling	SIBLING
Has a child	CHILD
No relation	OTHER

Table 4.2: Overview of the relations and their respective labels.

As such, the approach is divided into two steps: The first step involves the selection of the data sources, which is one of the most fundamental aspects of the approach (Section 4.2.1). The approach requires a source of textual data and a source of structured information that is related to the textual data. For the second step, the data sources need to be processed, and matching operations that allow for the automatic labelling process (Section 4.2.2) need to be defined. Different approaches to processing the textual data can potentially affect the final dataset (Section 4.2.3). These steps lead to the final dataset consisting of sentences, marked entities and their respective relation, and can be used to train neural models (Section 4.2.4). Figure 4.2 shows a diagram of the system architecture.

A key aspect of GDS is the restrictive nature of this approach. GDS cannot be compared to an open information extraction oriented approach, where the goal would be to extract as much information as possible, including any kind of relation. With GDS this would not be compatible, since there would be too many variables. Therefore, specific structured information in sentences is targeted

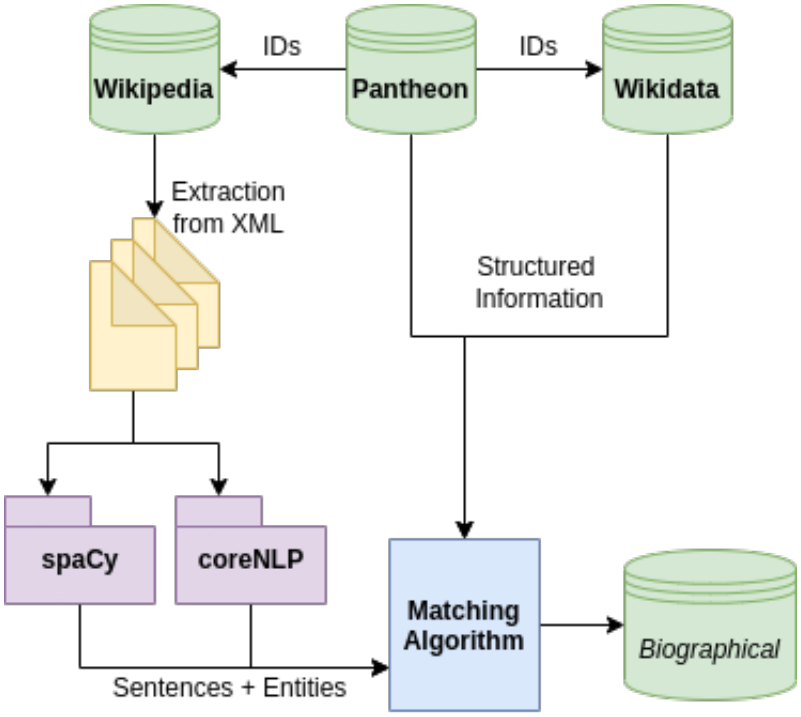

Figure 4.2: Diagram of the system architecture.

explicitly, and in turn limiting the relations that are extracted. In spite of the restrictions, this does not mean that further relations could be added to the dataset at a later point. This factor sets it apart from distant supervision, which targets relations in a more open information extraction-oriented approach.

4.2.1 Data Sources

Guided distant supervision combines data from various sources, three in the case of the example discussed here: Wikipedia, Wikidata and Pantheon. These datasets have been presented in Section 2.3. Wikipedia serves as the main source of textual data, in the form of sentences taken from specific articles. Pantheon and Wikidata serve as the sources of structured information. Additionally, Pantheon is used to

select the initial set of biographical articles from Wikipedia. Specific biographical articles in Wikipedia are targeted that are confirmed by the Pantheon dataset using Wikipedia IDs. The structured data from the Pantheon dataset is augmented with information from Wikidata. This expanded dataset is matched to sentences in Wikipedia, allowing for each sentence to be labelled according to the type of relation it contains. In the following sections, the data sources and their respective role in the process are discussed in further detail.

Wikipedia Wikipedia provides the textual data for the approach. For processing textual data from Wikipedia, parts of the previously established workflow, described in Section 3.2.1, are used. Wikipedia dumps are used, which are an exact copy of all Wikipedia articles of a given language at a specific point in time. For the dataset compiled here, the *enwiki-20190420* dump was used, which corresponds to the content of English Wikipedia on 20th of April 2019. Once downloaded, articles corresponding to the entries in the Pantheon dataset are extracted, which is done via the Wikipedia IDs. Extracting the text can be a complex task in itself, since the structure of the XML file is not uniform. In addition, the XML representations of the articles contain complex XML patterns which can be challenging to process (Attardi, 2015). Since the extraction of text from Wikipedia is not the main goal of this thesis and could warrant a separate project in itself, an existing tool was used for the extraction process: The wikiextractor[2] package for Python converts articles to plain text and is able to

[2] https://github.com/attardi/wikiextractor

select which complex XML structures to extract. Although the complex parts were left out for the purposes of this project, some extraction problems were observed. This includes XML-tag artefacts, mismatched quotation marks, and incomplete sentences, which are removed at the processing stage using regular expressions.

Pantheon The Pantheon dataset (Yu et al., 2016) is used as an indicative list of biographical entities that are available in Wikipedia and Wikidata. Although there are methods of determining biographical sections in Wikipedia articles (Aprosio and Tonelli, 2015), that was not the aim of the research at hand. With its criterion of availability of an entity across 15 languages, a certain notoriety of the entity and thus length of article is ensured. Longer articles are more desirable for this approach in order to gain the highest possible number of annotated sentences.

Each entity in the list includes some information such as dates of birth and death, places of birth and death, and main occupation. The included information allows the BIRTHDATE, DEATHDATE, BIRTHPLACE, DEATHPLACE and OCCUPATION relations to be labelled, as well as confirming the name of a person, which can vary slightly. Furthermore, each entity is linked to Wikipedia and Wikidata, making retrieval of further information from Wikidata straightforward, in addition to indicative text retrieval from Wikipedia.

Wikidata Wikidata is used to augment the information from Pantheon. Using the corresponding entries, the EDUCATED, PARENT, SIBLING and CHILD relations are added for this project. Finally, an important part of the labelling concerns the

OTHER tag. Without it, the system would always classify one of the nine relations, without being able to label relations that are not of interest. For this reason, any other information from Wikidata that is matched within a text is assigned the OTHER label. It is feasible that this approach could be refined for future iterations, in order to minimise overlap with actual important information, however, this was not further looked into for the purposes of this research.

4.2.2 Automatic Labelling

The next step in the approach is the automatic labelling of sentences, which is again derived from a previous pipeline (see Section 3.3.1). Once the text of each Wikipedia article is extracted, the texts are processed using spaCy[3] NER to tag persons, locations, organisations, dates, as well as Stanford CoreNLP (Manning et al., 2014) entity information to tag occupations in each article. It should be noted that here, spaCy is run during compilation, whereas one full annotation run with Stanford CoreNLP on all articles was carried out, which was stored and accessed subsequently. This is because Stanford CoreNLP was very slow for multiple runs, while spaCy was much faster (see Section 3.2.2).

Approach Each sentence of an article is processed in order to determine whether it features the main target entity or subject of a given article. To accomplish this, the script matches the name with the person tags in the sentence, and also allows some substring matches, such as first and last name excluding any other titles, or last name only. If a match is found, the sentence is regarded

[3] https://spacy.io

as containing some information about that person. This is ensured because the sentence is taken from that person's article and it includes that person's name.

After a positive match is made within a sentence, the other tagged entities (locations, organisations, dates and occupations) in the sentence are checked against the information provided by the Pantheon dataset and the respective Wikidata entry. Each matched pair, for instance a *name* and a *location*, is then marked with <eN> (begin) and </eN> (end) tags, where N is either 1 or 2, depending on the position of the entity (i.e. first or last). This is followed by the respective relation tag. Example 4.2 shows sample sentences that have been tagged in this way.

BIRTHPLACE **<e1>**William Shakespeare**</e1>** was born and raised in **<e2>**Warwickshire**</e2>**.

DEATHDATE **<e1>**Henry Baynton**</e1>** (23 September 1892 in Warwickshire **<e2>**2 January 1951**</e2>** in London) was a British Shakespearean actor of the early 20th century.

OCCUPATION Rolling Stone listed **<e1>**Burke**</e1>** as the 89th best **<e2>**singer**</e2>** of all time in 2008.

OTHER **<e1>**Titus**</e1>** was thereby a competitor for **<e2>**Vespasian**</e2>**.

Example 4.2: Examples annotated sentences, with the relation indicated, followed by the sentence with marked entity pair.

The assumption is made that this simple combination of named entity tagging and string matching works because of the controlled circumstances, which were mentioned at the beginning of this section. Only allowing matches involving

the person who is the main subject of an article ensures that statements made in sentences are most likely to be about this person. This may sound quite obvious at first. However, sentences taken from articles at random, matching random people, do not necessarily contain statements about that person. If the subject of the Wikipedia article is a certain person, most, if not all, statements made mentioning that person are likely to directly relate to that person. It is estimated that this approach could be extended to all relations where it would be possible to match the information in a sentence using the described method.

Another control mechanism involves the structure of Wikipedia. Often a number of opening paragraphs containing the most important information about a person (or other entity) are found. First mentions of certain facts are likely to be the key point of referral, such as the first date mentioned usually being the date of birth, first mentioned locations being the places of death and/or birth, job titles usually the corresponding (and main) occupation of the person etc. This structure can cause problems though, as will be explained in Section 4.3.1.

It is important to note that not every relation is always found for every entity. Therefore, different processing approaches for the textual data were tested, detailed in Section 4.2.3. A breakdown of the number of relations per set is presented in Section 4.3.1. Each relation also requires slightly different handling depending on the type of information. Tasks include date normalisation, partial matching for occupations, and exact location name matching. Exact details are presented in the following sections.

Date-based Relations This set of relations includes BIRTHDATE and DEATHDATE. In order to match these relations, *DATE* entities, which are normalised to *YYYY-MM-DD* format, are compared with the structured information. For this, the dateparser[4] package is used, with the date of processing as a relative date (in order to normalise rare cases such as *tomorrow* or *two weeks ago*). Only the first positive match for both date-based relations are used, discarding subsequent matches. This mode of processing, which assumes most pertinent information to be mentioned towards the beginning of a Wikipedia article rather than towards the end, is essential to the approach in general.

Name-based Relations This set of relations includes PARENT, SIBLING and CHILD, as well as BIRTHPLACE, DEATHPLACE and EDUCATED (the place of education). For these name-based relations, the *PER*, *LOC* and *ORG* tagged entities are compared with the structured information. It is ensured that only full matches are accepted, even though it may seem favourable to accept partial matches, at least for anything concerning persons. This is because with persons, it can be reasonable to allow just the first or last name to match. However, it was found during the manual annotation process (Section 4.3.1) that too many false matches occurred, caused by different persons having the same name, especially when members of the same family are mentioned in the same article.

Entity Information Relations Only the OCCUPATION relation is included in this group. Since spaCy's NER capabilities do not include any tags such as *title* or

[4] https://dateparser.readthedocs.io/en/latest/

job, Stanford CoreNLP's entity information processing was added to detect this relation. While the spaCy model could potentially have been trained to detect a new entity type, Stanford CoreNLP was chosen since adding a new relation to a model could also have introduced another layer of errors.

The matching algorithm for this relation functions in a similar way to the previous set of relations, except that the CoreNLP information is accessed for matching instead of the spaCy information. As mentioned, the initial CoreNLP processing is run separately, due to the increased run time. Again, only complete first matches are allowed for annotation. Potentially, this relation set could be extended by using further occupation information from Wikidata, which in most cases lists a number of different occupations for a person, rather than the one main occupation listed in Pantheon.

Other Relations This class of relations, labelled as OTHER in the dataset, is used for all other relations. It is essentially the zero class, that is labelled when all other lookups in a sentence have failed. The OTHER label is then applied to an entity pair that does not appear to be part of any of the other nine relations matched. Since the number of obtained OTHER relations can be disproportionately higher than all the other nine relations combined, a random selection of sentences is taken, balanced according to the total number of all other sentences containing relations. The OTHER relation class is balanced to make it equivalent in size with the remaining nine relations combined.

If in future more relations are added to the dataset, it would be vital to ensure

that these OTHER labelled sentences do not contain the new relations, since there could conceivably be overlap due to the nature of this class.

4.2.3 Processing Approaches

The process of automatically labelling each entity pair with a corresponding relation is carried out at the document and sentence levels of a relevant Wikipedia article. All the processing in terms of NLP (such as NER, POS-tagging, etc.) is done at the document level. Articles are then split into sentences, and each sentence is processed for automatic labelling, as described previously.

The correct processing of the articles plays a vital part in the approach, and as such, different approaches to processing may affect the final annotations. In order to investigate this hypothesis, two different approaches for processing were investigated next to the normal approach: First, the impact of adding running coreference resolution on the Wikipedia texts was tested, to ascertain whether it would yield more annotated sentences. Next, to increase sentence diversity, an approach was implemented that skips the first sentence of an article.

Coref Set With the *coref* set approach, the hypothesis is made that replacing co-referential entity mentions will result in the annotation approach finding more matches overall. This is due to the fact that more names would be matched because there are more in the text, compared to a non-treated sentence. Detecting more names could then potentially lead to more relation matches overall. To accomplish this, a coreference resolution algorithm[5] that works with spaCy was

[5] https://github.com/huggingface/neuralcoref

used to automatically replace named entity mentions with the most probable name. This is followed by the annotation step, which uses the same article text, but all co-referential entity mentions have been replaced with the target entity.

Table 4.3 shows the number of relations found across each of the compiled sets: *normal*, *coref*, and *skip* (which is described in the next section). The last line of the table shows the total number of relations found per set.

Relation	**normal**	**coref**	**skip**
birthdate	**52,083**	48,004	45,366
birthplace	**50,396**	46,552	19,746
deathdate	**17,376**	14,505	8,793
deathplace	19,055	**20,444**	11,202
occupation	**41,469**	**41,469**	17,642
parent	6,503	**10,301**	6,022
educated	5,738	**9,430**	5,694
child	2,343	**4,042**	2,215
sibling	2,189	**3,618**	2,098
other	197,952	**199,165**	119,578
Total	395,104	**397,530**	238,356

Table 4.3: Number of relations in each set, where *normal* was compiled with the standard processing method, *coref* with coreference resolution and *skip* omitting the first sentence of each article.

Comparing the total number of instances of each relation of the *normal* set and the *coref* set, a small increase can be observed. However, looking at the number of instances of the different relation types, it is clear that it is not a simple increase across each relation. In fact, fewer matches can be seen in certain cases. Upon further inspection, it was found that this was mainly due to the automatic replacement process producing illegible sentences through incorrect

replacements. The two main problems were mixed entity replacements and sentences being unintelligible because every single entity mention was replaced with one main entity, that was often too long in addition.

The main problems are shown in the examples below. In Example 4.3, an entity has been replaced many times, including an opening bracket. Cases like these were observed frequently, and with more characters added. These cases introduced matching errors in the set. In Example 4.4, a nested replacement can be observed, which causes similar matching problems.

Replaced: Born in **<e1>**Évreux **</e1>**, Eure, a great fan of **Paris Saint-Germain Paris Saint-Germain** since **<e2>Bernard Mendy </e2>(** childhood, **Bernard Mendy (** achieved **Bernard Mendy (** ambitions in 2000 when **Bernard Mendy (** joined PSG from SM Caen.

Original: Born in Évreux, Eure, a great fan of Paris Saint-Germain since his childhood, he achieved his ambitions in 2000 when he joined PSG from SM Caen.

Example 4.3: Example of multiple entity replacements.

Replaced: The hundreds of volumes contained **Queen Victoria's Queen <e1>Victoria</e1>'s's** personal views of [...]

Original: The hundreds of volumes contained **Queen Victoria's** personal views of [...]

Example 4.4: Example of nested entity replacements.

To understand better why this approach does not always work well, a manual annotation of 100 randomly selected sentences per relation from this set was carried out and is described in Section 4.3.1. A neural model was trained using this dataset, the evaluation of which is detailed in Section 4.3.3.

Skip Set The *skip* set was compiled to study the effects of omitting the first sentence of an article from Wikipedia. One problem with using Wikipedia texts stems from the first sentence of an article, or rather the structure of the first sentence of an article. Example 4.5 shows that the date of birth (and death) occur within parentheses after the name, followed by a short summary of the person.

> William Shakespeare (bapt. 26 April 1564 23 April 1616) was an English playwright, poet and actor, widely regarded as the greatest writer in the English language and the world's greatest dramatist.

Example 4.5: Typical first sentence of a biographical Wikipedia article.

This type of sentence structure (and content) is not only extremely frequent, but also quite specific to Wikipedia, suggesting that unnatural behaviour could be learned by a machine learning model. This was observed by Chisholm et al. (2017) who exploited this for their benefit. However, for this approach, the aim was to match as many representative sentences as possible. Therefore, a dataset was compiled that follows the previously described methodology, but skips the first sentence of each article. The hypothesis is that this forces more matches elsewhere in the article, where more representative sentences occur.

Table 4.3 (shown on page 105) shows the total and individual counts for each relation. Looking at the overall counts for each set, the skip set has much fewer matches than the other two sets, and it never has the highest number of individual counts in any category, although the numbers are comparable in some categories to the normal set. Regardless, some of the generally larger categories, such as BIRTHPLACE and BIRTHDATE are significantly smaller than the other two sets, generally pointing towards the fact that the assumption is true that this information is more common in the first sentence. It is not always certain that this information will appear later on in an article, therefore leading to a smaller number of matches.

As with the previous set, a manual evaluation of 100 randomly selected sentences per relation from this set is presented in Section 4.3.1, and the results of a trained neural model using this dataset in Section 4.3.3.

4.2.4 Neural Models

The machine learning model used for this project to perform relationship classification is based on transformers. Since their introduction, transformer models have shown excellent results in various NLP tasks (Devlin et al., 2019) such as text classification (Ranasinghe and Zampieri, 2020), NER (Jia et al., 2020), question answering (Yang et al., 2019) and RE (Yamada et al., 2020; Joshi et al., 2020; Wu and He, 2019; Alt et al., 2019). In this research, the architecture that was utilised was introduced by (Baldini Soares et al., 2019).

The input to the model is sentences with [E1] and [E2] markers for the respective entity positions. Then the output hidden states of transformer at the

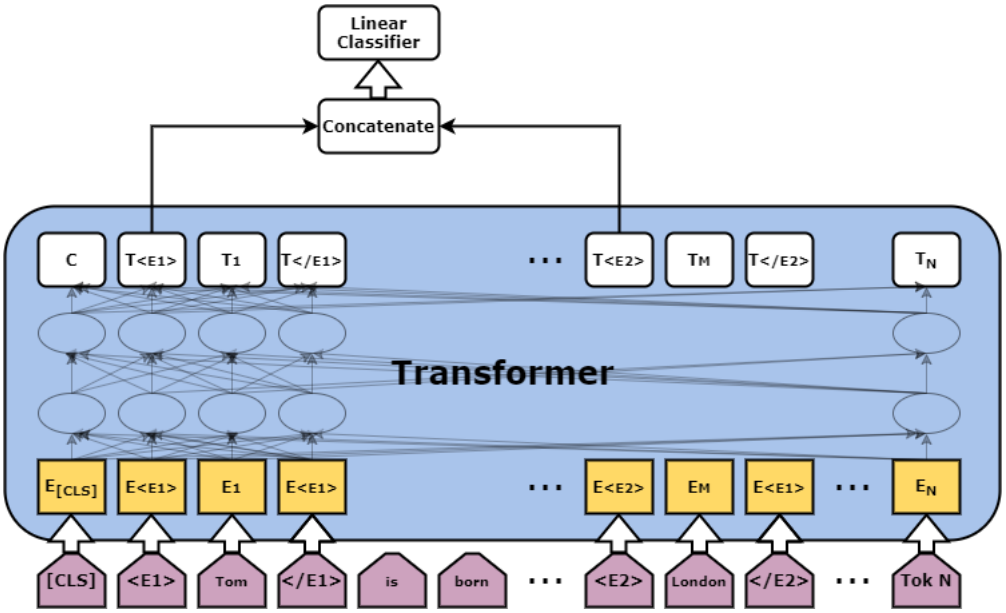

Figure 4.3: Diagram of the neural network architecture.

[E1] and [E2] token positions are concatenated as the final output representation of the relationship. Finally, a linear classifier is stacked on top of the output representation. A diagram of the architecture is displayed in Figure 4.3.

The parameters of the transformer and additionally the linear classifier are fine-tuned jointly by maximising the log-probability of the correct label. For all the experiments, the parameters were optimised with AdamW using a learning rate of $7e$-5, a maximum sequence length of 512, and a batch size of 32 samples. The models were trained using a 24 GB RTX 3090 GPU over five epochs. For the pre-trained transformer model, the *bert-base-uncased* model available in HuggingFace (Wolf et al., 2020) was used. For the pre-training experiments, the same model was used but re-initialised. For both pre-trained and non pre-trained experiments the parameters were kept the same to ensure maximum comparability.

It should be noted that better configurations may be found, but the purpose of this evaluation was to obtain a general insight into each training approach.

For training the pre-trained BERT-based classifier, each of the three sets (*normal*, *coref* and *skip*) was used separately. In addition, a combination of the three sets was used, referred to as *all*, where any duplicates caused by overlap were removed. The focus of this task was to ascertain whether the automatically compiled dataset works, rather than whether it achieves the best possible results.

4.3 Evaluation

Multiple experiments were carried out to estimate the quality and usefulness of this dataset. First, the effects of different processing approaches for the article texts were examined. The motivation for selecting different processing approaches was also discussed in Section 3.3.2. Next, a small sub-set of sentences was manually annotated to pinpoint potential problems and to create a gold-standard set for evaluation purposes. After re-running the compilation process, taking into account certain observations and minor processing improvements after manual annotation, various state-of-the-art ML models were trained using the dataset, and evaluated in terms of performance versus the gold-standard set.

4.3.1 Manual Annotation

The quality of the semi-supervised datasets was assessed before being used to train machine learning models, by means of manual annotation. This was important in order to find areas where the approach fails to match data accurately, where

processing methods do not work, or any other similar problems. In addition, a gold standard test set was required for benchmarking our neural models.

As pointed out in previous sections, 100 sentences per relation across the three datasets were extracted, equalling 3000 sentences in total that were manually annotated, and subsequently referred to as the Gold Set. The data was annotated by two persons, one native English speaker and one non-native but fluent English speaker, both postgraduate students. For each sentence, the task was to look at the relation assigned by our matching algorithm and add the correct relation if it had been labelled incorrectly. One of the nine indicative labels was to be used where appropriate, and the OTHER label if a different relation was expressed. The annotation guideline was that a human should understand by reading the sentence which relation is expressed, regardless of any prior knowledge they may have. This is demonstrated by Example 4.6.

Explicit: **<e1>**Renate Künast**</e1>** (born 15 December 1955) **is a** German **<e2>**politician**</e2>** of Bündnis 90/Die Grünen.

Implicit: A few months later **<e1>**Apollo Korzeniowski**</e1> died**, leaving **<e2>**Conrad**</e2> orphaned** at the age of eleven.

Unclear: Thus, **<e1>**Janaka**</e1>** tries to find the best husband for **<e2>**Sita**</e2>**.

Example 4.6: Examples of annotation scenarios.

The first example shows a sentence that clearly mentions **e2**, the OCCUPATION, of the entity **e1**. The second example shows an implicit relation, where the relation is not directly stated, but rather the word *orphaned* in relation to entity **e2** with

the statement that **e1** *died*, implies that **e1** is the *parent* of **e2**. The final example has been labelled as the PARENT relation. Although this may indeed be the case, and the annotator may have prior knowledge of this, or it has been expressed in a different sentence, it is not clearly stated in this sentence.

The inter-annotator agreement is 0.908 (Cohen's Kappa), indicating a very high agreement between the annotators. The annotations allowed for a number of observations. First, it is noticeable that two very similar relations work very differently. While BIRTHPLACE works extremely well across sets, DEATHPLACE does not. Upon further examination, it was found that the first mention of the place where someone died often was also the place where a person lived. In future, cases like these may warrant a different approach to processing by this approach, but for the purposes of the research at hand it was left unchanged. Second, it was observed that many relations in the *coref* set were incoherent and probably incorrect, due to imprecise replacements by the coreference resolution algorithm.

While Wikidata as a source does work quite well, categories can sometimes be ambiguous, such as the EDUCATED and PARENT relations. Here, it was observed that the Wikidata entries contained information that was not consistent with type of entry, such as EDUCATED containing a University that is the place of work rather than the place of education, or PARENT containing a person that the target entry is a parent of rather than being a person that the target entry has as a parent. Since this did not occur often during the manual evaluation, a strategy to solve this problem was not implemented.

Finally, a number of simple processing errors were found that were solved by

improving the regular expressions used for text cleaning. The matching procedure for the OCCUPATION relation was also adjusted, so that matches were avoided where the occupation related to a different entity. This led to a slightly smaller number of relations overall, with a detailed overview shown in Table 4.4.

	normal	**coref**	**skip**
birthdate	**51,524**	47,977	45,211
birthplace	**50,226**	46,551	17,537
deathdate	**17,197**	14,500	5,925
deathplace	18,944	**20,430**	10,790
occupation	**18,114**	**18,111**	8,716
parent	6,352	**10,291**	5,596
educated	5,639	**9,415**	3,858
child	2,209	**4,053**	2,123
sibling	2,083	**3,601**	1,997
other	173,969	**175,916**	103,248
Total	346,257	**350,845**	205,001

Table 4.4: Relations per set after processing improvements.

Overall, the following conclusions were drawn for each set. The *normal* approach works well, while not offering a very diverse set of sentences. As alluded to earlier, it is clear that this approach matches mainly the standard Wikipedia first sentence, as described in previous sections. The *coref* set, while seemingly the largest set, also includes a large number of unusable sentences and incorrectly replaced entities. During the course of the evaluation, it was found that the coref set is not always explicit and sometimes difficult to understand due to the incorrect replacements. Finally, it was found that the *skip* set is very mixed in terms of accurate annotations. While for some relations it seems that none of the matches

returned were usable, other relations seem to have worked very well, offering in addition a wide variety of different sentences demonstrating the desired effects.

	NORMAL			COREF			SKIP		
Relation	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
birthdate	1.0	1.0	**1.0**	.99	1.0	.99	1.0	1.0	**1.0**
birthplace	.84	.90	**.87**	.86	.88	**.87**	.79	.83	.81
deathdate	1.0	.99	**1.0**	.98	1.0	.99	.94	.99	.96
deathplace	.37	.95	**.53**	.31	1.0	.48	.36	.97	**.53**
educated	.88	1.0	.94	.92	.99	.95	.96	.99	**.97**
occupation	.80	1.0	**.89**	.90	1.0	.85	.68	1.0	.81
parent	.77	.99	.87	.73	1.0	.85	.80	1.0	**.89**
sibling	.75	.95	**.84**	.62	1.0	.77	.70	.92	.80
child	.80	.99	**.88**	.64	1.0	.78	.63	1.0	.77
other	.97	.37	**.54**	.98	.36	.53	.96	.33	.49
macro avg.	.82	.91	.83	.79	.92	.81	.78	.90	.80

Table 4.5: Evaluation of the automatic labels on the gold set versus the manually assigned labels.

In order to determine the performance of the automatic compilation approach, an evaluation of the automatically assigned labels versus the manually assigned labels was carried out on the gold set. Table 4.5 shows the results of the evaluation for each set. Sentences that contained processing errors caused by conversion to plain text, incorrect replacement of anaphors and spaCy tagging errors totalled about 100 and were removed. Since these would all have been annotated as OTHER, it was decided to remove these sentences since they could have caused an unbalanced test set. Looking at the results most of the matches found are correct, indicated by high precision and recall scores. However, the problem with DEATHPLACE that was observed during the evaluation is confirmed here.

In addition, recall drops down for the OTHER class, mainly due to incorrect classifications by the automatic labelling approach to begin with.

4.3.2 Baseline Results

To put the results of the neural models into context, Table 4.6 presents the results for two baseline approaches. The approach used for the first baseline was a Naïve Bayes system using a bag-of-words approach. The second approach was a SVM based system using TF-IDF for the words. In both cases, preprocessing was kept to a minimum, including only tokenisation and lower-casing the words. Each baseline system used whole sentences as input, including the entity markers.

	NB			**SVM**		
Relation	**P**	**R**	**F1**	**P**	**R**	**F1**
birthdate	.72	.50	.59	.45	.62	.52
birthplace	.58	.54	.56	.49	.77	.60
deathdate	.61	.45	.52	.69	.33	.44
deathplace	.25	.61	.36	.40	.41	.41
educated	.76	.74	.75	.90	.50	.64
occupation	.78	.44	.56	.61	.26	.37
parent	.49	.53	.51	.67	.09	.16
sibling	.32	.52	.40	.00	.00	.00
child	.23	.55	.32	1.0	.01	.02
other	.55	.38	.45	.47	.89	.62
macro avg.	.53	.52	.50	.57	.39	.38

Table 4.6: Baseline results with Naïve Bayes and SVM approaches.

Overall, comparing the baseline results with the neural results (see following section) indicates that the latter is much more accurate. Specifically, the baseline methods perform quite badly when identifying relations that require detection of

a single reference to a named entity (rather than multiple entities, such as a name and a date). Another interesting result is that the Naïve Bayes approach appears to outperform the SVM approach, particularly with regard to recall. While the SVM approach does score better in terms of performance, it seems to be unable to predict the SIBLING label, although it performs excellently on the CHILD label. The Naïve Bayes approach, on the other hand, is more balanced regarding the SIBLING and CHILD labels. While this may appear counter-intuitive, since SVM is generally regarded as a more performant approach, the better results of Naïve Bayes can be explained due to the absence of any feature-engineering for the SVM approach. In addition, the BoW approach of Naïve Bayes may achieve better results due to there being distinct trigger words for some relations, implying that this approach would work well.

4.3.3 Neural Model Results

To assess the impact of pre-training models on purpose built pre-training data versus models that were only fine-tuned (using already available pre-trained models), two approaches were taken: the results are split into *fine-tuning*, where a pre-trained BERT model was fine-tuned on the datasets compiled using GDS and *pre-training*, where blank BERT models were pre-trained on separately compiled pre-training sets.

Fine-tuning Table 4.7 shows the evaluation results of the pre-trained models fine-tuned on the three different sets, with Table 4.8 showing the results of the

model fine-tuned on all three sets combined. Overall, these results correspond to those of the evaluation of the human annotations against the automatic compilation process. This is to be expected since the behaviour is learned from the dataset by the model. Furthermore, the model performs relatively poorly in terms of recall when detecting some other relations, including CHILD, PARENT and SIBLING. When comparing to the counts per set (see Table 4.4) these relations are quite low in number compared to the others, possibly explaining the results.

	NORMAL			COREF			SKIP		
Relation	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
birthdate	1.0	.99	**.99**	1.0	.99	**.99**	.74	.95	.83
birthplace	.90	.91	**.91**	.89	.94	**.91**	.79	.92	.85
deathdate	.95	.97	**.96**	.94	.99	**.96**	.95	.71	.81
deathplace	.30	.73	**.41**	.30	.74	**.41**	.22	.44	.29
educated	.93	.82	.87	.91	.94	**.92**	.91	.93	**.92**
occupation	.70	1.0	.81	.70	1.0	**.82**	.70	1.0	**.82**
parent	.81	.73	.77	.79	.79	**.79**	.79	.76	.78
sibling	.85	.56	.67	.74	.63	**.68**	.77	.60	**.68**
child	.92	.51	.60	.76	.71	**.74**	.84	.57	.64
other	.77	.73	**.75**	.84	.68	**.75**	.77	.64	.70
macro avg.	.82	.80	.78	.79	.85	**.80**	.75	.76	.74

Table 4.7: Evaluation metrics for the relations in the *normal*, *coref* and *skip* sets.

Pre-training Blank models were trained in addition to the pre-trained BERT-model, to investigate the impact of training the model from scratch. For this, different datasets were used for pre-training. The first was based on the CNN dataset (Nallapati et al., 2016), the second and third based on the same Wikipedia articles from which the training dataset was compiled. The CNN and small

Relation	ALL		
	P	**R**	**F1**
birthdate	1.0	1.0	1.0
birthplace	.91	.90	.91
deathdate	.93	.96	.95
deathplace	.32	.99	.46
educated	.93	.94	.93
occupation	.70	1.0	.81
parent	.73	.84	.77
sibling	.92	.56	.67
child	.75	.63	.68
other	.70	.66	.68
macro avg.	.80	.86	.80

Table 4.8: Evaluation metrics for the relations in the *all* set

Wikipedia dataset were 3.6MB in size, while the larger Wikipedia dataset was 10MB. Originally, the aim was to test larger sized versions of the Wikipedia dataset, however, these attempts were abandoned due to very long training times, with the 10MB set taking roughly a month to pre-train. The justification for larger training sets would be improved performance. After pre-training, each model was fine-tuned using the *normal* training set as described earlier.

Table 4.9 shows the evaluation results of BERT models pre-trained on the three different pre-training corpora described above, and all fine-tuned on the normal set. While there are only minor differences between each model, the results indicate that the model trained on the smaller Wikipedia pre-training corpus performs best. It appears that increasing the size of the pre-training corpus does not result in better performance of the model overall, although some models are marginally better at individual relations, for instance the DEATHPLACE relation.

	CNN			SMALL			MEDIUM		
Relation	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
birthdate	1.0	.99	.99	1.0	1.0	**1.0**	.98	1.0	.98
birthplace	.88	.90	.89	.89	.93	**.91**	.86	.90	.88
deathdate	.94	.97	.95	.94	.99	**.96**	.95	.98	.96
deathplace	.28	.48	.35	.33	.73	.44	.36	.75	**.47**
educated	.90	.82	.86	.93	.92	**.92**	.90	.92	.91
occupation	.70	1.0	.82	.71	1.0	**.83**	.70	1.0	.82
parent	.78	.76	**.77**	.73	.76	.73	.65	.73	.67
sibling	.80	.46	.55	.84	.63	**.71**	.71	.59	.64
child	.85	.50	.61	.80	.59	.66	.94	.65	**.74**
other	.73	.70	.72	.79	.70	.74	.81	.69	**.75**
macro avg.	.79	.76	.75	.79	.83	**.79**	.78	.83	.78

Table 4.9: Evaluation metrics for the relations in each of the models that were pre-trained.

4.3.4 Error Analysis

Although the evaluation of the various models gives an indication of how well each model performs, it is evident when taking the results into consideration that the differences are in some cases marginal. Therefore, this error analysis aims investigate the smaller differences between the models, depending on each label. As before, the models are treated separately depending on the training category.

Fine-tuning Looking at the confusion matrices for all four fine-tuned sets, displayed in Figure 4.4, the trends are mostly similar for all four models. In particular, the OTHER label is most often misclassified, where it is most often assigned the DEATHPLACE and OCCUPATION tags. The fact that the models are unable to correctly detect this relation is probably due to its very wide scope, and the fact that DEATHPLACE is often assigned can be explained by the low number of samples in the training set. Other frequently misclassified tags

include the relations between persons, PARENT, SIBLING and CHILD, which were usually assigned the OTHER or one of the other person tags. These relations are very similar in terms of sentence structure, which probably leads to incorrect predictions. Finally, it is important to point out the fact that the *skip* model also misclassifies DEATHDATE very often, mainly due to the much smaller size of the training set.

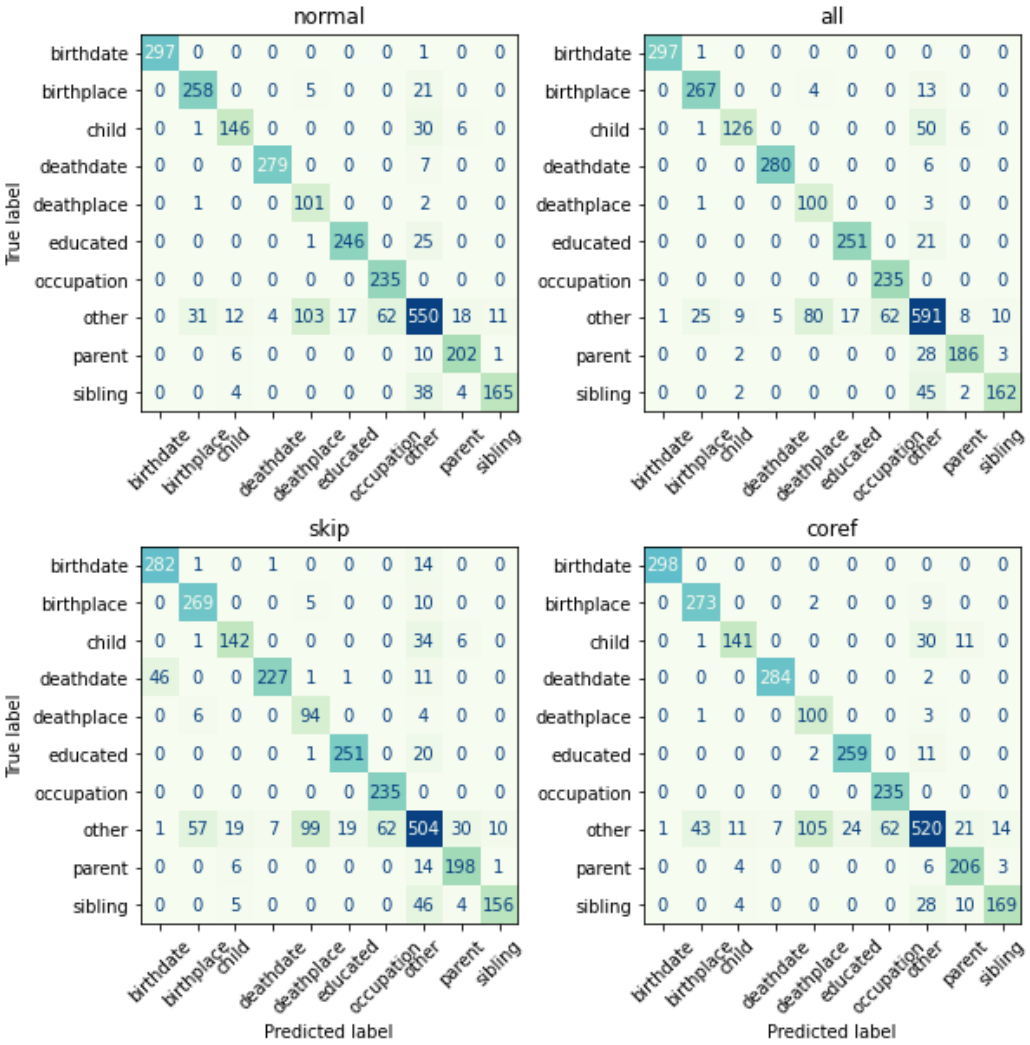

Figure 4.4: Confusion matrices for all four fine-tuning models.

Comparing each model in terms of agreement on classifications, there were 687 cases out of the 2900 gold sentences where at least one model predicted an incorrect label. Out of the 687 cases, there were 277 where all four sets predicted an incorrect label. Of these, 186 were with the OTHER label, with 33 for SIBLING, 25 for CHILD, 12 for PARENT, 10 for EDUCATED, 7 for BIRTHPLACE, as well as 2 for DEATHPLACE and DEATHDATE each.

The second most frequent cases, at 230, were those where only one model predicted an incorrect label. Here again, OTHER was most frequent with 69, followed by 56 for DEATHDATE, 29 for CHILD, 19 for PARENT, 16 for BIRTHDATE, 14 for SIBLING, 11 for EDUCATED, as well as 8 each for DEATHPLACE and BIRTHPLACE. The model that made the most errors while the others did not was the model trained on the *skip* set, with 130 incorrect labels. Interestingly, this was followed by the *all* set with 45, which is surprising since it is overall the best performing set in terms of F1-score. The *coref* set was incorrect in 29 cases, with the *normal* set in 26 cases. This quite clearly follows the reported precision values for the models, with 82% for the master set, which makes almost the least amount of single errors.

Taking a closer look at which model diverged from the other models the most in this category, Figure 4.5 presents a break down per label. The *skip* set is present for most labels, but mostly with OTHER and DEATHDATE, which does not produce many of the errors made by the other models. This could be explained by the lack of examples of this relation in the training set, which is due to the processing method. Another observation of note is the relatively high number of incorrect

labels produced by the *all* model on the CHILD and PARENT relations. In contrast to *skip*, this could be due to an overabundance of training examples, due to the fact that this model is trained on a combined dataset. As these relations are quite similar, this could indicate that too many examples of two similar labels confuses the classifier.

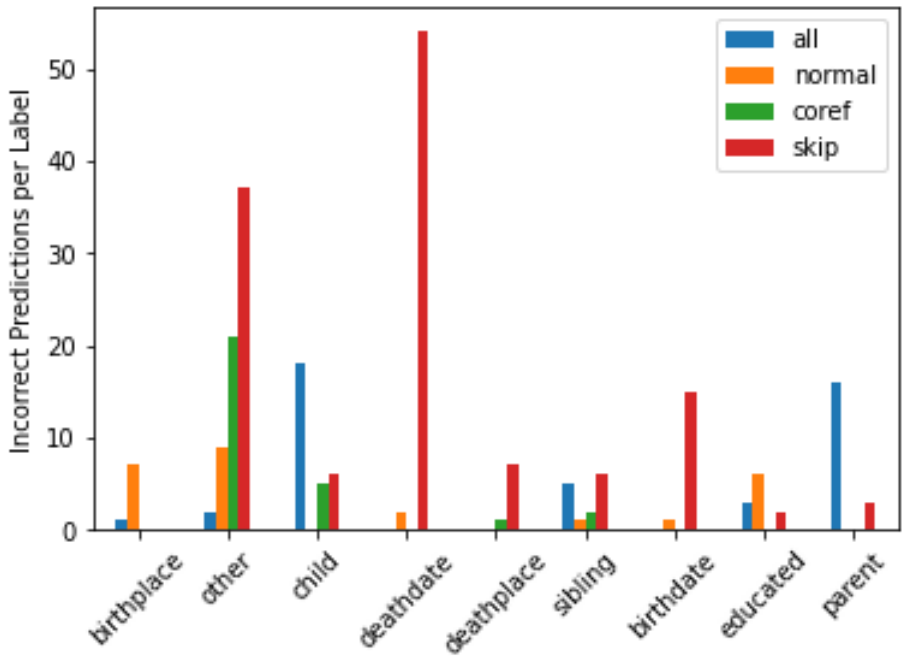

Figure 4.5: Number of times **one** model diverged from the other models (per label).

Next were cases where two models predicted an incorrect label, accounting for 95 cases. As before, OTHER was most frequent with 43 cases, followed by CHILD, SIBLING, EDUCATED, BIRTHPLACE, PARENT, DEATHDATE, DEATHPLACE and BIRTHDATE, all below 10. The models were almost equally represented, with 64 for *skip*, followed by 46 for *normal*, 43 for *coref* and 37 for *all*.

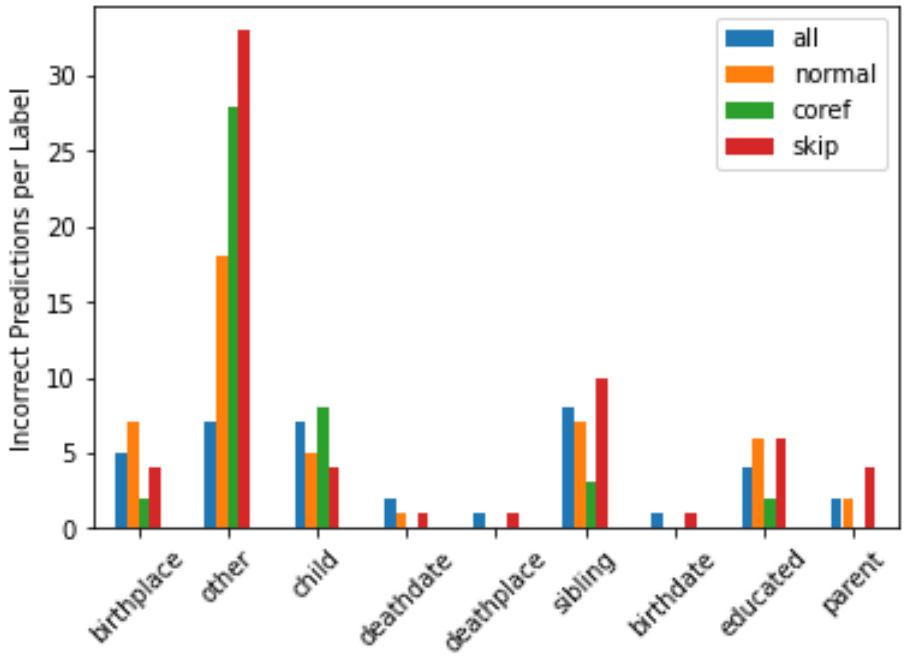

Figure 4.6: Number of times **two** models diverged from the other models (per label).

Finally, cases where three models made an incorrect prediction, occurred with a frequency of 85. Similarly to the mistakes made by two models, the most frequent by a large margin was OTHER with 56 cases. The labels CHILD, SIBLING, BIRTHPLACE, EDUCATED, PARENT, DEATHDATE and DEATHPLACE all followed with below 10. The models were also represented almost equally, this time with 72 for *normal*, 71 for *skip*, 66 for *coref* and 46 for *all*.

Figures 4.6 and 4.7 show the number of times each model diverged per label in the cases where two and three models diverged, respectively. These charts do not show any noteworthy trends, but underline the results discussed.

Overall, sentences labelled in the gold set with OTHER caused the most problems by far. This is unsurprising when taking into account the high precision

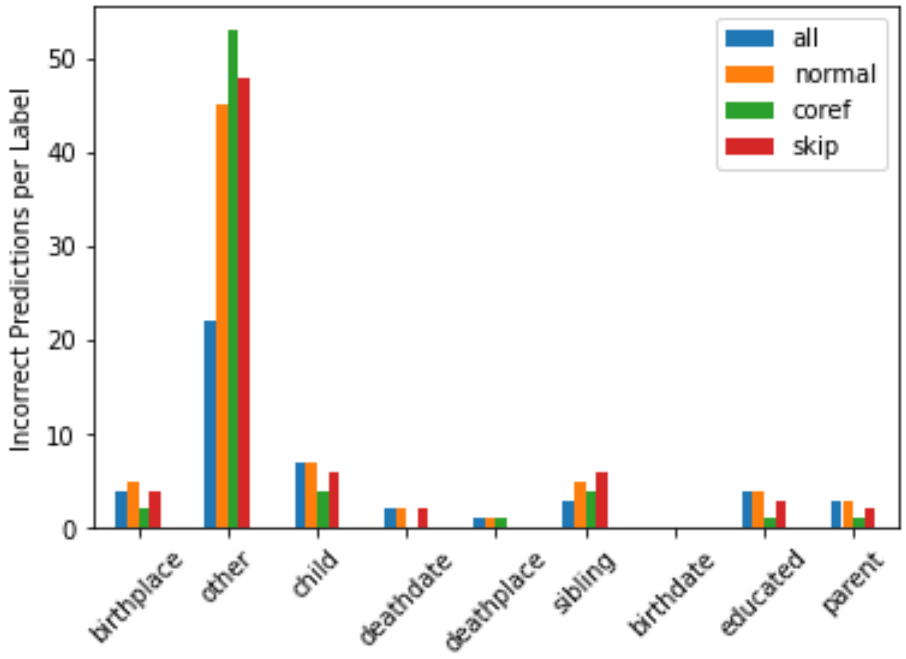

Figure 4.7: Number of times **three** models diverged from the other models (per label).

and quite low recall metrics reported from the gold set evaluation (see Table 4.5). It is also important to note the problems caused by labels that require two names, such as PARENT, SIBLING and CHILD, which overall accounted for the second largest number of mistakes. A possible explanation for this is the similarity of these labels and indeed the sentences themselves, which have almost equivalent structures.

Pre-training Figure 4.8 shows the confusion matrices for all three pre-trained models. The trends are almost identical to the four fine-tuned models, with the OTHER label being misclassified most often, where it is most often assigned the DEATHPLACE and OCCUPATION tags. Other frequently misclassified tags include the relations between persons, PARENT, SIBLING and CHILD, which were usually

assigned the OTHER or one of the other person tags.

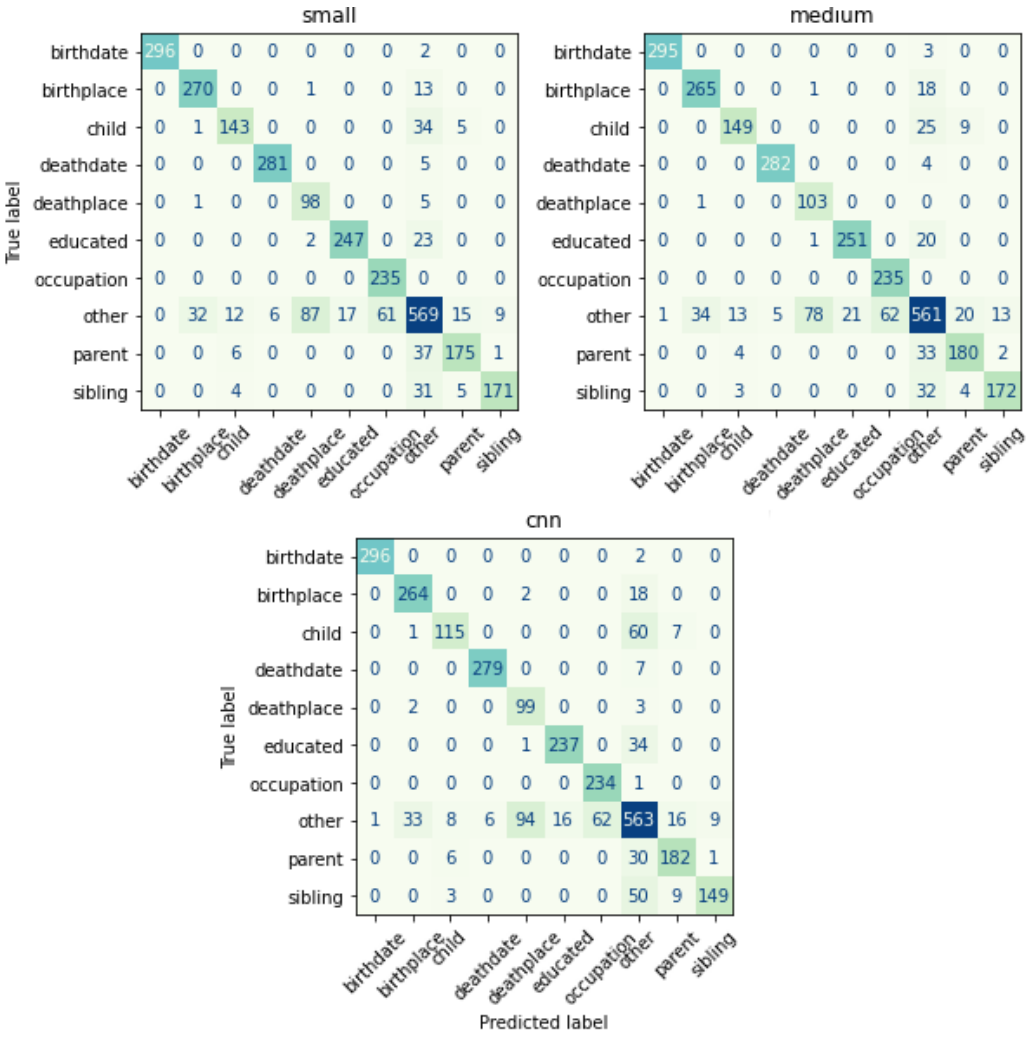

Figure 4.8: Confusion matrices for all three pre-trained models.

Comparing the three pre-trained sets in terms of agreement on classifications, in 583 cases out of 2900 at least one model predicted an incorrect label. Of these, 311 predictions occurred where all three models were incorrect. This included 191 cases with OTHER, 35 with SIBLING, 27 each for CHILD and PARENT, 15

for EDUCATED, 10 for BIRTHPLACE, 4 for DEATHDATE, as well as 1 each for BIRTHDATE and DEATHPLACE.

Second most frequent were cases were one of the three models made an incorrect prediction, with 173 cases. Of these, 58 were for OTHER, with 26 for SIBLING, 25 for CHILD, 23 for PARENT, 16 for EDUCATED, 15 for BIRTHPLACE, 4 for DEATHDATE, 3 for DEATHPLACE, 2 for BIRTHDATE and 1 for OCCUPATION. The model that diverged the most was the *cnn* model with 96 cases, followed by *small* with 40 cases and *medium* with 37 cases. Taking a closer look at which model diverged from the other models in this category, Figure 4.9 presents a breakdown per label. It is of note that the *cnn* model is present for almost every relation, and in many cases it diverges the most, by far. The *cnn* model is matched only in the OTHER relation, where *medium* diverges almost the same number of times. The *small* model is relatively low, diverging the most on PARENT.

Finally, there were 99 cases were two of the models were incorrect. This includes 50 for OTHER and below 10 each for CHILD, EDUCATED, PARENT, SIBLING, BIRTHPLACE, DEATHPLACE and BIRTHDATE. Again, *cnn* diverged most with 75 cases, although *small* and *medium* were closer this time with 64 and 59, respectively. Figure 4.10 presents a breakdown of where each model diverged the most for this category, although the results are as expected.

Overall, the most errors were cause by sentences labelled OTHER, followed by relations between two persons. Other labels were more problematic here than with the fine-tuned models, such as EDUCATED and BIRTHPLACE, which is unexpected due to the large amount of training examples.

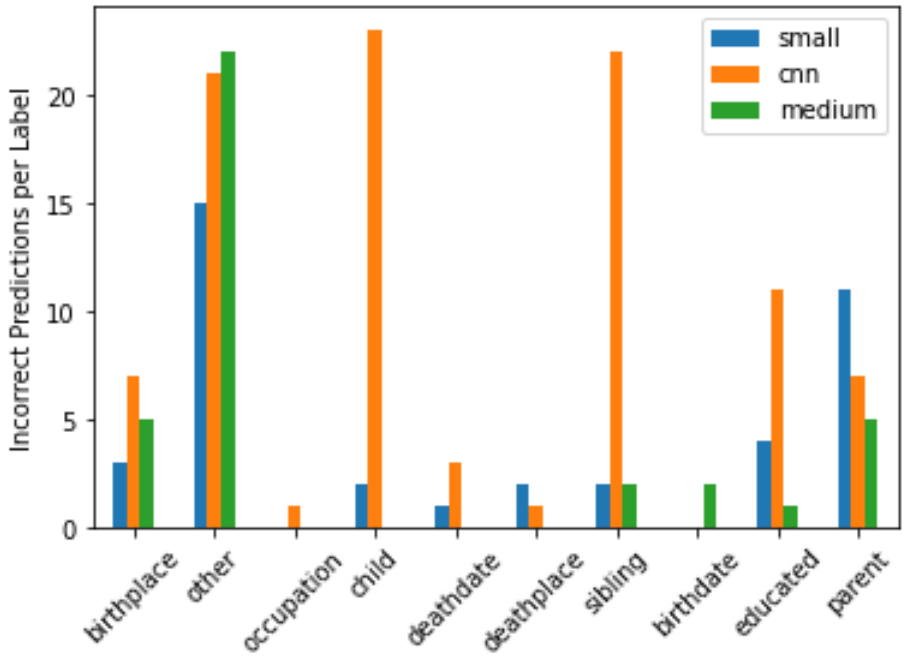

Figure 4.9: Number of times **one** model diverged from the other models (per label).

Fine-tuned versus pre-trained In order to make a final recommendation, this section will compare the best performing models across the mode of training. From the fine-tuned models, while *skip* can be ruled out, it is difficult to determine whether *normal* or *coref* is better, and in addition, although *all* performs arguably the best, it achieves this by combining the three processing methods. However, determining the best processing approach would be preferable. From the pre-trained models, it is clear that the *small* set performs best. Therefore, each fine-tuned model will be compared to the *small* model.

Confusion matrices for these models are shown in Figures 4.4 and 4.8. For each of the tested combinations the number of times one of the two models diverged was 190, 194 and 184 times with an almost even split each time. Figure 4.11 compares

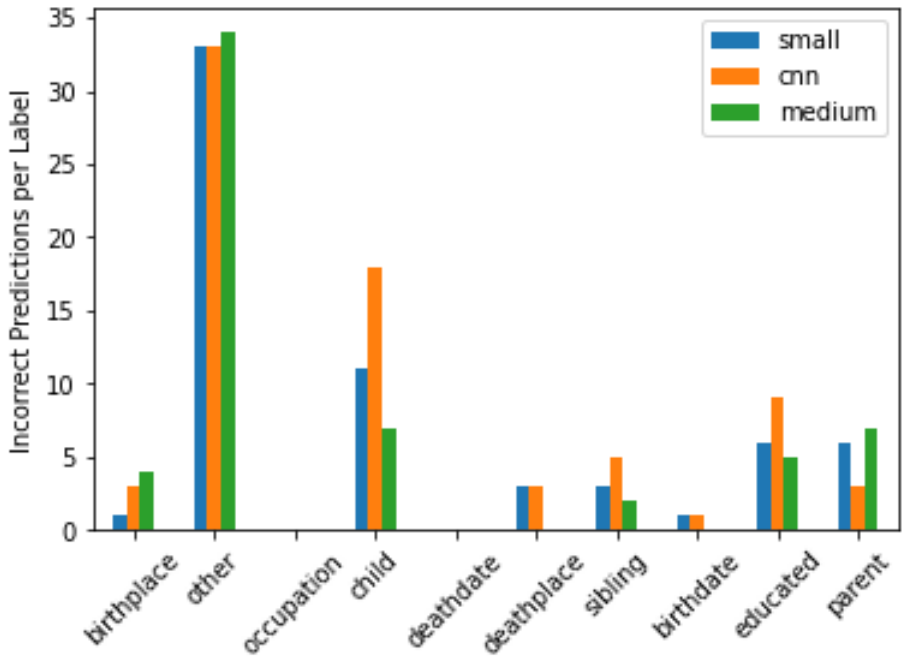

Figure 4.10: Number of times **two** models diverged from the other models (per label).

all to *small*. While *small* almost doubles the amount of times it produces incorrect labels in comparison to *all*, it appears that the models excel at different labels. The model *all* produces more errors with BIRTHPLACE, CHILD and SIBLING, *small* produces more errors with OTHER, EDUCATED and PARENT.

Figure 4.12 compares *coref* to *small*, with *coref* being arguably the best model in terms of processing approach. While the evaluation metrics indicate the best recall, at the cost of some precision, the number of times it diverges tend to agree with this. However, it is noteworthy that it produces more than triple the amount of errors for the OTHER relation when compared to *small*.

The other potentially] best model in terms of evaluation metrics is *normal*, which has the higher precision score. Figure 4.13 compares it to *small*. This comparison

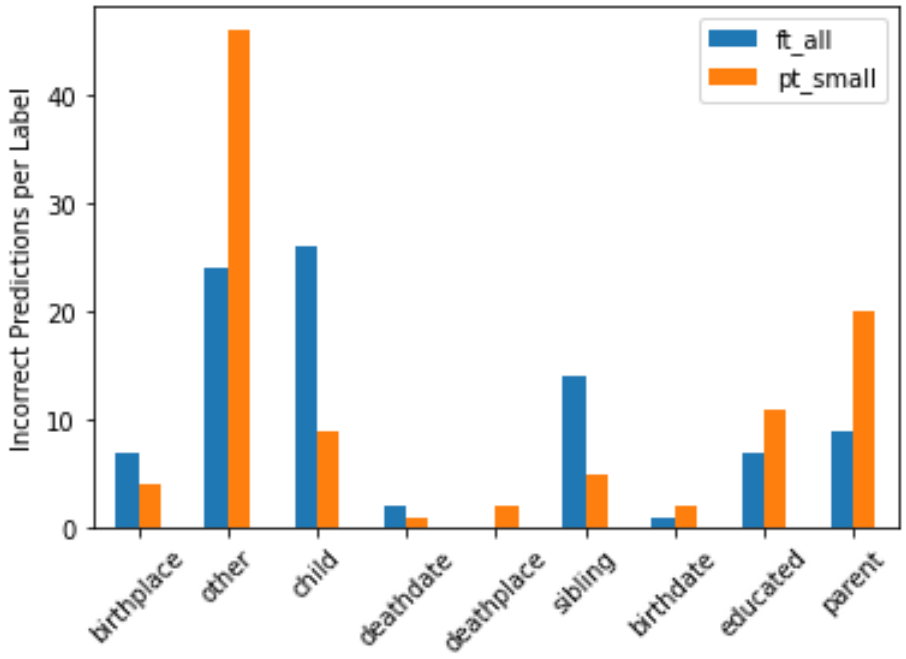

Figure 4.11: Number of times **one** model diverged from the other models (per label).

is highly useful, because it shows exactly the improvements of pre-training with a specialised set, rather than the default BERT model, as both these models were fine-tuned on the *normal* set. Overall, pre-training with the small set of Wikipedia sentences works favourably for the OTHER relation, with almost half as many divergences. The other relations are also reduced, with the exception of CHILD, which produces marginally more, and PARENT, which produces many more. It appears that there is a shift from OTHER to person based relations such as PARENT and CHILD.

Summary Overall, the evaluation shows that there are not many differences between the processing approaches, with the exception perhaps of the *skip* model. The error analysis has highlighted that the precision can vary for each label

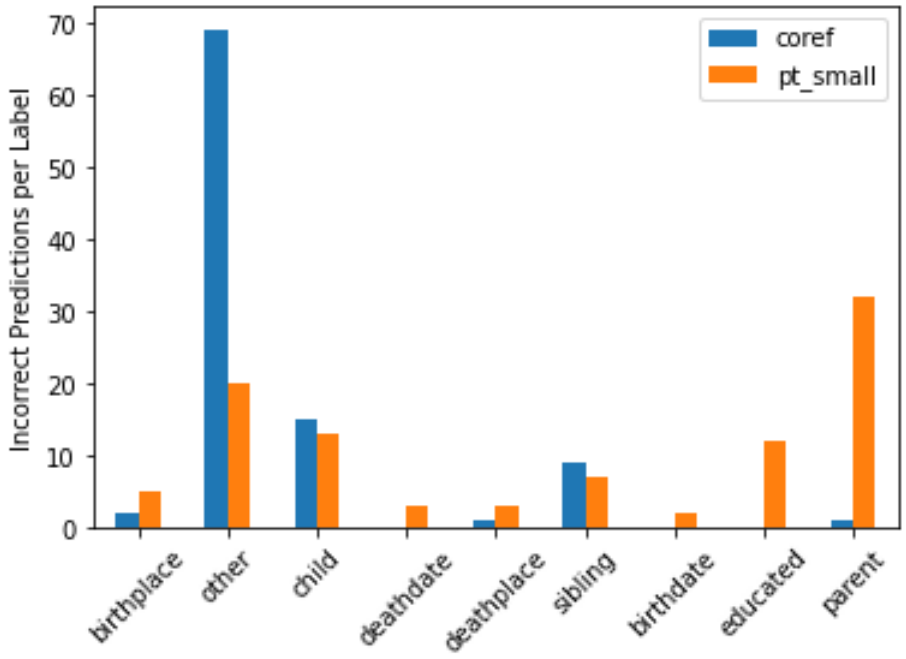

Figure 4.12: Number of times **one** model diverged from the other models (per label).

depending on the model. It would appear that pre-training the models, due to its time-consuming nature, is not preferable, and the *normal* model probably produces the best results for the least amount of effort.

A clear error pattern emerges after taking into account all the results, as well as comparing sentences where most models failed to assign a correct label. Relations that are similar in terms of sentence structure and tagged entities (i.e. two people) , but feature a word or phrase which changes the target entity. This becomes clearer when looking at the sentences in Example 4.7, which always resulted in an incorrect label. All three sentences were manually assigned the label OTHER, because the relation expressed in the sentence does not relate to the marked entity, but to a different entity.

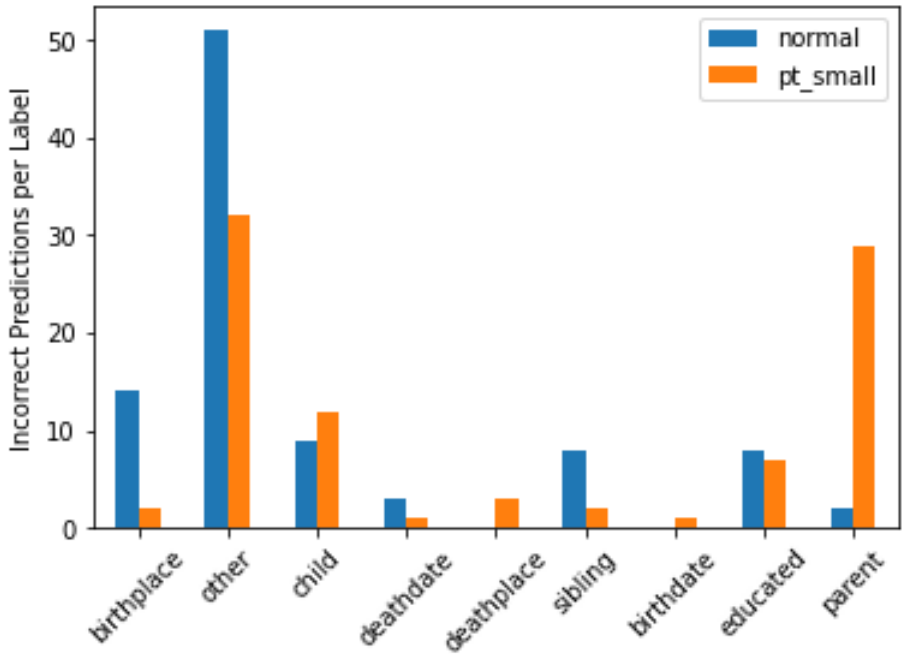

Figure 4.13: Number of times **one** model diverged from the other models (per label).

(1) **<e1>**Wyatt**</e1>** is the direct ancestor of famous **<e2>**poet**</e2>** Sir Thomas Wyatt.

(2) **<e1>**Cass Gilbert**</e1>** is sometimes also confused with his son, **<e2>**architect**</e2>** Cass Gilbert, Jr..

(3) **<e1>**Veselý**</e1>** 's younger sister is a **<e2>**basketball player**</e2>**.

Example 4.7: A selection of sentences where each model failed to provide a correct label.

Another frequent case was when two person entities were marked, and the sentence structurally looked similar to the *educated* and *occupation* relations. These labels are often structurally the same as other relations, with the only difference being a verb. Example 4.8 shows frequent examples of these cases, where one verb, e.g. *arrived* and *settled*, changes the meaning of the sentence and

does therefore not express a relevant relation.

(1) At approximately 11:35, **<e1>**Whitman**</e1>** arrived on the **<e2>**University of Texas at Austin**</e2>** campus.

(2) **<e1>**Lossky**</e1>** settled in **<e2>**Paris**</e2>** in 1924.

(3) **<e1>**Kulmbach**</e1>** probably arrived in **<e2>**Nuremberg**</e2>** around 1505.

Example 4.8: A selection of sentences where each model failed to provide a correct label.

The third case of misclassifications is caused by the similarity of certain labels. For the relations SIBLING, PARENT and CHILD, it seems that the models are confused by these three labels, because of the structural similarity of the sentences. Example 4.9 shows a selection of sentences that were always misclassified. The first three examples should be labelled SIBLING, and in each case feature a keyword that points to a different relation: *children*, *sons* and *mother*. The last two examples should be CHILD and PARENT, respectively, and show the same pattern. However, while the models are often confused between these relations, the relations are also often falsely assigned OTHER and BIRTHPLACE tags, which could be explained by the fact that these are the most common in the training sets.

(1) Their children were **<e1>**Waldo**</e1>**, Ellen, Edith, and **<e2>**Edward Waldo Emerson**</e2>**.

(2) He had two sons: **<e1>**Gnaeus Domitius Ahenobarbus**</e1>** and **<e2>**Lucius Domitius Ahenobarbus**</e2>**.

(3) **<e1>**Karl**</e1>** and **<e2>**Michael Polanyi**</e2>**'s mother was Cecília Wohl.

(4) **<e1>**Sigler**</e1>** gave birth to their son, **<e2>**Beau Kyle Dykstra**</e2>**, on August 28, 2013.

(5) **<e1>**Bakis**</e1>** was born in Kaunas to the family of **<e2>**Stasys Antanas Bakis**</e2>**, a Lithuanian diplomat.

Example 4.9: A selection of sentences where each model failed to provide a correct label.

Although the evaluation of the models indicates good performance, the error analysis raises the question of whether the problems that appear to be inherent to guided distant supervision itself can be solved. The most urgent problem involves the tag OTHER used to determine irrelevant sentences. It seems that this group contains many edge cases, that are close to existing relevant relations. If the performance of a trained model were to be improved in these cases, it seems that the OTHER tag could be split into more finer-grained categories. For instance, for each relevant relation there could be an equivalent relation that tags similar sentences that are however not relevant. This would require a rule in the GDS approach that separates these by using similar entity types (i.e. *Person/Date* or *Person/Person*). Another way to improve the approach would be to use probability scores for each predicted relation, so that secondary or further tags could be assigned.

4.4 Conclusion

This chapter presented *Biographical*, a relation extraction dataset that is an example of *guided distant supervision*. The compilation process has been described in detail, and experiments investigating the dataset have been carried out. This included different processing approaches, a manual annotation task, the training of different neural models and an error analysis. Not only have these experiments investigated different ways of optimising the compilation of the dataset for different goals, they have also validated the results in terms of machine learning. The research presented in this chapter aims to address **RQ-2**.

RQ-2 *Are semi-supervised datasets effective for training a biographical relation extraction model?*

RQ-2a *How do certain processing steps affect the model performance?*

RQ-2b *Does pretraining improve perfomance over fine-tuned models?*

With regard to the first part of the question, it can be said that models for biographical relation extraction can be trained using Wikipedia-based training data. Throughout the chapter, an automatic compilation method was used to annotate sentences taken from Wikipedia. This dataset was subsequently used to train various neural models to extract the relation between two entities.

With regard to sub-question **RQ-2a**, three different processing steps were used to compile the training data. The evaluation of the models trained on these different sets has shown that the performance can vary, although the differences

can be said to be marginal. However, with a deeper look into the results via the error analysis, it seems that including coreference resolution can improve the recall of the model. Reducing the number of sentences of the training set, as is the case with the *skip* model, does reduce performance. In addition, a real-world dataset would be needed to ascertain whether one of the alternate processing approaches (*coref* or *skip*) is better suited, as it would ideally be less similar to Wikipedia and therefore better show the general applicability.

With regard to sub-question **RQ-2b**, concerning pre-training, it is clear that this does not necessarily improve performance overall. Nevertheless, this step, while time-consuming, can adjust the predictions to be balanced differently.

Overall, this chapter has demonstrated that a dataset compiled using GDS is successful for RE tasks. The desired processing and training approach would potentially depend on the targeted texts and relations. Although rough guidance can be ascertained from this evaluation, with regard to precision and recall of a model, some testing would have to be completed depending on the application. Due to the success of this approach, the question that arises now is whether this approach can be successful in a language-independent setting.

CHAPTER 5

MULTILINGUAL BIOGRAPHICAL RELATION EXTRACTION METHODS

In recent years, distant supervision for relation extraction has been proposed to create a large amount of auto-generated labels (Mintz et al., 2009). It is based on the assumption that if there is a relation between entities, then every sentence containing them may also express that relation. The majority of the multilingual datasets released recently are based on this distant supervision paradigm (Seganti et al., 2021). Despite being popular, distant supervision has major flaws (Hoffmann et al., 2011) as it possesses an ideal hypothesis that all instances containing the same entity pairs express the same relation. However, this is far from reality because multiple relations may exist between a specific entity pair (Surdeanu et al., 2012). For example, both the relations *born in* and *died in* are valid between the entity pair *Shakespeare* and *Stratford-upon-Avon*. In addition, a relation between two entities may not be expressed in a sentence of multiple entities. Therefore, distant supervision can end up providing confusing auto-generated labels.

To overcome this, *guided distant supervision* was proposed in the previous chapter as a method which ensures correct labels for the relations identified by distant supervision, mainly by using external resources such as Pantheon (Yu et

al., 2016), and Wikidata (Vrandečić and Krötzsch, 2014). However, that research is limited to English and adapting the same approach to a different language can be challenging due to the availability of external resources. In this chapter, a novel large dataset for German biographical RE is proposed, while adapting *guided distant supervision* to overcome several challenges in the methodology for English. Furthermore, the challenge of adapting *guided distant supervision* to potential further low-resource languages is addressed by means of experimenting with cross-lingual RE. The approach will significantly reduce the burden on the manual annotation process while having more accurate sentences than traditional distant supervision. The main contributions of this chapter are as follows:

1. Introduction of the largest German dataset for biographical RE built using *guided distant supervision* with ten relationship categories. Including a manually annotated subset that can be used for evaluation.

2. Evaluation of several machine learning models to perform biographical RE for German, based on state-of-the-art transformer models such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020).

3. Evaluation of cross-lingual transfer learning between two datasets created using *guided distant supervision*.

The rest of the paper is structured as follows. Section 5.1 overviews related work. Section 5.2 describes the adaptations to *guided distant supervision* that were made to produce a German biographical dataset for RE. This includes the requirements

to the source datasets, the methodology for labelling and the neural models that were trained. Section 5.3 discusses the evaluation of the approach, including a description of the annotated evaluation data, the results for the baseline approach and the results for the neural models under various learning conditions. The evaluation is concluded by an error analysis. Finally, Section 5.4 summarises the conclusions from the research presented in this chapter.

5.1 Related Work

Biographical IE from online documents is a popular research area in the NLP community, given its wide range of applications. A number of early approaches to biographical fact extraction are presented by Garera and Yarowsky (2009), who test the effectiveness of six different approaches of the time. Using Wikipedia articles for textual data and the structured database NNDB[1] for training data, the authors first implement a pattern based approach first suggested by Hovy et al. (2002), which looks for <*Name,Attribute Value*> triples in Wikipedia. After this, five models using distinct approaches are tested, including a position-based model that matches words at specific positions, latent context models, and transitive models, as well as correlative and distributive models over different attributes. While it is concluded that the best performance is achieved by combining the position and transitive models, one of the stand-out findings is the cross-lingual use of the latent model. In arguably one of the first occurrences of multilingual targeting in biographical information extraction, the authors describe how they

[1] https://nndb.com

train the latent model on German Wikipedia text in order to detect occupations given in English.

Glazkova (2020) describes a system for classifying text fragments according to their topic, specifically in the area of biographical information. The authors use a manually annotated collection of biographical texts taken from Wikipedia, where sentences are grouped into one of ten groups, excluding, however, a non-biographical group unlike Zhou et al. (2004). Two context-aware neural models are trained, namely multilingual BERT (mBERT) and Russian BERT (RuBERT), as well as a Bag-of-Words based SVM approach. While the Transformers architecture is used for the neural models, Glazkova (2020) carries out experiments with a variety of layer configurations. The system is fed with and without context for each configuration. Overall, the system is capable of achieving an of 94% F1-score at the best performing configuration, while the SVM approach only scores around 66%, a much lower but expected value for this kind of approach. This approach demonstrates that context-aware models perform much better than context-free models, particularly true in this case for biographical data.

Recently, several neural network-based models (Katiyar and Cardie, 2016; Miwa and Bansal, 2016) have been proposed to extract entities and relations from a sentence jointly. The current state-of-the-art in RE, also used for this research, is based on neural transformers (Baldini Soares et al., 2019). These transformer models are trained using a language modelling task such as masked language modelling or next sentence prediction, which have then been used to perform RE

as a downstream NLP task. Results on recent RE datasets show that transformers outperform the previous architectures based on RNNs and CNNs (Xue et al., 2019; Baldini Soares et al., 2019).

The machine learning models referred to previously follow a supervised paradigm, where a dataset is required for training a model. Therefore, the NLP community has a growing interest in producing datasets capable of training machine learning models to perform RE. Several datasets in this area, such as NYT24 (Hoffmann et al., 2011), and TACRED (Zhang et al., 2017) have been released for this purpose. As mentioned previously, annotating data manually for RE is time-consuming and expensive, limiting manually annotated datasets largely to English.

As a solution, (Mintz et al., 2009) proposed distant supervision to automatically label RE datasets. Even though distant supervision is flawed, several multilingual datasets for biographical RE have been released under the distant supervision paradigm (Seganti et al., 2021). Various solutions such as multi-instance learning (Hoffmann et al., 2011) and sentence-level attention (Lin et al., 2016) have been proposed to overcome the problems of distant supervision. Nevertheless, these solutions are limited in their own ways, giving rise to the introduction of *guided distant supervision* to create biographical RE datasets automatically (Section 4.2). However, the chapter focused on English, with heavy reliance on external resources, making it difficult to port to different languages.

Within the application scope of transformer models, many of the tasks have been focused on English language, due to the fact that the majority of pre-trained

transformer models are trained on English data (Ranasinghe and Zampieri, 2020). In addition, with the multilingual models that do exist, such as mBERT (Devlin et al., 2019), there has been speculation about the ability of these models to accurately represent every single language (Pires et al., 2019). While multilingual models are more of an amalgamation of models across different languages, cross-lingual models are designed to transfer knowledge between languages, and apply it where the model determines is the most appropriate. Although mBERT does show some cross-lingual characteristics, it has not been trained on cross-lingual data, which is one of the main requirements of a cross-lingual model (Karthikeyan et al., 2020). In order to address this problem, Conneau et al. (2019) introduced the XLM-R model. XLM-R is a cross-lingual model trained on 104 languages, which is able to achieve state-of-the-art results on cross-lingual benchmarks (Conneau et al., 2019). Similar to other transformer architectures, it has been used for text classification tasks (Conneau et al., 2019). Most importantly, however, it also performs comparably on monolingual tasks, making it extremely versatile and therefore a natural choice for the approach at hand.

Multiple arguments indicate that the use of a cross-lingual model would yield positive results. First, a robust sentence classifier can be trained for English. This is ensured by the fact that cross-lingual models perform well in monolingual tasks, as has been pointed out. Moreover, finding data is possibly easiest for English, as it is not only well represented in NLP (if not the best represented) and one of the more straight-forward to process. With a robust classifier for English, most of the ground work for this approach is already laid, as the next step is zero-shot transfer

of the English knowledge. Hereby, the pre-trained cross-lingual model is simply trained on the English data for each classification task, which can each be applied to the same classification task in another language.

In this chapter, the gap in multilingual (biographical) IE is addressed by adapting *guided distant supervision* to create a German biographical RE dataset. This should also serve as a way of demonstrating that the methodology described can be applied to other languages. In the case of low-resource languages which may not be able to make use of guided distant supervision, cross-lingual RE models, trained on English for example, are investigated as a substitute.

5.2 Adapting Guided Distant Supervision

The approach described here follows the methodology introduced in the previous chapter (*guided distant supervision*) to create a German biographical RE dataset. As highlighted previously in this chapter, adapting data compilation methods to other languages can be difficult due to resource and labour requirements. The approach presented here shows that these limitations can be overcome using resources that are multilingual by design (Wikipedia and Wikidata) and combining these with an automatic labelling approach. The steps are split into data source selection (Section 5.2.1), and automatic labelling (Section 5.2.2).

5.2.1 Data Sources

As described in the previous chapter, the data compilation method requires a source of textual data and a structural counterpart, much like distant supervision

(Mintz et al., 2009). For the purposes of this research, and in line with Chapter 4, Wikipedia, Wikidata and Pantheon Yu et al. (2016) were used. Each data source provides a different part for the automatic compilation process.

Pantheon Pantheon serves as the main source of structured data, as well as an indicative list of entities to target, as in the previous chapter. Wikidata and Wikipedia links are provided, as well as meta information (birthdate, birthplace, occupation, etc.). As a partly manually curated list of entities that all have biographical entries across a minimum of 15 languages in Wikipedia, this resource is particularly useful for multilingual application. Working with entities that have articles across multiple languages, a certain level of comparability of the article contents is ensured. While this should by no means be regarded as parallel data, it is a step in that direction.

Wikidata In addition to Pantheon, Wikidata is used, which contains structured information from Wikipedia. As before, Wikidata is used to extend some of the information about the relevant entities, as well as retrieving certain parts of information in German, such as occupation and place of birth. Relations like these can be difficult to translate automatically, as for instance German has three genders for nouns, which can only be inferred from context information. Locations, on the other hand, may be complex since they can often be ambiguous, therefore again relying on context information to determine the correct translation. However, it should be noted that while Wikidata has entries for most of the used entities in

German, this is not always the case. It is most common for Wikidata to provide all information in English, with other languages only depending on availability.

Wikipedia Wikipedia serves as the source of textual data, and is linked directly to the indicative list of entities, taken from Pantheon. Processing the textual data requires a number of steps in itself, established previously. In summary, all relevant articles from a Wikipedia dump are extracted using article IDs. As this process involves German Wikipedia articles, the Wikipedia IDs from Pantheon (which are provided for English articles only) need to be mapped to the German article IDs by using a database dump of Wikipedia, which contains this information. The ID mapping can be done via regular expressions, searching for the English ID and looking for a nearby keyword indicating the German ID. After matching the relevant articles, the articles are extracted using the wikiextractor[2] package for Python, which converts from XML to plain text.

5.2.2 Automatic Labelling Adaptations

The automatic labelling approach used here is adapted from the approach presented in Section 4.2.2. The sentences taken from each German Wikipedia article are labelled by processing them using spaCy[3], running NER to tag entities, and then matching the entities to the available structured data. For this, the non-neural spaCy model for German is used, since the latest neural model does not support NER, which is crucial to the approach. For the previously mentioned

[2] https://github.com/attardi/wikiextractor
[3] https://spacy.io

English counterpart to this dataset, the neural spaCy model was used. Compared to the English dataset, the German pipeline is not as precise, and in addition, far fewer entity types are supported: *location* (LOC), *person* (PER), *organisation* (ORG) and *miscellaneous* (MISC).

For each article, the matching algorithm identifies for the main article entity, then checks whether other entities are part of the structured information that is available from the structured sources. If a positive match is made, the sentence is labelled according to the match. For all relations, it is vital that only the first occurrence of a match is accepted. This is due to the underlying assumption that the selected relations are of an importance such that they would appear towards the beginning of a Wikipedia article.

BIRTHDATE: Im Alter von fast 77 Jahren starb <e1>**Lorenzo Ghiberti**</e1>am <e2>**1 Dezember 1455**</e2>in Florenz.

BIRTHDATE: At the age of almost 77 <e1>**Lorenzo Ghiberti**</e1>died on <e2>**1 Dezember 1455**</e2>in Florence.

Example 5.1: Example of multiple entity replacements in German (with English translation).

EDUCATED: <e1>**Menger**</e1>lernte bei Hans Hahn und promovierte 1924 an der <e2>**Universität Wien**</e2>.

EDUCATED: <e1>**Menger**</e1>studied with Hans Hahn and received his doctorate from the <e2>**University of Vienna**</e2>in 1924.

Example 5.2: Example of nested entity replacement in German (with English translation).

As the spaCy NER model does not detect dates or titles (occupations), regular expressions are used to detect these entities. For dates, a regular expression to match date variations was used. For the titles, a list of occupations as gathered from the meta information from Pantheon was used and combined. Because the information in Pantheon is in English, the list was translated using deepL[4], although since the list is quite short (around 100 items) some manual revisions were made where MT was not accurate. In addition, both masculine and feminine versions of each occupation were added, since German has distinct forms for these. In order to match the correct location names, all places of birth and death were translated using the *alternate names* table provided by GeoNames. Although Pantheon does provide a GeoNames identifier for each location, this did not seem to match GeoNames. Instead, the names were retrieved via the alternate names table, with the use of partial coordinate matching. This is because the coordinates used to distinguish locations between the two sources were found to rarely match exactly. Therefore, to ensure the correct location approximately, the first three digits of the coordinates were used.

For matching the named entities to the structured Pantheon and Wikidata information, list matching was preferred, instead of just string matching used in the previous chapter. This was due to the fact that both the English and German version of each relation should result in a positive match, since it was found that the German counterpart of some relations (mainly the location names) was not always used.

[4] https://deepl.com

In terms of processing approaches, two versions of the German dataset were compiled (instead of three in the previous chapter). While the normal method was performed as described above, the skip method involves skipping the first sentence of a Wikipedia article. The aim of this method is to find more diverse sentences, since many times the first sentence of Wikipedia contains much of the desired information. As this sentence has a very standardised structure (see Example 5.3), it stands to reason that having many sentences of this kind could over-fit a neural model unnecessarily, making it less precise with sentence types of a different structure. A coref version of the German set was not compiled since this part of the spaCy pipeline was not available at the time of carrying out this research. In addition, Section 4.3.3 has highlighted that the addition of coreference resolution to the English approach only improves the performance in terms of recall.

German: Bernard Tomic (*21 Oktober 1992 in Stuttgart, Deutschland) ist ein australischer Tennisspieler.

English: Bernard Tomic (*21 October 1992 in Stuttgart, Germany) is an Australian tennis player.

Example 5.3: Typical first sentence of a German biographical Wikipedia article.

Table 5.1 shows the counts for each relationship type in the German dataset, which are compared with the English dataset from the previous chapter. Overall, it is clear that for the German set a much smaller number of relations was gathered, which is to be expected when taking into account the relative size of German

Relation	EN normal	DE normal	DE skip
birthdate	51,524	8,777	770
birthplace	50,226	12,833	5,816
deathdate	17,197	922	454
deathplace	18,944	4,059	3,263
educated	5,639	610	607
occupation	18,114	10,861	4,836
parent	6,352	3,704	3,565
sibling	2,083	917	890
child	2,209	718	701
other	173,969	39,782	20,469
Total	**346,257**	**83,183**	**41,380**

Table 5.1: Number of Relations in English vs. German Sets

Wikipedia in comparison to English Wikipedia. This could potentially also be due to the fact that the performance of the German model for spaCy is not as precise as the English model. Furthermore, in the case of the *skip* set, significantly fewer relations were matched. The cause for this could be the fact that the German articles are much shorter in some places, and feature more complex sentences in other places.

This goes against the theory that roughly the same number of relations as for the English dataset would be found. This assumption was based on the choice of Pantheon as an indicative list of persons to extract information about. This would ensure the same number of articles across languages, since this is one of the main aims of the Pantheon dataset (Yu et al., 2016).

5.2.3 Neural Models

Using the datasets compiled for English (see previous chapter) and German (described in this chapter), a number of different neural models were trained to

examine relation extraction performance. All the neural models used for this research are based on the neural transformers (Devlin et al., 2019), and the architecture first shown by Baldini Soares et al. (2019). Exact details of the architecture have been presented previously in Section 4.2.4, which also describes the parameters and hardware used for training.

As stated previously, a variety of pre-trained transformer models were tested: *bert-base-uncased* (Devlin et al., 2019), *bert-base-multilingual-cased* (Devlin et al., 2019), *bert-base-german-cased* (Chan et al., 2020) and *xlm-roberta-base* (Conneau et al., 2020). All models are available via the HuggingFace website[5] (Wolf et al., 2020).

5.3 Evaluation

In this section, the evaluation results of the various neural models trained using different dataset combinations are presented. The aim of this evaluation is to find out where the strengths and weaknesses of each data/model combination are, in order to verify whether the approach for data compilation is successful. The annotated data used for evaluation is described in Section 5.3.1, the baseline used for comparison is described in Section 5.3.2, with the following sections dedicated to three separate learning paradigms: monolingual, multilingual and cross-lingual. The final part of the evaluation presents an error analysis, looking into the most common incorrect classifications (Section 5.3.4).

[5] https://huggingface.co

Relation	P	R	F1	Supp.
birthdate	.98	1.0	.99	196
birthplace	.69	.83	.76	167
deathdate	.92	1.0	.96	184
deathplace	.20	1.0	.33	39
educated	.92	.99	.95	184
occupation	.90	1.0	.94	179
parent	.84	.95	.89	178
sibling	.65	.99	.78	130
child	.69	.99	.81	139
other	.94	.31	.47	604
macro avg.	.77	.91	.79	2000

Table 5.2: Evaluation of the automatic labels on the gold set versus the manually assigned labels.

5.3.1 Manual Annotation

For the evaluation data, 2000 sentences in total were separated from the total dataset, with 100 sentences per relation from both processing methods, which were then manually annotated. The manual annotation was also used to obtain an insight into the performance of the automatic compilation method. Table 5.2 shows the evaluation results of the automatic compilation method.

Two German native speakers performed the annotation. The guidelines were the same as the guidelines for the English dataset (See Section 4.2.2), where each sentence should either explicitly or implicitly convey the relation, and no prior knowledge should have influence over the decision. The Cohen's Kappa for the inter-annotator agreement is 0.92, which indicates a very high agreement between the annotators.

5.3.2 Baseline Results

To establish a baseline for this task, the idea was to use machine translation to English, so that an already trained model could classify the sentences. This approach also functions as an alternative to the adaptations to GDS for another language if good quality machine translation to English is available for the source language. For the case at hand, the baseline results were obtained by translating each sentence in the evaluation set from German to English using deepL[6]. The translated sentences were then classified using a pre-trained BERT model for English, fine-tuned with the *normal* English Biographical set (see Section 4.2 for compilation, and Section 4.3.3 for evaluation results). Table 5.3 shows the results

Relation	**P**	**R**	**F1**
birthdate	.97	.99	.98
birthplace	.79	.91	.85
deathdate	.95	.91	.93
deathplace	.30	.89	.45
educated	.96	.77	.85
occupation	.88	.87	.88
parent	.75	.92	.83
sibling	.92	.84	.88
child	.86	.67	.75
other	.73	.66	.69
macro avg.	.81	.84	.81

Table 5.3: Baseline results with Machine Translation + English model approach.

of the evaluation using the translated German sentences. Overall, the results show this approach to be quite good, although it should be pointed out that a manual

[6] https://www.deepl.com/pro-api?cta=header-pro-api

check of some translated sentences show the quality of the machine translation to be very good. This raises the question of whether some of the sentences may have been among those used to train deepL, since they are readily available from Wikipedia. While Wikipedia is by no means considered to be parallel, it is conceivable that these sentences were part of a monolingual set of data, which was later translated to be used for training.

5.3.3 Neural Model Results

For the evaluation, three different learning paradigms were tested: Monolingual learning, where the source and target languages of a model are the same, zero-shot learning, where the source language and target languages are different, and multilingual learning, where the model is trained on two or more languages. The evaluation results are followed by an error analysis, to pinpoint the relations that are commonly incorrectly labelled by the models, as well as to determine the best model for this task by comparing the best models from each learning setting.

Monolingual Learning Results The transformer models were trained on the German *guided distantly supervised* dataset and evaluated using the German evaluation data. Several models were used, including *bert-base-multilingual-cased* (Devlin et al., 2019), *bert-base-german-cased* (Chan et al., 2020) and *xlm-roberta-base* (Conneau et al., 2020). Table 5.4 shows the precision, recall and F1-score metrics for each relation, and on an overall macro level. *Xlm-roberta-base* (Conneau et al., 2020) performed better than the other two transformer models in

the majority of relationship types and overall F1-score. Furthermore, the results are compared to the best results achieved for *guided distantly supervised* English dataset. Despite having less training data in the German dataset, the results are comparable between English and German.

Monolingual	**GBERT**			**MBERT**			**XLM-R**			**BERT-EN**		
Relation	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
birthdate	.93	.96	**.95**	.93	.96	.94	.94	.96	**.95**	1.0	1.0	1.0
birthplace	.56	.88	.67	.52	.80	.62	.64	.85	**.78**	.91	.91	.91
deathdate	.92	.97	.94	.91	.96	.93	.94	.95	**.95**	.96	.97	.97
deathplace	.15	.98	.24	.15	.98	.25	.20	.86	**.30**	.31	.73	.41
educated	.91	.88	**.89**	.92	.79	.85	.93	.81	.86	.94	.83	.88
occupation	.89	1.0	.94	.89	.98	.93	.92	1.0	**.95**	.70	1.0	.82
parent	.60	.79	.66	.58	.80	.65	.70	.70	**.75**	.81	.74	.78
sibling	.64	.79	.70	.63	.87	.72	.72	.75	**.77**	.86	.56	.67
child	.61	.54	.56	.67	.54	.60	.87	.56	**.63**	.93	.52	.61
other	.75	.49	.59	.76	.49	.60	.79	.76	**.76**	.70	.66	.68
macro avg.	.69	.84	.72	.69	.83	.71	.75	.79	**.75**	.82	.80	.78

Table 5.4: Monolingual results for GBERT, mBERT and XLM-R when trained on German dataset and evaluated on German gold dataset. Includes evaluation results for best EN model for comparison purposes.

Zero-shot Learning Results As mentioned previously, the *guided distant supervision* relies on external data sources such as Pantheon and Wikidata which might not be available in low resource languages. Therefore, this section demonstrates how the transformer models trained on a *guided distant supervised* dataset perform in a different language under zero-shot setting. The transformer models were trained on the English *guided distantly supervised* dataset (Section 4.2.4) and evaluated using the German evaluation data. The models included

bert-base-multilingual-cased and *xlm-roberta-base* (Conneau et al., 2020). Table 5.5 shows the precision, recall and F1-score metrics for each relation, and on an overall macro level.

Zero-Shot	**mBERT**			**XLM-R**		
Relation	**P**	**R**	**F1**	**P**	**R**	**F1**
birthdate	.93	.93	.93	.95	.94	**.94**
birthplace	.63	.75	.68	.65	.73	**.70**
deathdate	.93	.86	**.89**	.93	.86	**.89**
deathplace	.15	.49	.23	.25	.86	**.47**
educated	.88	.93	**.91**	.93	.95	.87
occupation	.93	.82	.87	.93	.85	**.91**
parent	.61	.66	.62	.61	.85	**.71**
sibling	.66	.70	.67	.68	.75	**.73**
child	.59	.58	.58	.72	.58	**.65**
other	.70	.66	**.68**	.78	.60	.65
macro avg.	.70	.73	.70	.72	.81	**.74**

Table 5.5: XLM-R and mBERT trained on English dataset, evaluated on German gold dataset under zero-shot learning.

According to the results, *xlm-roberta-base* outperforms the *bert-base-multilingual-cased* model in overall macro and weighted F1 scores. However, the important finding is that the zero-shot results are comparable with the monolingual results in Table 5.4. This shows that the transformer models trained with datasets created under *guided distant supervision* are capable of performing cross-lingual transfer learning. Therefore, even in the low resource language scenarios where the external resources that were used to perform *guided distant supervision* are not available, zero-shot transfer learning will provide comparably good results for RE. This finding can be beneficial for a multitude of low-resource languages.

Multilingual Learning Results Since zero-shot learning is successful, the performance of using a combined dataset of both English and German was also assessed. The idea is that this could boost performance, with the model learning

Combination	**mBERT**			**XLM-R**		
Relation	**P**	**R**	**F1**	**P**	**R**	**F1**
birthdate	.97	.95	**.96**	.95	.97	**.96**
birthplace	.72	.95	**.81**	.66	.87	.80
deathdate	.93	.96	.94	.94	.96	**.96**
deathplace	.24	.83	.35	.34	.85	**.40**
educated	.90	.96	**.93**	.95	.82	.87
occupation	.78	.99	.87	.93	1.0	**.96**
parent	.63	.82	.71	.71	.72	**.77**
sibling	.64	.67	.65	.75	.77	**.78**
child	.67	.60	.63	.90	.58	**.65**
other	.84	.53	.65	.81	.78	**.77**
macro avg.	.72	.84	.75	.77	.80	**.77**

Table 5.6: XLM-R and mBERT trained on English dataset concatenated with the German dataset, evaluated on German gold dataset under multilingual learning.

from both datasets (but still only targeting one language for evaluation). Overall, an increase in both macro and weighted F1-scores is observable. Especially with regard to precision for relations where the model is relatively inaccurate, such as DEATHPLACE and SIBLING, the results show an increase in performance. In terms of models, it is clear that *xlm-roberta-base* benefits more from this data, as opposed to only minor improvements with *bert-base-multilingual-cased*. These findings show that multilingual learning is beneficial with the datasets created under *guided distant supervision*, highlighted by Table 5.6.

5.3.4 Error Analysis

As in Chapter 4, because of the mostly marginal differences in evaluation results per model, this error analysis section will examine the differences between each model. Each label will be taken into account individually per model learning setting (monolingual, zero-shot, and multilingual). The final comparison will look at the best model overall, the best zero-shot model, and the baseline using MT and English BERT (as described in Section 4.3.3).

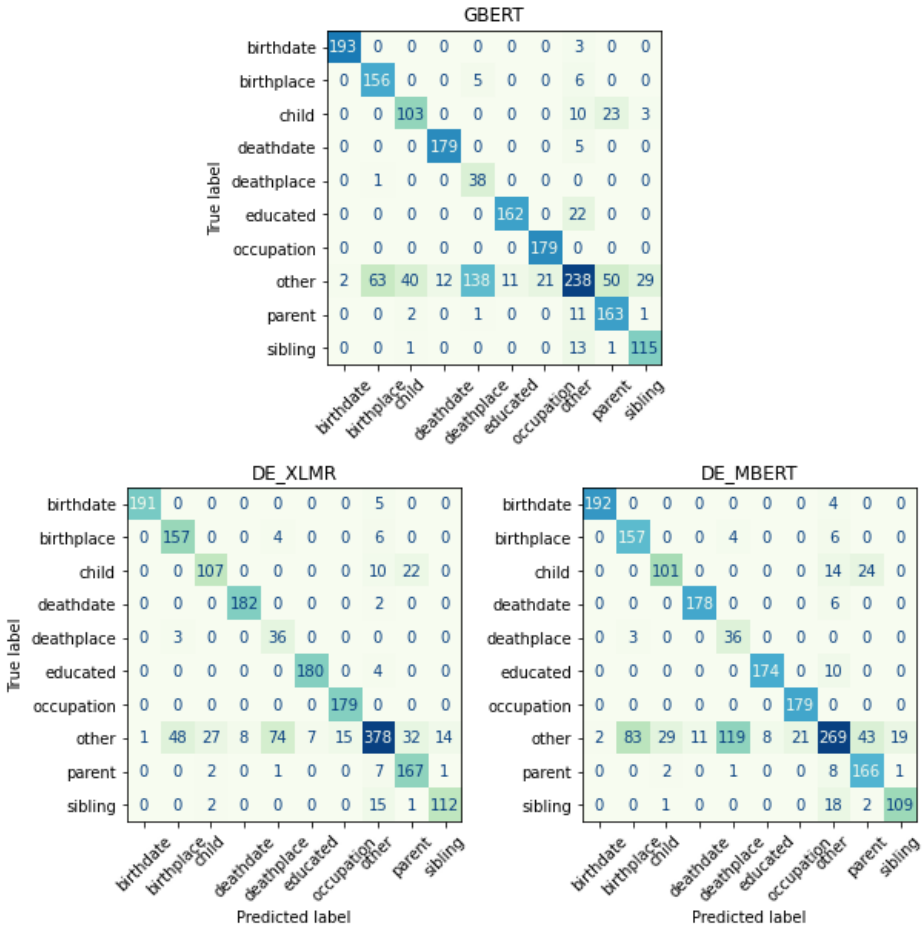

Figure 5.1: Confusion matrices for all three monolingual models.

Monolingual Learning Looking at the confusion matrices of the monolingual models (Figure 5.1) the trends are mostly similar for all three models. Particularly the OTHER label is most often misclassified, being most often assigned the DEATHPLACE and BIRTHPLACE tags. The fact that the models are unable to correctly detect this relation is probably due to its very wide scope. Both DEATHPLACE and BIRTHPLACE are often incorrectly assigned to these sentences, a major difference to the models trained with English datasets.

Other frequently misclassified tags include the CHILD relation, which is usually assigned the OTHER or PARENT tags instead. The other two relations between persons, PARENT and SIBLING, are not misclassified as often, especially in comparison to the English models. Comparing each model in terms of

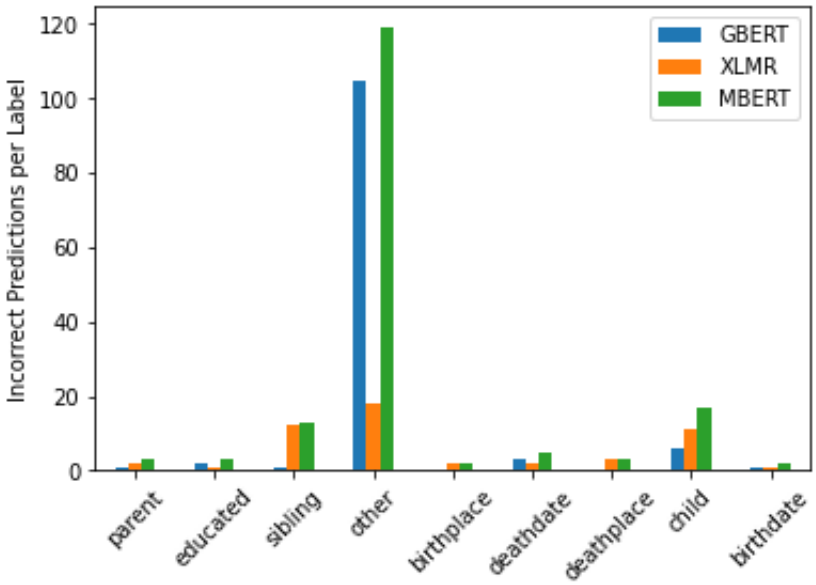

Figure 5.2: Number of times **two monolingual** models diverged from the other models (per label).

agreement on classifications, there were 549 cases out of the 2000 gold sentences where at least one model predicted an incorrect label. Out of the 549 cases, there were 253 where all three sets predicted an incorrect label. In order, 207 were with the OTHER label, 20 for CHILD, and 10 or less for PARENT, BIRTHPLACE, SIBLING, EDUCATED and BIRTHDATE.

Next were cases where two models predicted an incorrect label, of which there were 169 cases. As before, OTHER was most frequent by a large margin with 121 cases, followed by 20 for CHILD, 13 for SIBLING, and 5 or less for DEATHDATE, PARENT, DEATHPLACE, EDUCATED, BIRTHPLACE and BIRTHDATE. Figure 5.2 shows which model diverged most from the other models in this category. The chart shows that, while XLMR performs very well, all the model divergences happen in the same relations, namely OTHER, as well as CHILD and PARENT.

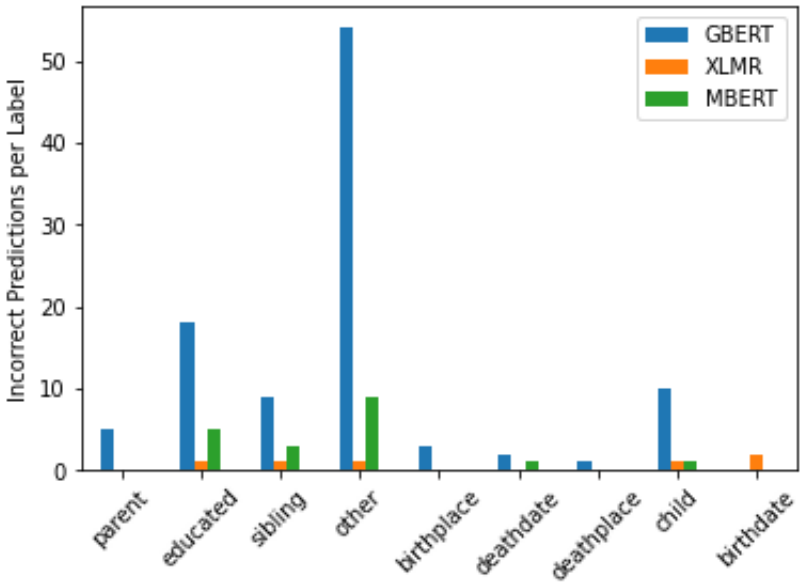

Figure 5.3: Number of times **one monolingual** model diverged from the other models (per label).

In 127 cases only one model predicted an incorrect label. Here again, OTHER was most frequent with 64, followed by 24 for EDUCATED, 13 for SIBLING, 12 for CHILD, and below 5 for PARENT, DEATHDATE, BIRTHPLACE, BIRTHDATE and DEATHPLACE. Overall, the model that made the most errors while the others did not was the *GBERT* model with 102.

Taking a closer look at which model diverged from the other models the most in this category, Figure 5.3 presents a breakdown per label. It clearly demonstrates that the *GBERT* model performs quite badly compared to the other models, especially in the OTHER relation.

Zero-Shot Learning The confusion matrices for both zero-shot learning models show very few differences (Figure 5.4). In fact, in some cases, such as BIRTHDATE and BIRTHPLACE, the models make the exact same mistakes (which is confirmed by manual comparison of the sentences in the relevant groups). The deciding relation is OTHER, which the *XLMR* model seems to predict more precisely.

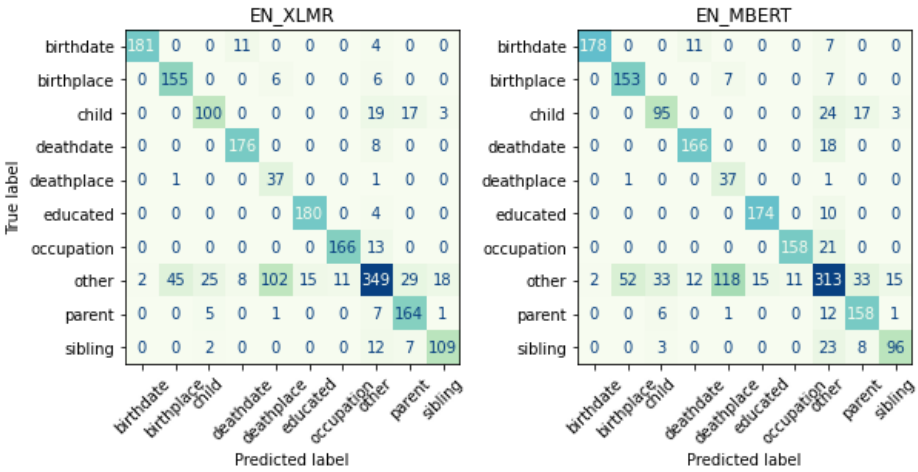

Figure 5.4: Confusion matrices for both zero-shot models.

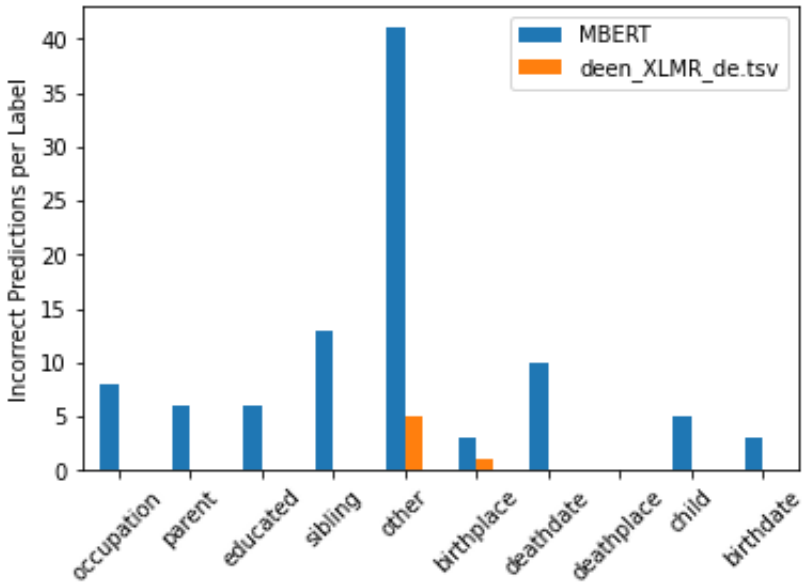

Figure 5.5: Number of times **one zero-shot** model diverged from the other models (per label).

In terms of agreement per label, in 478 cases out of 2000 the models diverged from the correct label. Of these, both models diverged in 377 cases, with 250 occurrences for OTHER, 39 for CHILD, 21 for SIBLING and 15 or less for BIRTHDATE, PARENT, OCCUPATION, BIRTHPLACE, DEATHDATE, EDUCATED and DEATHPLACE.

In 101 cases, one model disagreed, with 46 cases for OTHER, 13 for sibling and 10 or less for DEATHDATE, OCCUPATION, PARENT, EDUCATED, CHILD, BIRTHPLACE and BIRTHDATE. In 95 cases this was the *MBERT* model, with only 6 for XLMR. Figure 5.5 shows a breakdown per label. Overall, it is clear that *XLMR* performs better than *MBERT*, which is backed-up by the evaluation results described previously.

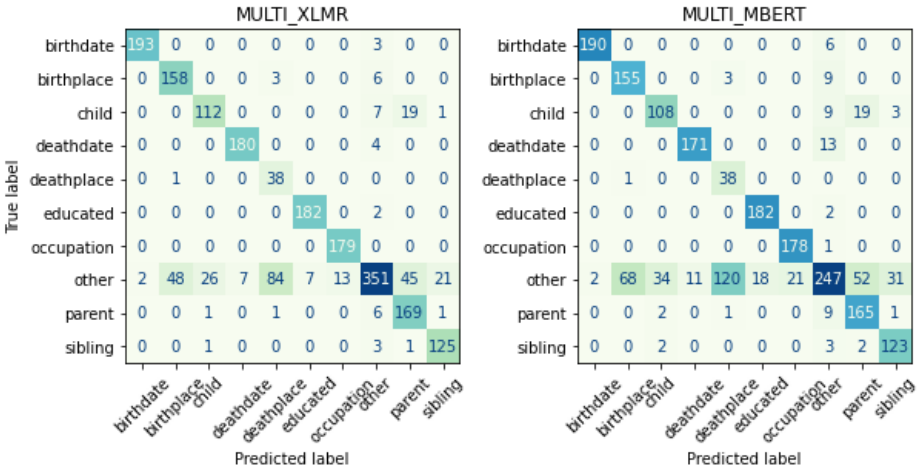

Figure 5.6: Confusion matrices for both multilingual models.

Multilingual Learning As with the zero-shot models, the multilingual models show very few differences (Figure 5.6), although direct comparison shows that the numbers are not exactly equal. As before, the most problems are caused by the OTHER relation, which also decides the final outcome of the evaluation.

In terms of agreement per label, in 447 cases out of 2000 the models diverged from the correct label. Of these, both models diverged in 309 cases, with 251 occurrences for OTHER, 27 for CHILD, and 9 or less for BIRTHPLACE, PARENT, SIBLING, DEATHDATE, BIRTHDATE, DEATHPLACE and EDUCATED.

In 138 cases, one model disagreed, with 108 cases for OTHER, and 10 or less for DEATHDATE, PARENT, CHILD, BIRTHPLACE, BIRTHDATE, EDUCATED, SIBLING and OCCUPATION. In 134 cases this was the *MBERT* model, with 4 for XLMR. Figure 5.7 shows a breakdown per label and clearly shows that *XLMR* performs better than *MBERT*. Compared to the zero-shot setting, the multilingual setting is marginally better overall, which is probably not worth the computational

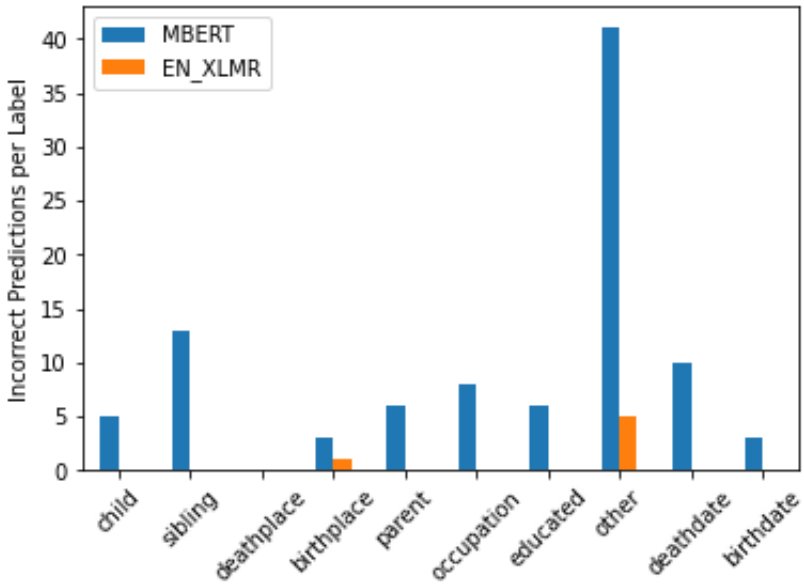

Figure 5.7: Number of times **one multilingual** model diverged from the other models (per label).

cost of combining two datasets (as opposed to only one for English under zero-shot). It is also clear that the *XLMR* architecture is much better in these settings than *MBERT*.

Learning Approaches Compared Figure 5.8 shows the confusion matrices of the *MT baseline* with BERT, *multilingual XLMR* and *zero-shot XLMR* models. While the two *XLMR* models perform comparably, with the multilingually trained version performing slightly better (fewer incorrect predictions for BIRTHDATE and OTHER), the MT BASELINE actually outperforms MULTILINGUAL XLMR for the OTHER label. It makes more mistakes, however, in other relations, such as CHILD, DEATHDATE, EDUCATED and OCCUPATION, mainly assigning the OTHER tag in these cases. Out of the 2000 gold sentences, at least one model diverged 1015

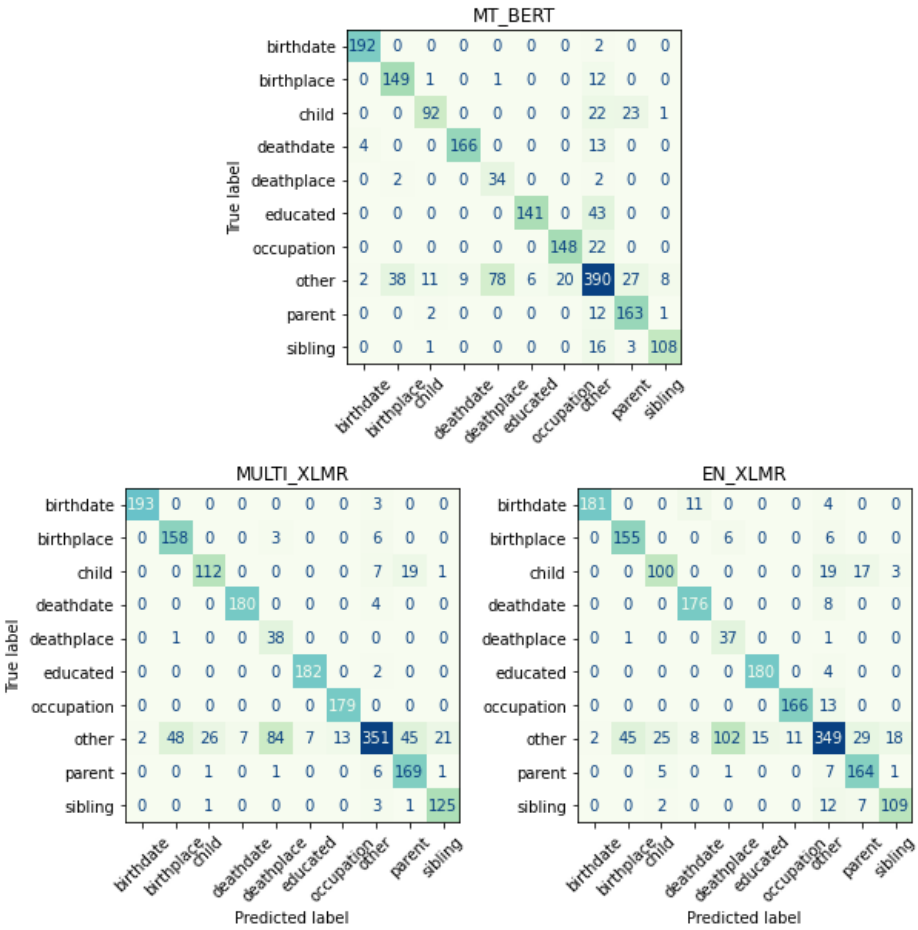

Figure 5.8: Confusion matrices for three best models (from each category).

times. In 184 cases, all three models were incorrect, with 130 of these cases in the OTHER relation, 16 for the CHILD relation and 10 or less for EDUCATED, BIRTHPLACE, PARENT, SIBLING, OCCUPATION and DEATHPLACE.

In 313 cases, one model diverged, with 90 of these being for OTHER, 39 each for CHILD and EDUCATED, 31 for OCCUPATION, 29 for SIBLING, 22 for DEATHDATE, 20 for PARENT, 16 each for BIRTHPLACE and BIRTHDATE and 11 for DEATHPLACE. Figure 5.9 shows the breakdown per label for each model. The *MT baseline* model always diverges, and is similar with the OTHER relation.

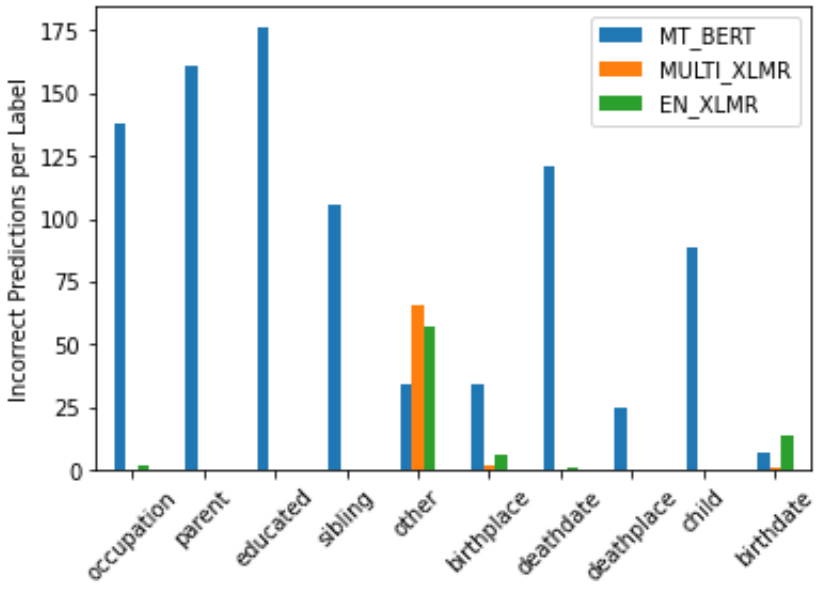

Figure 5.9: Number of times **one** model diverged from the other models (per label).

In 518 cases, two models diverged, with 212 cases for OTHER, 51 for SIBLING, 47 for CHILD, 43 for BIRTHPLACE, 41 for EDUCATED, 32 for BIRTHDATE, 31 for OCCUPATION, 24 for PARENT, 22 for DEATHDATE and 15 for DEATHPLACE. Figure 5.10 shows the break down per label for each model, which appears much more uniform for this category.

Summary The sentences labelled in the gold set with OTHER caused the most problems by far. As with the English dataset in Chapter 4, this is unsurprising when taking into account the high precision and quite low recall metrics reported from the gold set evaluation (See Table 5.2). It is also important to note the problems caused by labels that require two names, such as PARENT, SIBLING and CHILD, which overall accounted for the second largest number of mistakes.

Taking the results of the error analysis into account and comparing sentences

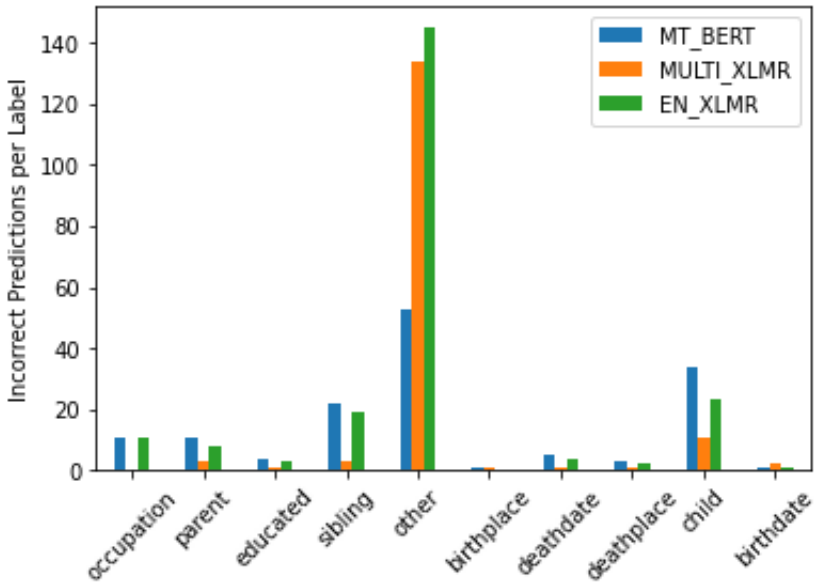

Figure 5.10: Number of times **two** models diverged from the other models (per label).

where all or most models failed to assign a correct label, a clearer pattern emerges as to which sentence structures cause the most problems.

(1) After Jeanne's death at the stake in Rouen in 1431, **<e1>**Gilles**</e1>** retired to his estates near **<e2>**Nantes**</e2>**.

(2) **<e1>**Ira Levin**</e1>** was born the son of a toy merchant and grew up in the Bronx and **<e2>**Manhattan**</e2>**.

(3) In 1868, **<e1>**Halévy**</e1>** married Louise Breguet in **<e2>**Paris**</e2>** and had two sons with her.

Example 5.4: A selection of sentences where each model failed to provide a correct label (translations of the German original).

The sentences in Example 5.4 should all have been assigned the OTHER label, contrary to the sentences presented in the same section of the previous chapter (Section 4.3.4). Here, it seems that irrelevant sentences (i.e. with the OTHER)

label that have a person and a location as target entities cause the most problems. This is confirmed by some of the numbers of the error analysis and performance evaluation, which point to the fact that the BIRTHPLACE and DEATHPLACE relations, that both have the same structure, do not perform well and have a smaller number of samples in the training set.

Furthermore, cases where either BIRTHPLACE or DEATHPLACE should have been assigned and were not are also very frequent. Example 5.5 shows some examples from this category. The third type of cases, parallel to the third most frequent type of cases of the previous chapter, are those relating to two persons, PARENT, SIBLING and CHILD.

(1) Paul David Hewson KBE (10 May 1960 in Dublin **<e1>**Ireland**</e1>**) known as **<e2>**Bono**</e2>** is an Irish musician

(2) Raised in Newport **<e1>**Rhode Island**</e1>**, **<e2>**Stuart**</e2>** first took lessons from the Scottish painter Cosmo Alexander (1724-1772).

Example 5.5: A selection of sentences where each model failed to provide a correct label (translations of the German original).

Overall, it seems there is quite a strong case for using machine translation with a model trained using English data. The baseline outperforms every other model in this chapter, particularly in terms of precision. Of course, this can be explained by the extremely poor performance of the DEATHPLACE relation with all models trained on German data, since it is very infrequent in the training set. Looking at the error analysis, however, it becomes clearer that a recommendation for the MT

baseline system should not be made in every case. This is due to its performance being extremely good with the OTHER relation, while it suffers somewhat with other relations. Should the focus be more precision with the other relations, rather than determining relations that are not relevant, all three XLMR models (monolingual German, zero-shot English and multilingual German + English) offer better results in the relevant labels, while losing with the OTHER relation.

(1) **<e1>**Richard Hamilton**</e1>** (**<e2>**Basketball player**</e2>**)

(2) **<e1>**15 April 1764**</e1>** in Versailles) briefly **<e2>**Madame de Pompadour**</e2>** was a mistress of the French king Louis XV

Example 5.6: A selection of sentence fragments that caused an incorrect prediction by the MT baseline model.

In addition, it should be mentioned that MT models for German to English are quite good, which may not be the case for other source languages. The quality of the sentences to be classified may be important, although this could not be fully tested here because the quality of the translated sentences was good. However, as shown by the example sentences in Example 5.6, the English model does struggle with sentence fragments caused by the translation, indicating that lower quality may confuse the model.

5.4 Conclusion

This chapter demonstrated a methodology to adapt guided distant supervision to compile a German dataset for biographical relation extraction. The steps

taken to adapt the already existing methodology for a new language have been described. Additionally, a thorough investigation has been carried out to assess the performance of such a dataset for relation extraction models. An evaluation was presented, comparing the performance to a baseline using MT and an English model, zero-shot learning and the combination of multiple language datasets. Not only has this demonstrated that *guided distant supervision* is a valid approach for compiling data in multiple languages, but also that in low-resource settings, zero-shot learning can perform just as well. In addition, it has further highlighted the usefulness of a well-engineered English set, as it can be used in settings where high-quality MT is available for the source language, making the adaptations of this chapter unnecessary. The research presented in this chapter addresses **RQ-3**:

RQ-3 *Is it possible to adapt guided distant supervision to another language?*

- **RQ-3a** *Can the data compilation approach be adapted to German?*
- **RQ-3b** *Do cross-lingual and monolingual models perform comparably?*
- **RQ-3c** *Is machine translation a more effective alternative?*

With regard to **RQ-3**, the chapter has described the necessary steps to adapt the approach established previously (See Section 4.2) to another language. The source language for this was German, therefore also addressing **RQ-3a**. Although some difficulties were shown in the area of finding translations for certain relations, as well as having to work with a limited processing pipeline, the evaluation has confirmed that a semi-supervised relation extraction approach can be adapted to another language, and that this is the case in particular with German.

With regard to **RQ-3b** and cross-lingual models, the evaluation of various learning approaches included two zero-shot settings, with multilingual BERT and XLM-R networks. The evaluation has shown that these models do perform comparably to monolingual models. However, regarding **RQ-3c**, it appears that machine translation is also a viable alternative, as shown in the evaluation. The only problem is that the machine translation pipeline used here seemed to perform extremely well on Wikipedia source texts, which may not be true if machine translation quality suffers on text from other domains.

In the future, some problems arising from the conversion of the methodology from English to different low resource languages could be addressed. In addition, it may be productive to explore how the multilingual models perform in a low-resource environment for RE. Furthermore, for the MT approach, research into how the quality of the translations affects the performance of an English model should be investigated. If MT works at lower qualities of translation, it may be prudent to improve GDS for English even further, since adaptations for other languages would not strictly be necessary.

CHAPTER 6

CONCLUSIONS

In conclusion, this work marks a significant step in applying data processing in the area of digital humanities. The application of more computational resources to mine data can be of immense value to historians and those working in related fields. The value has been proven not only by the results discussed in this thesis, but also by the acceptance of two papers stemming from this work at peer-reviewed conferences.

This thesis has focused on using Wikipedia as a source of data for information extraction purposes, as well as developing a methodology to automatically compile training data for biographical relation extraction models. This has been achieved by using Wikipedia as a target database to extract information from, and comparing the extracted data to Wikipedia's structured counterpart, Wikidata. Using both of these databases to extract text and label it accordingly, by using a parallel structured database to corroborate information, a new methodology for automatic compilation of data was engineered. The methodology, called *guided distant supervision*, was also adapted to work for another language, to demonstrate that it can be language independent. An evaluation of various monolingual, multilingual and zero-shot neural models, trained on the compiled data, has

highlighted how well the approach can work in certain settings, even if the quality of modern machine translation has made it a viable alternative.

The following sections revisit the research questions (Section 6.1) posed at the beginning of this thesis and summarise the contributions made through the research presented in this thesis (Section 6.2). The final section offers some directions for future work (Section 6.3).

6.1 Research Questions Revisited

The research conducted throughout this thesis was aimed at answering various research questions:

RQ-1 *Is Wikipedia a suitable data source to facilitate the extraction of biographical information?*

RQ-1a *Which processing steps are required to use Wikipedia?*

RQ-1b *Could Wikipedia and Wikidata be used in an automatic annotation approach for training data?*

Chapter 3 presented work that allowed for a preliminary answer to **RQ-1**: In the context of a limited project, Wikipedia is useable as a source of data from which to extract biographical information. Section 3.2 presented a processing approach for Wikipedia and Wikidata to extract limited biographical information for a biographical dictionary. The chapter also demonstrated the processing steps required, in answer to **RQ-1a**. To answer **RQ-1b**, Section 3.3 also presented a

pilot project to train a sentence classifier on automatically labelled data, and in doing so also contributed to **RQ-1**, again highlighting that Wikipedia can be used to facilitate the extraction of biographical information. Moreover, both Chapters 4 and 5 showed in further detail how Wikipedia can be used to obtain automatically annotated data for training neural models to extract biographical relations, across two languages.

RQ-2 *Are semi-supervised datasets effective for training a biographical relation extraction model?*

RQ-2a *How do certain processing steps affect the model performance?*

RQ-2b *Does pretraining improve perfomance over fine-tuned models?*

The pilot project described in the second half of Chapter 3, as well as the entirety of Chapter 4 have contributed to answering **RQ-2**, showing in-depth how effective a semi-supervised dataset, compiled using *guided distant supervision*, can be for training a biographical relation extraction model. Additionally, Chapter 5 has even proven that it can be effective in a multilingual setting. Chapter 4 answered **RQ-2a** by demonstrating how certain processing steps affect the model performance. The processing experiments highlighted that coreference resolution may not be worth the additional computational cost over more minimal processing. The experiments also showed that skipping the first (most informative) sentence of a Wikipedia article leads to a much smaller training set and in turn lower performance of models trained on this set. Furthermore, in relation to **RQ-2b** four models were pre-trained using data compiled from Wikipedia and CNN, and showed that

pretraining is almost certainly not worth the computational cost, since training at this scale took a large amount of time. It is also interesting to note in this context that the zero-shot learning experiments from Chapter 5 have shown that this relation extraction task can even be learned from another language, maybe indicating that its relative complexity does not require pre-training.

RQ-3 *Is it possible to adapt guided distant supervision to another language?*

RQ-3a *Can the data compilation approach be adapted to German?*

RQ-3b *Do cross-lingual and monolingual models perform comparably?*

RQ-3c *Is machine translation a more effective alternative?*

Chapter 5 investigated the use of GDS in a multilingual setting, in order to answer **RQ-3**. To do so, GDS was adapted to German successfully, confirming that the relation extraction approach can be adapted to another language, and that this is possible for German (**RQ-3a**). Going beyond this, a variety of models were trained to answer **RQ-3b** and demonstrated that cross-lingual models do compare well with monolingual models. Finally, for **RQ-3c** the German gold data was translated to English, and then classified using a previously trained English model. This clearly showed that machine translation can be a more effective alternative in certain settings, depending on the desired outcome of the task, and probably also on the machine translation quality.

6.2 Contributions

The research presented in this thesis includes a number of contributions. First, a detailed approach to process Wikipedia and Wikidata for information extraction purposes. This description is useful not only for biographical information extraction in the area of digital humanities, but can also serve as a general guideline for IE using these data sources. Since Wikipedia offers more than just biographical information, and as its potential uses go beyond IE, the processing guidelines can be useful for many approaches in NLP. Any kind of NLP application that can make use of Wikipedia sentences that have been processed in the way described can be targeted.

Next, *guided distant supervision*, a methodology to automatically compile annotated datasets for information extraction models, can serve as an alternative to distant supervision, if more control over the target relations is needed. This is conceivably true not only for biographical information, but also other areas where structural characteristics of Wikipedia text can be exploited in a similar manner. Areas of research that require IE, particularly in the area of digital humanities, can make use of this automated approach, rather than using costly human annotators.

Lastly, should this approach be desired for languages other than English, an original method for adapting GDS has been described, highlighting the prerequisites and necessary steps for this approach. An alternative using machine translation was also presented, should the adaptation not be possible for the desired source language, although this does involve English as an intermediate

language. This extends the potential areas of application of GDS to many languages, again with a focus on areas of research that may not have the resources to provide human annotated data.

In addition to the methodological contributions, there are a number of contributions in the form of resources: The first dataset compiled using GDS for English including nine biographical relations. Furthermore, the first dataset compiled using GDS for German, including the same nine relations. These datasets will be made available via a Github repository[1]. This not only enables researchers to train models for extracting biographical relations, but also opens the door to further research in order to potentially improve or expand the datasets with other methods. Finally, the trained models, including monolingual, multilingual and cross-lingual models, can also be made available, allowing researchers to extract information without having to train any models.

6.3 Directions for Future Work

In the future, further research would be needed to make GDS, as it has been proposed here, a common approach within the field of DH. GDS should be improved by addressing some of the inherent problems discussed in Chapter 4. As pointed out, the most urgent problem involves the OTHER relation, used to determine irrelevant sentences, which contains many edge cases. Were the performance of a trained model to be improved in these cases, this relation may have to be split into more finer-grained categories, requiring a rule in

[1] https://github.com/plumaj/biographical

the GDS approach that separates the relation by using similar entity types (i.e. *Person/Date* or *Person/Person*). For instance, each relation could have an irrelevant counterpart, where every similar sentence (e.g. sentence structure, entity pairs, etc.) could be captured.

Having more relations in general could also solve this problem, and would probably be preferable for many projects in any case. For biographical information, it could be useful to have relations that show where a person spent a lot of time, or where the place of work was. These relations could help to identify difficult relations better, such as DEATHPLACE in the research at hand. Another way to improve the approach would be to use probability scores for each predicted relation, so that secondary or further tags could be assigned. This could allow for decisions to be made after all relations about one person were extracted and clustered, where some labels could be ruled out.

Furthermore, the language independence for multilingual use should be tested further. This would include adapting GDS to more languages, most importantly including non-European languages, as these show a lot of variety. In tandem with this, the impact of machine translation quality on a well-trained English model needs to be tested further. For extremely low-resource languages, it would be preferable to have the option of either adapting GDS to the source language, or choosing machine translation, as it stands to reason that in these cases one of the two options may not exist. Moreover, if both options should not be available, it may be more straightforward to train a machine translation model, since creating a new database of text and having a structured derivative in parallel may be

much more time-consuming to construct. Therefore, knowing the threshold of acceptable machine translation performance is essential, so that only the minimum level has to be achieved.

Finally, once mentioned areas for improvement have been addressed, a transversal pilot project, involving experts from digital humanities and NLP, could be carried out to assess the usability of GDS in a real-world setting. This would show how useful this approach can truly be, by looking at in-depth feedback from real users, and potentially tweaking certain aspects. Nevertheless, the research described in this thesis indicates that the potential uses for *guided distant supervision* are manifold and could enable research in a variety of areas as a result.

REFERENCES

Agerri, Rodrigo and German Rigau (2016). "Robust Multilingual Named Entity Recognition with Shallow Semi-Supervised Features". In: *Artificial Intelligence* 238, pp. 63–82. ISSN: 00043702. DOI: 10.1016/j.artint.2016.05.003. URL: http://dx.doi.org/10.1016/j.artint.2016.05.003.

Alt, Christoph, Marc Hübner, and Leonhard Hennig (2019). "Improving Relation Extraction by Pre-trained Language Representations". In: *Automated Knowledge Base Construction (AKBC)*. URL: https://openreview.net/forum?id=BJgrxbqp67.

Appelt, Douglas E, Jerry R Hobbs, John Bear, David Israel, and Mabry Tyson (1993). "FASTUS: A Finite-State Processor for Information Extraction from Real-World Text". In: *IJCAI*. Vol. 93, pp. 1172–1178.

Aprosio, Alessio Palmero and Sara Tonelli (2015). "Recognizing Biographical Sections in Wikipedia". In: *Empirical Methods in Natural Language Processing, EMNLP 2015*. The Association for Computational Linguistics, pp. 811–816.

Attardi, Giusepppe (2015). *WikiExtractor*. https://github.com/attardi/wikiextractor.

Ayers, Phoebe, Charles Matthews, and Ben Yates (2008). *How Wikipedia Works: And How You Can Be a Part of It*. No Starch Press. 536 pp. ISBN: 978-1-59327-176-3. URL: http://ia700603.us.archive.org/31/items/HowWikipediaWorks/HowWikipediaWorks.pdf.

Azzam, Saliha, Kevin Humphreys, Robert Gaizauskas, and Yorick Wilks (1999). "Using a Language Independent Domain Model for Multilingual Information Extraction". In: *Applied Artificial Intelligence* 13.7, pp. 705–724. ISSN: 10876545. DOI: 10.1080/088395199117252.

Balasuriya, Dominic, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran (2009). *Named Entity Recognition in Wikipedia*, pp. 10–18. URL: http://delivery.acm.org/10.1145/1700000/1699767/p10-balasuriya.pdf?ip=134.220.228.

206&id=1699767&acc=OPEN&key=BF07A2EE685417C5.300EF34F9F006C06.4D4702B0C3E38B35.6D218144511F3437&__acm__=1548759982_7bf226b8427f0a48716a90b033c86fed.

Baldini Soares, Livio, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski (July 2019). "Matching the Blanks: Distributional Similarity for Relation Learning". In: *Proceedings of ACL 2019*. Florence, Italy: ACL, pp. 2895–2905. URL: https://aclanthology.org/P19-1279.

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni (2007). "Open Information Extraction from the Web". In: *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2670–2676. ISSN: 10450823.

Basaldella, Marco, Fangyu Liu, Ehsan Shareghi, and Nigel Collier (2020). "COMETA: A corpus for medical entity linking in the social media". In: *arXiv preprint arXiv:2010.03295*.

Baumann, Antonia (2019). "Multilingual Language Models for Named Entity Recognition in German and English". In: *Proceedings of the Student Research Workshop Associated with RANLP 2019*. Varna, Bulgaria: INCOMA Ltd., pp. 21–27. DOI: 10.26615/issn.2603-2821.2019_004. URL: https://aclanthology.org/R19-2004.

Biadsy, Fadi, Julia Hirschberg, and Elena Filatova (2008). "An Unsupervised Approach to Biography Production Using Wikipedia". In: *Proceedings of ACL-08: HLT*, pp. 807–815.

Bonch-Osmolovskaya, Anastasia and Matvey Kolbasov (2015). "Tolstoy Digital: Mining Biographical Data in Literary Heritage Editions." In: *BD*, pp. 48–52.

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko (2013). "Translating embeddings for modeling multi-relational data". In: *Advances in neural information processing systems* 26.

Brin, Sergey (1998). "Extracting Patterns and Relations from the World Wide Web". In: *International Workshop on the World Wide Web and Databases*. Springer, pp. 172–183.

Bunescu, Razvan C and Raymond J Mooney (2005). "A Shortest Path Dependency Kernel for Relation Extraction". In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*. HLT '05. Stroudsburg, PA, USA:

Association for Computational Linguistics, pp. 724–731. DOI: 10.3115/1220575.1220666. URL: https://doi.org/10.3115/1220575.1220666http://portal.acm.org/citation.cfm?doid=1220575.1220666.

Chan, Branden, Stefan Schweter, and Timo Möller (2020). "Germans next Language Model". In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6788–6796.

Cheng, Xiao and Dan Roth (2013). "Relational Inference for Wikification". In: *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1787–1796. ISBN: 978-1-937284-97-8.

Chisholm, Andrew, Will Radford, and Ben Hachey (2017). "Learning to Generate One-Sentence Biographies from Wikidata". In: *CoRR* abs/1702.0. URL: http://arxiv.org/abs/1702.06235.

Coates-Stephens, Sam (1991). "Automatic Acquisition of Proper Noun Meanings". In: *International Symposium on Methodologies for Intelligent Systems*. Springer, pp. 306–315.

Collins, Michael and Yoram Singer (1999). "Unsupervised Models for Named Entity Classification". In: *Proceedings of EMNLP/VLC-99*, pp. 100–110.

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli (2020). "Unsupervised Cross-Lingual Representation Learning for Speech Recognition". arXiv: 2006.13979.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Unsupervised Cross-lingual Representation Learning at Scale". In: *CoRR* abs/1911.0. URL: http://arxiv.org/abs/1911.02116.

Conneau, Alexis, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov (2018). "XNLI: Evaluating Cross-Lingual Sentence Representations". arXiv: 1809.05053.

Crichton, Gamal, Sampo Pyysalo, Billy Chiu, and Anna Korhonen (2017). "A neural network multi-task learning approach to biomedical named entity recognition". In: *BMC bioinformatics* 18.1, pp. 1–14.

Cui, Lei, Furu Wei, and Ming Zhou (2018). "Neural Open Information Extraction". In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 2, pp. 407–413. DOI: 10.18653/v1/p18-2065.

Dang, Hoa Trang, Karolina Owczarzak, et al. (2008). "Overview of the TAC 2008 update summarization task." In: *TAC*.

Dang, Hoa Trang, Lucy Vanderwende, Catherine Blake, Julia Kampov, Andreas Orphanides, David West, Cory Lown, Massih Amini, Nicolas Usunier, Fabrizio Gotti, et al. (2007). "Document understanding conference duc 2007". In: *Document Understanding Conference*.

De Araujo, Denis, Sandro Rigo, Carolina Muller, and Rove Chishman (2013). "Exploring the Inference Role in Automatic Information Extraction from Texts". In: *Proceedings of the Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction*, pp. 33–40.

Del Corro, Luciano and Rainer Gemulla (2013). "ClausIE: Clause-based Open Information Extraction". In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13. New York, NY, USA: ACM, pp. 355–366. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488420. URL: http://doi.acm.org/10.1145/2488388.2488420.

Derczynski, Leon, Kalina Bontcheva, and Ian Roberts (2016). "Broad twitter corpus: A diverse named entity recognition resource". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1169–1179.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL 2019: HLT*. Minneapolis, Minnesota: ACL, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423.

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel (2004). "The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation." In: *4th International Conference on Language Resources and Evaluation, LREC-2004*. Vol. 2, pp. 837–840. ISBN: 2-9517408-1-6. DOI: doi=10.1.1.78.8442. pmid: 24815381. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

Embley, David W., Stephen W. Liddle, Deryle W. Lonsdale, and Yuri Tijerino (2011). "Multilingual Ontologies for Cross-Language Information Extraction and Semantic Search". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6998 LNCS, pp. 147–160. ISSN: 03029743. DOI: 10.1007/978-3-642-24606-7_12.

Erjavec, Toma, Joh Dokler, and Petra Vide Ogrin (2018). "Slovenian Biography". In: *CEUR Workshop Proceedings* 2119, pp. 16–21. ISSN: 16130073.

Etzioni, Oren, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates (2005). "Unsupervised Named-Entity Extraction from the Web: An Experimental Study". In: *Artificial Intelligence* 165.1, pp. 91–134. ISSN: 0004-3702. DOI: 10.1016/j.artint.2005.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0004370205000366.

Etzioni, Oren, Anthony Fader, Janara Christensen, Stephen Soderland, et al. (2011). "Open information extraction: The second generation". In: *Twenty-Second International Joint Conference on Artificial Intelligence*.

Falke, Tobias, Gabriel Stanovsky, Iryna Gurevych, and Ido Dagan (2016). "Porting an Open Information Extraction System from English to German". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, pp. 892–898. ISSN: 1476-1122. DOI: http://dx.doi.org/10.1038/nmat1849. pmid: 17330084.

Faruqui, Manaal and Shankar Kumar (2015). "Multilingual Open Relation Extraction Using Cross-Lingual Projection". In: *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 1351–1356. DOI: 10.3115/v1/n15-1151.

Finkel, Jenny Rose, Trond Grenager, and Christopher D Manning (2005). "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pp. 363–370.

Florian, Radu, Abe Ittycheriah, Hongyan Jing, and Tong Zhang (2003). "Named Entity Recognition through Classifier Combination". In: *Proceedings of CoNLL-2003*. Ed. by Walter Daelemans and Miles Osborne. Edmonton, Canada, pp. 168–171.

Fokkens, Antske, Serge Ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber (2014). "BiographyNet: Methodological Issues When NLP Supports Historical Research". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pp. 3728–3735.

Freitag, Dayne (1998). "Information extraction from HTML: Application of a general machine learning approach". In: *Proceedings of AAAI/IAAI*.

Gamallo, Pablo and Marcos Garcia (2015). "Multilingual Open Information Extraction". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9273, pp. 711–722. ISSN: 16113349. DOI: 10.1007/978-3-319-23485-4_72.

Garera, Nikesh and David Yarowsky (2009). "Structural, Transitive and Latent Models for Biographic Fact Extraction". In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pp. 300–308.

Glazkova, Anna Valer'evna (2020). "Topical Classification of Text Fragments Accounting for Their Nearest Context". In: *Automation and Remote Control* 81.12, pp. 2262–2276.

Gotti, Fabrizio and Philippe Langlais (2017). "From French Wikipedia to Erudit: A test case for cross-domain open information extraction". In: *Computational Intelligence* 34.2, pp. 420–439. DOI: 10.1111/coin.12120. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/coin.12120. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12120.

Grishman, Ralph (2019). "Twenty-Five Years of Information Extraction". In: *Natural Language Engineering* 25.6, pp. 677–692. ISSN: 14698110. DOI: 10.1017/S1351324919000512.

Grishman, Ralph and Beth Sundheim (1996). "Message Understanding Conference - 6: A Brief History". In: *Proceedings of the International Conference on Computational Linguistics*. Vol. 1, pp. 466–471. DOI: 10.3115/992628.992709.

Gruber, Christine and Eveline Wandl-Vogt (2017). "Mapping Historical Networks: Building the New Austrian Prosopographical Biographical Information System (APIS). Ein Überblick". In: *Europa Baut Auf*

Biographien. Aspekte, Bausteine, Normen Und Standards Für Eine Europäische Biographik. Pp. 271–282.

Hachey, Ben, Will Radford, and James R Curran (2011). "Graph-Based Named Entity Linking with Wikipedia". In: *International Conference on Web Information Systems Engineering*. Springer, pp. 213–226.

Hachey, Ben, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran (2013). "Evaluating Entity Linking with Wikipedia". In: *Artificial Intelligence*. Vol. 194, pp. 130–150. ISBN: 0004-3702. DOI: 10.1016/j.artint.2012.04.005. pmid: 1873413. URL: http://www.sciencedirect.com/science/article/pii/S0004370212000446.

Han, Jiabao and Hongzhi Wang (2021). "Transformer Based Network for Open Information Extraction". In: *Engineering Applications of Artificial Intelligence* 102, p. 104262. ISSN: 0952-1976. DOI: 10.1016/j.engappai.2021.104262. URL: https://www.sciencedirect.com/science/article/pii/S0952197621001093.

Hearst, Marti A. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora". In: *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*. COLING '92. USA: Association for Computational Linguistics, pp. 539–545. DOI: 10.3115/992133.992154. URL: https://doi.org/10.3115/992133.992154.

Hendrickx, Iris, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz (July 2010). "SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: ACL, pp. 33–38. URL: https://aclanthology.org/S10-1006.

Hoffmann, Raphael, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld (June 2011). "Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations". In: *Proceedings of ACL 2011: HLT*. Portland, Oregon, USA: ACL, pp. 541–550. URL: https://aclanthology.org/P11-1055.

Hogue, Alexander, Joel Nothman, and James R Curran (2014). "Unsupervised Biographical Event Extraction Using Wikipedia Traffic". In: *Proceedings of the Australasian Language Technology Association Workshop 2014*, pp. 41–49.

Hovy, Eduard, Ulf Hermjakob, and Deepak Ravichandran (2002). "A Question/Answer Typology with Surface Text Patterns". In: *Proceedings of the Human Language Technology Conference (HLT)*, pp. 247–251.

Huang, Kuan-Hao, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang (2021). "Improving zero-shot cross-lingual transfer learning via robust training". In: *arXiv preprint arXiv:2104.08645*.

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). "Bidirectional LSTM-CRF Models for Sequence Tagging". In: *CoRR*. ISSN: 1098-6596. DOI: 10.18653/v1/P16-1101. pmid: 25246403. URL: http://arxiv.org/abs/1508.01991.

Ji, Baijun, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo (2020). "Cross-lingual pre-training based transfer for zero-shot neural machine translation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01, pp. 115–122.

Ji, Heng and Ralph Grishman (2011). "Knowledge Base Population: Successful Approaches and Challenges". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1148–1158.

Jia, Chen, Yuefeng Shi, Qinrong Yang, and Yue Zhang (Nov. 2020). "Entity Enhanced BERT Pre-training for Chinese NER". In: *Proceedings of EMNLP 2020*. Online: ACL, pp. 6384–6396. DOI: 10.18653/v1/2020.emnlp-main.518. URL: https://aclanthology.org/2020.emnlp-main.518.

Jiang, Jing (2012). "Information Extraction from Text". In: *Mining Text Data*. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Boston, MA: Springer US, pp. 11–41. URL: https://doi.org/10.1007/978-1-4614-3223-4_2.

Jiang, Jing and ChengXiang Zhai (2007). "A Systematic Exploration of the Feature Space for Relation Extraction". In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 113–120.

Jin, Zhuoran, Yubo Chen, Dianbo Sui, Chenhao Wang, Zhipeng Xue, and Jun Zhao (2021). "CogIE: An Information Extraction Toolkit for Bridging Texts and CogNet". In: *Proceedings of the 59th Annual Meeting of the Association*

for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pp. 92–98.

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy (2020). "SpanBERT: Improving Pre-training by Representing and Predicting Spans". In: *Transactions of ACL* 8, pp. 64–77. URL: https://aclanthology.org/2020.tacl-1.5.

Jurafsky, Daniel and James H Martin (2018). "Information Extraction". In: *Speech and Language Processing*. Unpublished draft of 12th August 2018. Available at: https://web.stanford.edu/jurafsky/slp3/.

Kambhatla, Nanda (2004). "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction". In: *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 178–181.

Karthikeyan, K, Zihan Wang, Stephen Mayhew, and Dan Roth (2020). "Cross-Lingual Ability of Multilingual BERT: An Empirical Study". In: *International Conference on Learning Representations*.

Katiyar, Arzoo and Claire Cardie (2016). "Investigating Lstms for Joint Extraction of Opinion Entities and Relations". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 919–929.

Kazama, Jun'ichi and Kentaro Torisawa (2007). "Exploiting Wikipedia as External Knowledge for Named Entity Recognition". In: *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (March), pp. 698–707.

Kim, Jin-Dong, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier (2004). "Introduction to the Bio-Entity Recognition Task at JNLPBA". In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*. Citeseer, pp. 70–75.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (2016). "Neural Architectures for Named Entity Recognition". In: *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pp. 260–270. DOI: 10.18653/v1/n16-1030.

Li, Yaoyong, Kalina Bontcheva, and Hamish Cunningham (2005). "Using Uneven Margins SVM and Perceptron for Information Extraction". In: *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pp. 72–79.

Lin, Yankai, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun (2016). "Neural Relation Extraction with Selective Attention over Instances". In: *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*. Vol. 4, pp. 2124–2133. ISBN: 9781510827585. DOI: 10.18653/v1/p16-1200.

Liu, Yudong, Zhongmin Shi, and Anoop Sarkar (2007). "Exploiting Rich Syntactic Information for Relation Extraction from Biomedical Articles". In: *Proceedings of NAACL 2007: HLT*. NAACL-Short '07. Rochester, New York: Association for Computational Linguistics, pp. 97100.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60. ISBN: 978-1-941643-00-6. DOI: 10.3115/v1/P14-5010. pmid: 25246403. URL: http://www.aclweb.org/anthology/P/P14/P14-5010http://aclweb.org/anthology/P14-5010.

Maynard, Diana (2003). "Multi-Source and Multilingual Information Extraction". In: *Money* 97.98.2, pp. 98–0.

McCallum, Andrew and Wei Li (2003). "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, pp. 188–191.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv: 1301.3781.

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky (2009). "Distant Supervision for Relation Extraction without Labeled Data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, pp. 1003–1011.

Mitkov, Ruslan (1999). *Anaphora Resolution: The State of the Art*. Citeseer.

Miwa, Makoto and Mohit Bansal (Aug. 2016). "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures". In: *Proceedings of ACL 2016*. Berlin, Germany: Association for Computational Linguistics, pp. 1105–1116. DOI: 10.18653/v1/P16-1105. URL: https://aclanthology.org/P16-1105.

Moussallem, Diego, Thiago Castro Ferreira, Marcos Zampieri, Maria Claudia Cavalcanti, Geraldo Xexéo, Mariana Neves, and Axel-Cyrille Ngonga Ngomo (2018). "RDF2PT: Generating Brazilian Portuguese Texts from RDF Data". In: *Proceedings of LREC*.

Nadeau, David, Peter D. Turney, and Stan Matwin (2006). "Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity". In: *Advances in Artificial Intelligence*. Ed. by Luc Lamontagne and Mario Marchand. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 266–277. ISBN: 978-3-540-34630-2.

Nallapati, Ramesh, Bowen Zhou, Cicero dos Santos, Çalar Gulçehre, and Bing Xiang (2016). "Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290.

Nasar, Zara, Syed Waqar Jaffry, and Muhammad Kamran Malik (2021). "Named Entity Recognition and Relation Extraction: State-of-the-Art". In: *ACM Comput. Surv.* 54.1. ISSN: 0360-0300. URL: https://doi.org/10.1145/3445965.

Nayak, Tapas and Hwee Tou Ng (2020). "Effective Modeling of Encoder-Decoder Architecture for Joint Entity and Relation Extraction". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 8528–8535. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6374.

Neelakantan, Arvind and Michael Collins (2014). "Learning Dictionaries for Named Entity Recognition Using Minimal Supervision". In: *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pp. 452–461. DOI: 10.3115/v1/e14-1048.

Ni, Jian, Taesun Moon, Parul Awasthy, and Radu Florian (2020). "Cross-Lingual Relation Extraction with Transformers". In: URL: http://arxiv.org/abs/2010.08652.

Nie, Yuyang, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai (2020). "Named Entity Recognition for Social Media Texts with Semantic Augmentation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1383–1391.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran (2013). "Learning Multilingual Named Entity Recognition from Wikipedia". In: *Artificial Intelligence* 194, pp. 151–175. ISSN: 00043702. DOI: 10.1016/j.artint.2012.03.006.

Pan, Xiaoman, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji (2017). "Cross-Lingual Name Tagging and Linking for 282 Languages". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1946–1958.

Park, Sungjoon, Sungdong Kim, Jihyung Moon, Won Ik Cho, Kyunghyun Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, et al. (2021). "KLUE: Korean Language Understanding Evaluation". In: *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*. Advances in Neural Information Processing Systems.

Petram, Lodewijk, Jelle Van Lottum, Rutger Van Koert, and Sebastiaan Derks (2018). "Small Lives, Big Meanings Expanding the Scope of Biographical Data through Entity Linkage and Disambiguation". In: *CEUR Workshop Proceedings* 2119, pp. 22–26. ISSN: 16130073.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019). "How Multilingual Is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001.

Piskorski, Jakub, Peter Homola, Magorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, and Marcin Woliski (2004). "Information Extraction for Polish Using the SProUT Platform". In: *Intelligent Information Processing and Web Mining*. Ed. by Mieczysaw A. Kopotek, Sawomir T. Wierzcho, and Krzysztof Trojanowski. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 227–236. ISBN: 978-3-540-39985-8.

Piskorski, Jakub and Roman Yangarber (2013). "Information Extraction: Past, Present and Future". In: *Multi-Source, Multilingual Information Extraction and Summarization*. Springer, pp. 23–49.

Plum, Alistair (2018). "Rule-based Information Extraction Using Wikipedia and Wikidata". MA thesis. University of Wolverhampton.

Plum, Alistair, Tharindu Ranasinghe, Spencer Jones, Constantin Orasan, and Ruslan Mitkov (2022). "Biographical Semi-Supervised Relation Extraction Dataset". In: *45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22, pp. 3121–3130.

Plum, Alistair, Marcos Zampieri, Constantin Orasan, Eveline Wandl-Vogt, and Ruslan Mitkov (2019). "Large-Scale Data Harvesting for Biographical Data". In: *Biographical Data in a Digital World 2019*. Vol. 3152. CEUR Workshop Proceedings, pp. 66–72.

Rahimi, Afshin, Yuan Li, and Trevor Cohn (2019). "Massively Multilingual Transfer for NER". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 151–164.

Ranasinghe, Tharindu and Marcos Zampieri (Nov. 2020). "Multilingual Offensive Language Identification with Cross-lingual Embeddings". In: *Proceedings of EMNLP 2020*. Online: ACL, pp. 5838–5844. DOI: 10.18653/v1/2020.emnlp-main.470. URL: https://aclanthology.org/2020.emnlp-main.470.

Reinert, Matthias and Bernhard Ebneth (2017). "Interfaces: Accessing Biographical Data and Metadata". In: *Biographical Data in a Digital World 2017*.

Reinert, Matthias, Maximilian Schrott, Bernhard Ebneth, and Malte Rehbein (2015). "From Biographies to Data Curation the Making of Www.Deutsche-Biographie.De". In: *Biographical Data in a Digital World 2015*, pp. 13–19.

Riedel, Sebastian, Limin Yao, and Andrew McCallum (2010). "Modeling Relations and Their Mentions without Labeled Text". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 148–163.

Russo, Irene, Tommaso Caselli, and Monica Monachini (2015). "Extracting and Visualising Biographical Events from Wikipedia". In: *CEUR Workshop Proceedings* 1399 (November 2017), pp. 111–115. ISSN: 16130073.

Sang, Erik Tjong Kim (2002). "Memory-Based Named Entity Recognition". In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Schlögl, Matthias and Katalin Lejtovicz (2018). "A Prosopographical Information System (APIS)". In: *CEUR Workshop Proceedings* 2119, pp. 53–58. ISSN: 16130073.

Schmitz, Michael, Stephen Soderland, Robert Bart, Oren Etzioni, et al. (2012). "Open Language Learning for Information Extraction". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–534.

Seganti, Alessandro, Klaudia Firlg, Helena Skowronska, Micha\ l Sat\ lawa, and Piotr Andruszkiewicz (2021). "Multilingual Entity and Relation Extraction Dataset and Model". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1946–1955.

Shen, Yatian and Xuanjing Huang (Dec. 2016). "Attention-Based Convolutional Neural Network for Semantic Relation Extraction". In: *Proceedings of COLING 2016: Technical Papers*. Osaka, Japan, pp. 2526–2536. URL: https://aclanthology.org/C16-1238.

Singhal, Ayush, Michael Simmons, and Zhiyong Lu (Apr. 2016). "Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature". en. In: *J Am Med Inform Assoc* 23.4, pp. 766–772.

Smirnova, Alisa and Philippe Cudré-Mauroux (2018). "Relation Extraction Using Distant Supervision: A Survey". In: *ACM Comput. Surv.* 51.5. ISSN: 0360-0300. URL: https://doi.org/10.1145/3241741.

Steinberger, Josef, Jenya Belyaeva, Jonathan Crawley, Leonida Della Rocca, Mohamed Ebrahim, Maud Ehrmann, Mijail Kabadjov, Ralf Steinberger, and Erik Van Der Goot (2011). "Highly Multilingual Coreference Resolution Exploiting a Mature Entity Repository". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 254–260.

Stevenson, Mark and Mark A Greenwood (2006). "Comparing Information Extraction Pattern Models". In: *Proceedings of the Workshop on Information Extraction Beyond The Document - IEBeyondDoc '06*. IEBeyondDoc '06. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 12–19. ISBN: 1-932432-74-4. DOI: 10.3115/1641408.1641410. pmid: 10198117. URL: http://dl.acm.org/citation.cfm?id=1641408.1641410http://portal.acm.org/citation.cfm?doid=1641408.1641410.

Subburathinam, Ananya, Di Lu, Heng Ji, Jonathan May, Shih Fu Chang, Avirup Sil, and Clare Voss (2020). "Cross-Lingual Structure Transfer for Relation and Event Extraction". In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 313–325. DOI: 10.18653/v1/d19-1030.

Sukthanker, Rhea, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu (2020). "Anaphora and Coreference Resolution: A Review". In: *Information Fusion* 59, pp. 139–162.

Sundheim, Beth M (1995). "Overview of Results of the MUC-6 Evaluation". In: *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Surdeanu, Mihai, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning (2012). "Multi-Instance Multi-Label Learning for Relation Extraction". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 455–465.

Tanabe, Lorraine, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur (2005). "GENETAG: a tagged corpus for gene/protein named entity recognition". In: *BMC bioinformatics* 6.1, pp. 1–7.

Toral, Antonio and Rafael Mu (2006). "A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by Using Wikipedia". In: *Proceedings of EACL 2006*, pp. 56–61. URL: http://acl.ldc.upenn.edu/W/W06/W06-2809.pdf.

Tsai, Chen Tse and Dan Roth (2016). "Cross-Lingual Wikification Using Multilingual Embeddings". In: *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pp. 589–598. ISBN: 978-1-941643-91-4. DOI: 10.18653/v1/n16-1072.

Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio (July 2010). "Word Representations: A Simple and General Method for Semi-Supervised Learning". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: ACL, pp. 384–394. URL: https://aclanthology.org/P10-1040.

Vrandečić, Denny and Markus Krötzsch (2014). "Wikidata: a free collaborative knowledgebase". In: *Communications of the ACM* 57.10, pp. 78–85.

Vrandei, Denny and Markus Krötzsch (Sept. 2014). "Wikidata: A Free Collaborative Knowledgebase". In: *Communications of The Acm* 57.10, pp. 78–85. ISSN: 0001-0782. DOI: 10.1145/2629489. URL: https://doi.org/10.1145/2629489.

Walker, Christopher, Stephanie Strassel, Julie Medero, and Kazuaki Maeda (2006). "ACE 2005 Multilingual Training Corpus". In: *Linguistic Data Consortium, Philadelphia* 57, p. 45.

Wang, Hailin, Guoming Lu, Jin Yin, and Ke Qin (2021). "Relation Extraction: A Brief Survey on Deep Neural Network Based Methods". In: *2021 The 4th International Conference on Software Engineering and Information Management*. ICSIM 2021. Yokohama, Japan: ACM, pp. 220228. ISBN: 9781450388955. DOI: 10.1145/3451471.3451506. URL: https://doi.org/10.1145/3451471.3451506.

Wang, Liwen, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu (2022). "Instructionner: A multi-task instruction-based generative framework for few-shot ner". In: *arXiv preprint arXiv:2203.03903*.

Wang, Xinyu, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu (2020). "Automated Concatenation of Embeddings for Structured Prediction". arXiv: 2010.05006.

Wissenschaften, ÖA der (2013). *Österreichisches Biographisches Lexikon 18151950*. Vol. 63. Verlag der Österreichischen Akademie der Wissenschaften.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of EMNLP 2020: System Demonstrations*. Online: ACL, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6.

Wu, Fei and Daniel S Weld (2010). "Open information extraction using wikipedia". In: *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 118–127.

Wu, Shanchan and Yifan He (2019). "Enriching Pre-Trained Language Model with Entity Information for Relation Classification". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. New York, NY, USA: Association for Computing Machinery, pp. 2361–2364. ISBN: 978-1-4503-6976-3. DOI: 10.1145/3357384.3358119. URL: https://doi.org/10.1145/3357384.3358119.

Xie, Jiateng, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell (2018). "Neural Cross-Lingual Named Entity Recognition with Minimal Resources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 369–379. DOI: 10.18653/v1/D18-1034. URL: https://aclanthology.org/D18-1034.

Xu, Peng and Denilson Barbosa (2019). "Connecting Language and Knowledge with Heterogeneous Representations for Neural Relation Extraction". In: *Proceedings of NAACL-HLT*, pp. 3201–3206.

Xue, Kui, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He (2019). "Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 892–897. DOI: 10.1109/BIBM47256.2019.8983370.

Yadav, Vikas and Steven Bethard (2019). "A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2145–2158.

Yamada, Ikuya, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto (Nov. 2020). "LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention". In: *Proceedings of EMNLP 2020*. Online: ACL, pp. 6442–6454. DOI: 10.18653/v1/2020.emnlp-main.523. URL: https://aclanthology.org/2020.emnlp-main.523.

Yang, Wei, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin (June 2019). "End-to-End Open-Domain Question Answering with BERTserini". In: *Proceedings of NAACL 2019*. Minneapolis, Minnesota: ACL, pp. 72–77. DOI: 10.18653/v1/N19-4013. URL: https://aclanthology.org/N19-4013.

Yangarber, Roman (2003). “Counter-Training in Discovery of Semantic Patterns”. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 343–350.

Yates, Alexander, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland (2007). “TextRunner : Open Information Extraction on the Web”. In: *NAACL-Demonstrations ’07 Proceedings of Human Language Technologies*. NAACL-Demonstrations ’07. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 25–26. DOI: 10.3115/1614164.1614177. pmid: 15283663. URL: http://dl.acm.org/citation.cfm?id=1614164.1614177http://portal.acm.org/citation.cfm?id=1614177.

Yu, Amy Zhao, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A Hidalgo (2016). “Pantheon 1.0, a Manually Verified Dataset of Globally Famous Biographies”. In: *Scientific data* 3.1, pp. 1–16.

Yu, Dian, Heng Ji, Sujian Li, and Chin-Yew Lin (2015). “Why Read If You Can Scan? Trigger Scoping Strategy for Biographical Fact Extraction”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1203–1208.

Zaila, Yisleidy Linares and Danilo Montesi (2015). “Geographic Information Extraction, Disambiguation and Ranking Techniques”. In: *ACM International Conference Proceeding Series* 26-27-Nove. DOI: 10.1145/2837689.2837695.

Zeng, Daojian, Kang Liu, Yubo Chen, and Jun Zhao (2015). “Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1753–1762.

Zeng, Daojian, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao (Aug. 2014). “Relation Classification via Convolutional Deep Neural Network”. In: *Proceedings of COLING 2014: Technical Papers*. Dublin, Ireland: Dublin City University and ACL, pp. 2335–2344. URL: https://aclanthology.org/C14-1220.

Zhang, Shaodian and Noémie Elhadad (2013). “Unsupervised Biomedical Named Entity Recognition: Experiments with Clinical and Biological Texts”. In: *Journal of Biomedical Informatics* 46.6, pp. 1088–1098. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2013.08.004. URL: https:

/ / www . sciencedirect . com / science / article / pii / S1532046413001196.

Zhang, Yuhao, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning (Sept. 2017). "Position-aware Attention and Supervised Data Improve Slot Filling". In: *Proceedings of EMNLP 2017*. Copenhagen, Denmark: ACL, pp. 35–45. URL: https://aclanthology.org/D17-1004.

Zhao, Shubin and Ralph Grishman (2005). "Extracting Relations with Integrated Information Using Kernel Methods". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (Acl05)*, pp. 419–426.

Zheng, Suncong, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu (2017). "Joint entity and relation extraction based on a hybrid neural network". In: *Neurocomputing* 257. Machine Learning and Signal Processing for Big Multimedia Analysis, pp. 59–66. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2016.12.075. URL: https://www.sciencedirect.com/science/article/pii/S0925231217301613.

Zhou, GuoDong and Jian Su (2002). "Named Entity Recognition Using an HMM-based Chunk Tagger". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 473–480.

Zhou, Liang, Miruna Ticrea, and Eduard Hovy (2004). "Multi-Document Biography Summarization". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 434–441.

Zhou, Peng, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu (Aug. 2016). "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification". In: *Proceedings of ACL 2016*. Berlin, Germany: ACL, pp. 207–212. DOI: 10.18653/v1/P16-2034. URL: https://aclanthology.org/P16-2034.