Master Thesis

Automatically building annotated training data in other languages

S. Chatzopoulou

Master Thesis DKE-20-08

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science of Artificial Intelligence
at the Department of Data Science and Knowledge Engineering
of the Maastricht University

Thesis Committee:

Dr. J.C. Scholtes
Dr. J. Spanakis

Maastricht University
Faculty of Science and Engineering
Department of Data Science and Knowledge Engineering

February 28, 2020

Acknowledgement

I would like to express my gratitude and appreciation to Dr. Johannes C. Scholtes, Mr. Jeroen Smeets and Ms Zoe Gerolemou for their excellent guidance and inspiring mentorship that helped me enrich my knowledge and grow. Along with them, I want to thank everyone at ZyLAB who welcomed me to their team so warmly. Finally, I would like to thank my family and friends for standing next to me every step of the way. I, wholeheartedly, dedicate this thesis to them. Nothing would have been possible without their priceless support.

Abstract

In machine learning, and especially in text mining or information retrieval applications, the availability of enough annotated training data is important. Multiple machine learning applications are language sensitive, and as a result the need for annotated data sets in multiple languages is critical for their performance. Even though annotated data sets in English are common, this is not the case for all languages, with having some languages not represented at all. Manual creation of data sets in all available languages is too slow, too expensive, too arduous and prone to human errors. The goal of this assignment is to develop automatic processes that create annotated data sets for other languages from English annotated dataset. Two foreign languages are investigated, Dutch and German, and models to automatically create annotated corpora in these languages are presented. The baseline of just matching the text of an english named-entity on the Dutch corpus for the English - Dutch pair achieves 49% when evaluated on the golden standard CoNLL dataset the same result for the English-German corpus is 46% when evaluated on the golden standard GermanNER. In this thesis, this is improved by using similarity metrics to increase the projection and a translation step to translate words to the foreign language. Additionally, when a BiLSTM model is used it achieves 66% on CoNLL for the English-Dutch pair and 57% on GermanNER for the English-German pair.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter a description of the problem is defined followed by the motivation for creating an automatic mechanism to create training data sets for other languages from an English language training dataset model for this purpose. The research questions are described with a brief explanation of the motivation and the goals.

In machine learning, and especially in text mining or information retrieval, the availability of enough annotated training data is important. Text analysis is facilitated by such annotated training data also known as corpus.

The source of the text varies. It can be obtained from newspapers, Wikipedia articles, tweets and, in general it can include written or spoken language. The structure, the scheme and the format of the annotated text differs depending on the task and the availability as explained on Mason, 1999 [17]. The annotation of a textual corpus can be done manually by an expert in the field depending on the task subsequently a machine learning model can then be trained on this dataset.

By training machine-learning models on annotated corpora, textual analysis is possible and, subsequently, the extracted information can be used to build systems for various applications e.g. question answering, document classification, machine translation etc.

1.1 Problem Description

Multiple machine learning applications are language sensitive and as a result the need for annotated datasets in multiple languages is critical for their performance. Even though annotated datasets in English are abundantly available, that is not the case for all languages, with some languages being not represented at all. Manual creation of datasets in all available languages is too slow, too expensive, arduous and prone to human errors.

The goal of this assignment is to develop automatic processes that create annotated datasets for languages other than English. An overview of the prob-

lem this assignment aims to solve and its main goal is on image 1.1. As it is shown the goal is to transfer the annotation information (that can be any kind of annotations in English text) to a foreign language and then use this data-set to build a machine learning model based on those annotations. The main focus of this thesis is to explore different procedures for German and Dutch as target languages and create machine learning models that are evaluated on the task of named-entity recognition (NER). The reason for the choice of these two languages are the availability of parallel corpus, evaluation datasets and pretrained named-entity recognition models.

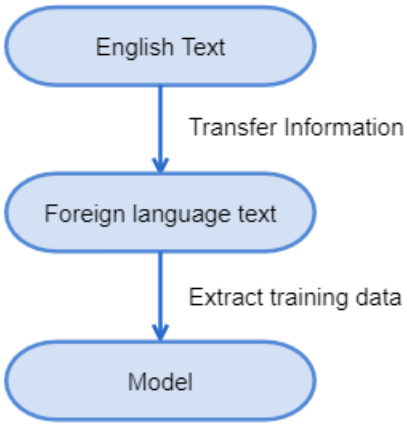


Figure 1.1: Problem Description

Named-entity recognition is the task of identifying phrases or words that contain the names of persons, organizations, locations, times or quantities as defined in Tjong Kim Sang, 2002 [33]. In this assignment we only focus on the named entities corresponding to person names, locations and organisations.

The key challenges in this project are the ability to deal with the ambiguity and variability of the different languages that may or may not have the same alphabet, syntax or grammar. Furthermore, the quality of annotations should be equivalent or comparable to the original data. In the baseline configuration the English word is simply looked-up in the foreign text without any preprocessing, or the use of any translation between English and the foreign language and using plain string matching.

1.2 Research Questions

The research questions this thesis aims to address, are formulated and described below.

- **Which models perform better when it comes to the automatic**

creation of annotated datasets for other languages? Multiple models and techniques are investigated to create annotated corpora in Dutch and German by transferring the annotations from an English parallel aligned corpus. Then the annotated corpora are used to train named-entity recognition models and evaluate their performance.

- **What is the quality of such models for a classification task like named-entity recognition and what is the performance difference on the original language compared to other languages?** An evaluation and comparison between the trained named-entity recognition system and the state-of-the-art for the specific language and model are presented.

- **What is the exact quality improvement of such models measured in different well established metrics (e.g. Accuracy, Recall, Precision etc)?** A comparison of the different models tested to increase the number of correctly projected entities and how well the various models performed are presented in the form of tables including well-established metrics.

- **What are limitations of the approach, what are the main causes of the final error and how could this be improved?** A presentation of the errors incorporated in each step of the approach is reported and suggestions to overcome those are being made.

- **Which pre- and post-processing techniques can be applied to improve the performance?** Different pre-processing and post-processing techniques are applied to identify the ones that increase the correctly projected entities from one language to another with the final goal being the performance improvement of the named-entity recognition model.

Chapter 2

Related Work

This chapter consists of four sections. In the first section the paper in which this thesis is based on is presented with other multilingual approaches whilst the approaches that are used for domain adaptation are described in the second one. Data augmentation and methods to handle to tackle the long-tail problem are introduced in sections three and four.

2.1 Multilingual Approaches

2.1.1 Named Entity Annotations

The research conducted in this thesis is based on the work of Ehrmann et *al.*, 2011 [8]. In this paper a parallel corpus, available in multiple languages including English, is used. The English text is annotated and subsequently these annotations are transferred to other languages with the aim of creating annotated training data. A phrase-based Statistical Machine Translation model is trained on some aligned corpora between a source language and a target language; English and Dutch for instance. To ensure the word-to-word alignment IBM alignment models and proximity rules were applied to obtain the final phrases. This model is used to translate previously extracted entities into the target language and then correct them using stop-word lists and multilingual named-entities database. Later, these entities are projected in the target text using string matching, consonant signature matching or/and similarity-distance metrics. The innovative suggestions of this paper is that it is possible to increase projection recall to a large degree using a non-supervised method.

The approaches for the projection of entities using plain string matching or similarity metrics are also used in this study enriched by the use of preprocessing and postprocessing techniques. This thesis differs by the use of precision and the f-score for the evaluation of the model on a sample of the dataset. In detail, a sample from the aligned corpus is taken and annotated for both languages (source and target) and subsequently the precision and F-score is calculated for every step. Furthermore, instead of training a translation system from scratch, a

pretrained one is used and different machine-learning approaches for the named-entity recognition are explored.

On paper Shah et *al.*, 2010 [32] machine translation from Swahili to English using available services of Google translate or Bing is explored. The translation is done sentence-by-sentence and after smaller sentences are preprocessed if needed. For the named-entity recognition on the English corpus, Stanford's Conditional Random Field named-entity recognition system is used and UIUC'S learning-based Java named-entity tagger. The word alignment between English and the new language makes use of a brute-force approach to translate every entity back to the source language. This is achieved by comparing it with a window of words and identifying the named entities in English in the translated document. This approach uses GIZA++ to align the words from the translated input in English and the source. Two main post-processing techniques are applied; if a word annotated as a named-entity occurs in non-named entities but not in named ones then the annotation is removed and if a word is not annotated as named-entity but appears in a named-entity but not in a non-named one then it is annotated. A token is categorized as being part of an entity if it is classified as such by either Stanford or UIUC system.

2.1.2 Semantic Annotations

In the Bentivogli and Pianta, 2005 [4] the authors attempt to transfer the annotations from one annotated corpus to its translated text in a target language by leveraging the fact that they are aligned. The annotations on the initial document are semantic. The workflow is first to annotate the initial document, translate the words and by projecting those annotations on the target text by aligning the two texts. Experts are used for the translation component whilst Knowa is employed for aligning the words. By following these steps, the annotations are projected and subsequently enriched by the addition of lemmas and part-of-speech (PoS) tags. The annotator performance is measured using the Dice coefficient. This approach is similar to what other papers presented previously but it varies in the way it handles the sentence and word alignments as well as the errors that originate from it. Because of the use of already-aligned corpora, these methods are not used in this assignment.

In the Pado and Lapata, 2005 [24] the FrameNet tool is used in the English corpus, a method which employs parallel corpora for acquiring frame elements and their syntactic realizations. Firstly, given a pair of sentences E and L that are translation of each other, it annotates E with semantic roles and then projects these roles into L. In the paper, a word-based model is presented which uses source and target word tokens as entities for the projection. A constituent-based model is evaluated while projecting from and to components (obtained from full parse trees). Two main subtasks are explored. A real-valued similarity function between source and target elements is employed to align relevant constituents and project the role info. The problem of determining the similarity between category-based representations is tackled by defining it as the proportion of target tokens aligned to source tokens which implies that larger

sentence slices tend to be less similar because of missing alignments. A forward constituent alignment which aligns the source constituents that form the span of a role to a single target constituent and a backward constituent alignment are explored.

The evaluation of the English-German sentence pairs with a manual FrameNet frame and role annotation is done using GIZA++ tool to induce word alignments, gather sentences that with at least one pair of aligned words which were listed in FrameNet and SALSA (which produces realistic corpus' sample for the role of projection task) and had at least one name in common. One alternative is to manually annotate the dataset. To measure the agreement: Kappa score, ratio of common frames between 2 sentences and validity of common roles with identical spans are used to investigate the degree of semantic parallelism between the 2 languages. The conclusions are that cross-lingual data exhibit more than twice the amount of frame differences than monolingual data when comparing word-alignment models against more resource-intensive models.

In the paper Yarowsky et *al.*, 2001 [36] approaches for automatically creating annotations for part-of-speech, noun phase, named entities and morphological for a foreign language are explored. These approaches are tested on bilingual text with the the English text annotations being projected onto the second language via statistically-derived word alignments. The corpora are word-aligned using the EGYPT system as presented on Al-Onaizan et *al.*, 1999 [23]. For the Part-of-Speech projection, the tags are manually corrected. Later, standard bigram taggers are trained on automatically-projected data but overfit. Finally, lexical prior estimation is also utilised. This is defined as the hierarchically smoothing the lexical prior model bias for the two most common tags and the expansion of model probability of majority part-of-speech tag. Eventually, the probabilities of all tags ranked second or lower are reduced. The evaluation of the procedure is based on its accuracy when applied on two evaluation datasets. For the noun-phrase bracketer, word alignments are used for the projection of tagging and bracketing English data. For each word in a noun phrase-subscripted with the number of noun phrase in the sentence subscript projected onto aligned target text to the corresponding target noun phrase is simply the max span of the projected subscript. The model focuses on training on highest quality projected data and excluding parts with the largest word-alignment error. For the evaluation, the exact match of the brackets is calculated using well-established metrics like Precision, Recall and F1-score. For the named-entity recognition, a per-word classifier and a transitive projection model are used. The final training data are then used to train the tagger for the target language. The performance of projection and the accuracy of the model are then assessed.

2.1.3 Approaches using Wikipedia data

Ghaddar and Langlais, 2017 [9] propose the creation of annotated corpora using Wikipedia. This is achieved by gathering anchored strings from Wikipedia and their type according to Freebase and expanding these by adding annotations

for texts that are not anchored in Wikipedia with their mentions as candidate annotations. Freebase is a knowledge base which includes structured data from multiple sources while Wikipedia is an online encyclopedia. Their main contribution is in how to deal with non-anchored strings in Wikipedia by analysing outline structure and removing mentions that do not have a page in Freebase. The named entities are defined as direct out-links or out-links of out-links of a target article both labelled using Freebase and as if they are coreferent mentions of any of the above two. The approach is only applied in English text. To assess the performance the out-of-domain F1 score is used that measures how well a NER performs on out-domain material as well as on cross domain datasets. This is also used for the evaluation in this thesis since they are tested on datasets from different distributions than the ones that the models are trained on. Furthermore, the miscellaneous tag is not included in the training or validation data to be able to assess the performance independently from it.

On Nothman et *al.*, 2013 [22] a different approach is followed, they first explore monolingual models to classify Wikipedia's articles and try to project the entities using the Wikipedia's outgoing inter-language links. Finally, each link is labelled based on the entity type of the target. Three classification approaches are tested and extended to nine languages. The first one uses key phrases from English Wikipedia category names that correspond to named-entity tags as described in Richman and Schone, 2008 [30], while the second is a semi-supervised classification approach using features in bootstrapping as proposed in Nothman et *al.*, 2008 [21]. Finally, classification using features from the first couple of sentences is explored. For each approach, both monolingual and multilingual classification approaches for each language are explored. They are evaluated on manually-annotated Wikipedia articles; either picked from the most popular ones or randomly chosen.

Even though these studies make use of interesting approaches they are not employed in this research because of the need of Wikipedia data and their assumption of Wikipedia structure, of anchored links and text. The evaluation on a manually-annotated or a golden-standard corpus is used as well as on annotated Wikipedia articles.

2.1.4 Using time distributions

Klementiev and Roth, 2008 [14] suggest the use of two characteristics to identify named-entities; namely, a transliteration-based and a time/ sequence/ similarity-based one. For a given named-entity in one language, the model chooses a list of top-ranked transliteration candidates in another language. Then, a Discrete Fourier Transform-based metric is used to re-rank the list and choose the candidate that is best temporally-aligned with a given named-entity. The candidates are used for the next iteration of the transliteration model which is generative. For a given named-entity, the transliteration model constructs a candidate list for each constituent word and if a dictionary is available, each list is augmented with translations.

In the temporal-based approach, the score is used to re-rank the aforemen-

tioned candidate lists for each token and the alignment resolves many ambiguities. The top candidates from each re-ranked list are merged into possible target language and are verified that candidate named-entity occurs in target. The temporal alignment is used as a supervision signal to iteratively train a transliteration model whilst the F-index is used to measure the similarities. Discrete Fourier Transform is applied to a time sequence to extract its Fourier expansion coefficients and the score of a pair of time sequences is computed as an Euclidean distance between expansion coefficient vectors.

For the test phase, candidate lists are collected for every constituent word for each source entity using a trained model, and augmented with dictionary translations of this particular word. The lists are then re-ranked without a threshold, and grouped into a multi-word target named-entity candidate. Those that do not actually occur in target corpus are discarded.

Although this approach is interesting, it is proved to be too complicated and not really needed since simpler approaches are easier to implement and faster to obtain results without getting into complex mathematics.

2.1.5 Treebanks

On the paper by Volk et *al.*, 2010 [35] a semantic annotation in parallel treebanks is investigated. A treebank is defined in the paper as an assortment of syntactically-annotated sentences. This approach builds monolingual trees which are aligned on word- and phrase-level using a tree aligner that was prior proposed by the authors. By aligning two trees, the translation equivalence of each word or phrase can be identified while the edge labels that contain function labels information is not lost. When it comes to named-entity classification, the PERSON or LOCATION entities can be found by referring to their Wikipedia page or, depending on the language and the available data, to other Gazetteers for confirmation. These alignments lead to the advantage that the verified information needs to be saved only once and used in multiple languages. This approach depends on publicly available gazetteers and aligned trees for each language and was not used as the main approach but ideas were considered in this research.

2.2 Domain Adaptation

In the work Kitoogo et *al.*, 2008 [13] domain-independent features are proposed as a way to tackle the problem of domain-dependent named-entity recognition. A machine-learning model using the maximum entropy is trained and the optimal feature selection is done using a multi-objective algorithm. The features that are suggested in this paper are divided in three main categories: the local features, the global features and the domain-independent features. The latter include information about the proportion of capitalized in a document from another domain, if this word appears with the same class suffix or unigram or bigram there. The global features contain the same information extracted from

the specific document while the local features are of orthographic, contextual, part-of-speech, lexicon features etc. This is the rationale behind adding Lexical features later in our approach.

Additional preprocessing techniques are explored and motivated by the thesis of Biswas, 2019 [5] which employ string matching using similarity metrics. In this thesis, Levenshtein and token-based techniques like soft TF-IDF and cosine similarity are explored as well as soft soft TF-IDF and Jaccard. The preprocessing techniques that are tested strip the punctuation and the extra white spaces, convert to lower case, deduplicate of the strings, remove of any token in a string which contains a number, normalize all other alphabets to English, remove designations and replace abbreviations. Furthermore, numerical representation of strings is performed using distance metrics and features as boolean values or percentages. Finally a Support Vector Machine, an XgBoost model and a Siamese Convolution Neural Network are trained and evaluated.

In this section papers where embedding approaches are investigated will be explored. Firstly, on Akbik et *al.*, 2018a [2] contextual string embeddings are presented. A word may be associated with several meanings depending on its surrounding context. This class of embeddings is capable of expressing that with the use of context-dependent representations. The contextual string embeddings work by encoding each word of a sentence using a bidirectional character-level language model which outputs internal character states and forms word-level embeddings which are then passed to a sequence-labelling model. To extract the contextual string embeddings Long Short-Term Memory neural networks are fed with a sequence of characters with the purpose of predicting the next character. For each word, the hidden states of the model are concatenated and output after the last character of the word. For the sequence-labelling task a Bidirectional Long Short-Term Memory neural network is used. Different combinations of architectures and embeddings are tested and this approach outperforms the rest in the task of named-entity recognition.

An improvement on the previously mentioned contextual string embeddings is presented by Akbik et *al.*, 2019 [1]. A problem that the previous word embeddings are introducing is that in the case of rare words in underspecified context, there is a large probability that these words will be misclassified on tasks such as named-entity recognition. The solution is given by the use of pooled contextual embeddings. These work by extracting the embeddings of each word and then adding them to a memory. Then a pooling function is applied to the memory list of this specific word and then the final embedding of this word is extracted by concatenating the pooled and the extracted contextual embedding. Pooling methods that are tested are the mean, min and max pooling, while the embeddings are not reset and the memory is reset in every epoch. The approach outperforms the previous model. Judging from the published results, the contextual embeddings are used to train a named-entity recognition system as described in the previous paper.

On M. E. Peters et *al.*, 2018 [25] Embeddings from Language Models (ELMo) are introduced. Vectors derived from a bidirectional Long Short-Term Memory are used and the final embeddings are calculated using two-layer Bidirectional

language models. The forward language model outputs a probability of the token given the history while the backward language model does the opposite work, trying to predict the previous token. After training the ELMo embeddings, they can be used for any natural-language processing task. The method presented is not explored in this thesis as it is outperformed by other approaches.

On Devlin et al., 2018 [7] the language-representation model of Bidirectional Encoder Representations from Transformers (BERT) is demonstrated. BERT consists of pre-training and fine-tuning. On pre-training the model is trained on unlabelled data while on fine-tuning the parameters extracted from pre-training are used as the initial values to be finetuned. The architecture used to achieve this is a multi-layer bidirectional Transformer encoder and two model sizes are explored. BERT's inputs are the tokens after applying WordPiece embeddings. A masked language model and a next-sentence-prediction model are used to pretrain it. By masked language model the authors mean that they cover some of the input randomly and then predict it but do not replace the actual words. For the fine-tuning self-attention, is used to encode a pair of sentences.

Finally, on Pires et *al.*, 2019 [28] the multilingual BERT model is evaluated. The main concept of multilingual BERT is that all languages are embedded in a single shared space. The conclusions of the paper are that multilingual BERT generalizes well while it is able to capture multilingual representations even for languages that do not share the same alphabet. The multi-lingual BERT is a 12-layer transformer trained on 104 languages with a shared vocabulary. The tokenized sentences are fed to BERT and the output is given as input to a labelling model. Depending on the similarity of the languages, multilingual BERT generalizes accordingly. An example that is given, is the generalization between Japanese and English where the performance is lower than in languages which have similar word-order features. BERT embeddings are explored and evaluated with contextual string embeddings in the form of a stacked embedding and lead to notable results.

2.3 Data Augmentation

The ideas from this paper could be used to augment a foreign dataset containing projected annotations or even to augment the initial English dataset. Moreover its proposed techniques for the automatic annotation, training and evaluation of the classifier could be utilised.

The main focus of Mesbah et *al.*, 2019 [18] is the generation and annotation of those data. The data are health-related and variational auto-encoders are used to learn the distribution of input data to be able to generate similar text. Subsequently, these are automatically annotated with adverse-drug-reactions mentions by measuring and aggregating the semantic relatedness between terms relevant to these reactions and positive terms excluding stopwords and negative examples. For the semantic relatedness, Word2Vec implementation of skip-n-gram word embeddings are used. The semantic relatedness of a term can be positive or negative and is defined as the sum of the individual semantic

relatedness of each term and every positive/negative term divided by the number of these terms. In case of ambiguity, each term is annotated as positive if it appears in the positive term or if its semantic relatedness and positive term is greater than the term and negative term or if threshold is higher than a given threshold (th). Finally, a Conditional Random Field sequence model is trained and evaluated on both the generated and initial data. The results are compared with state-of-the-art models and other alternative approaches.

2.4 Long-Tail in named-entity Recognition

In the paper by Mesbah et *al.*, 2018 [19] an iterative approach for long-tail entity extraction named Term and Sentence Expansion Named-Entity Recognition(TSE-NER) is described. Long tail of a distribution in named-entity recognition is defined as the portion of the examples having less occurrences from the average. TSE-NER works by taking into consideration that there are patterns when identifying domain-specific named entities. Firstly training data are created using seed terms, then the set is expanded and annotated. After a new named-entity recognizer is trained, the extracted named entities are filtered and used as seed terms. On all occurrences Word2Vec by Mikolov et *al.*, 2013 [20] of skip-n-gram word embeddings and the cosine distance between two vectors for semantic relatedness are employed to identify more terms. Then K-mean clustering is used to identify entities of the same type. On the contrary Doc2Vec embeddings are utilised for sentence expansion. For training a new named-entity recognition model conditional random fields are used. Finally the terms are filtered by stopwords, similar terms, pointwise mutual information, as well as using lookup tables and ensembles.

On the Vliegenthart et *al.*, 2019 [34] the problem of Long-Tail in named-entity recognition is addressed and some approaches that are interesting for our case are presented. By incorporating user feedback on the relevance of the classified entities, an increase in the performance is observed. For named-entity extraction, a TSE-NER system is used, a small seed of known entity instances is set. For each type, it is sufficient to have one or two domain experts to denote 5-50 known entities. The sets are then heuristically expanded and annotated to generate training data. These data are then used to train a named-entity classifier. The next step is to heuristically remove false positives to create entity set for the next iteration. This results to an extension of the TSE-NER with incremental, collaborative feedback from human contributors to support the heuristic filters. The heuristic filtering involves the use of wordnet, stopwords, similar terms, pointwise mutual information, knowledge base lookup, and ensemble majority voting like the initial algorithm. The resulting trained model is applied to all documents and users can then interact with the recognized entities using the Coner tool. This human feedback is what differentiates this paper from previous studies.

These two approaches are interesting as they present models and preprocessing methods to use when building a named-entity recognition model for a

specific language. They do not introduce ideas though that could be used in a multi-lingual named-entity recognition system.

Chapter 3

Approach

This chapter provides an overview of the main approach that is explored, its alternatives and the different machine learning models that are trained and tested with the purpose of increasing the performance of the overall system. The baseline and enhancements applied on the baseline are mostly based on Ehrmann et *al.*, 2011 [8].

3.1 Pipeline phases

The approach is composed of three main parts as it is shown on image 3.1. Two main corpora are used for building, training and evaluating the system. A multilingual dataset with aligned sentences is required for the construction of the model and a gold-standard corpus is needed for the assessment of the named-entity recognition. The reason behind using an aligned dataset is to be able to project subsequently the named-entities from 'source' language to 'target' language(e.g. Dutch or German)

3.1.1 Source Language named-entity Recognition

The first component uses a pre-trained named-entity recognition model to extract the named entities from the source language corpus. After their extraction a postprocessing is important.

3.1.2 Projection Step

The second part of the model projects the named-entities in source language to the aligned sentences in target language(e.g. Dutch or German). Some approaches for the normalization and preprocessing before the projection are investigated along with the use of similarity metrics instead of exact string matching.

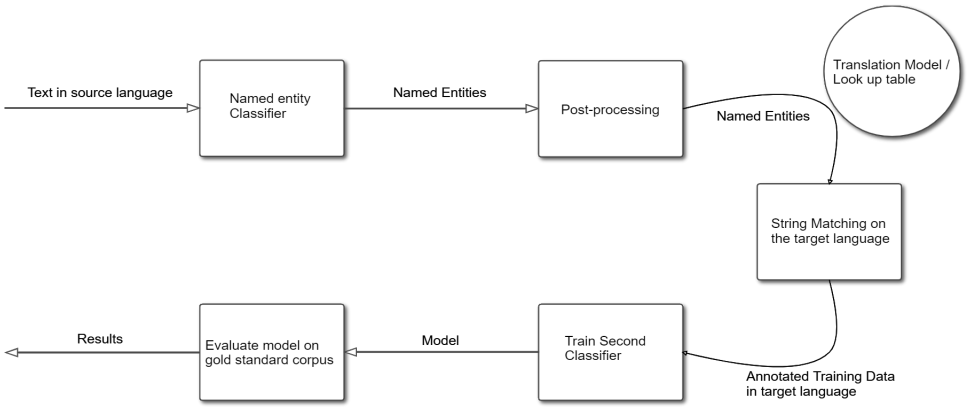


Figure 3.1: Language Model

For example let's assume that we have two sentences: one in English "The unilateral decision by Belgium to re-establish border controls is a clear illustration of this." and one in Dutch: "Kijk maar naar het unilaterale besluit van België om de controle aan zijn grenzen te herstellen.". The workflow will be to first apply an English named-entity recognition (first component) on the English sentence to extract "Belgium" as a named-entity of type LOCATION. Next an attempt will be made to project it to the Dutch sentence. In this case, plain string matching cannot be successfully used for identifying the entity in the Dutch sentence. This is the main motivation for using a look-up table between these two steps that maps the word in 'source' language to the corresponding 'target' language one or alternatively use a different projection method.

For the projection (part 2) two main approaches are explored: exact string matching and string matching using similarity metrics. For the exact string matching seven pre-process techniques are investigated on both the initial text and normalised text. The normalization is applied both in the text from the source language and the text in the target language before trying to project the first to the latter. For the normalization of the text six approaches are explored: a)removal of accents and lowercasing, b)use of the soundex algorithm, c)use of stemming, d) a) and removal of double letters, e)use of consonant signature, f) combination of c) and e). For projection using similarity metrics four metrics are tested: Levenshtein, TF-IDF secondary similarity function with one of the Levenshtein , Jaro and Jaro-Winkler.

In the example used above where the goal is to project from English "Belgium" to Dutch 'België ', if we try lowercasing and then removing the accents we will end up with "belgium" and "belgie" which is not possible to match. In case the Soundex algorithm for "Belgium" is used, it will output 'B425' and for "België" 'B42', so again they will not match. Even if we use the stemming, or remove the double letters (which in this case there is none) there will be no match between the two words. Finally, in the case of using only the consonant

signature 'blgm' and 'blg' are produced which again do not match. The previous example drives the decision to explore similarity metrics.

One additional thing that could be done between the English Named-Entity Recognition and projection steps is the translation of the named entities before trying to project them on the target language to avoid mistakes due to transliteration. This is implemented in the model named "Translation & Similarity metrics projection" that was tested by using similarity metrics in addition to plain string matching. For the translation component a look-up table including all cities, countries and organisation names in both languages was constructed as well as, already implemented automated translation modules and pre-trained translation models.

3.1.3 Target Language Named-Entity Recognition

For the third and last component of the model a new named-entity recognition model is trained on the data that were created from the projected entities in the target language. Two main models were investigated for named-entity recognition: a Conditional Random Field and a bi-directional Long Short-Term Memory Network with a Conditional Random Field decoder. A Conditional Random Field is a statistical model which predicts a tag by taking into consideration context and nearby samples as it is described in Lafferty et *al.*, 2001 [16]. Long Short-Term memory network is a recurrent neural network with feedback connections and it is used successfully for sequences as described in Hochreiter and Schmidhuber, 1997 [10]. The bidirectional Long Short-Term Memory Network with a Conditional Random Field decoder as described in Akbik et *al.*, 2018b and Huang et *al.*, 2015 [12, 3] was considered to be the state-of-the-art for named-entity recognition. To improve the last model, the use of character embeddings was investigated.

3.1.4 Baseline and Enhancements

The baseline system used an existing model pretrained on OntoNotes corpus (Hovy et *al.*, 2006 [11]) for the English Named-Entity Recognition. Alternatively, a model trained on the English version of CoNLL 2003 corpus or the same corpus with manual annotations to improve the precision and recall could be used. After the postprocessing that was explained before, the projection step with plain string matching is used and then the training of the Conditional-Random-Field-based named-entity recognition model.

3.2 Datasets

For the purpose of this thesis publicly available corpora from multiple domains and languages were used. For the implementation of the model the European Parliament Proceeding Parallel corpus is mainly used that was introduced on Koehn and Monz, 2005 [15]. This corpus includes aligned text in all languages

spoken in the countries belonging to the European Union. For this thesis the last version (version 7) was retrieved and for the experiments the German English and Dutch English subsets are used. The size of the German English corpus is 1,920,209 sentences while the Dutch English corpus is 1,997,775 sentences.

Name	Language	Type	Size
European Parliament Proceedings Parallel Corpus 1996-2011	English-Dutch	European Parliament Agendas	1,997,776 sentences
European Parliament Proceedings Parallel Corpus 1996-2011	English-German	European Parliament Agendas	1,920,209 sentences
Tatoeba	English-Dutch	Sentences and their Translations	44,445 sentences
GlobalVoices	English-Dutch	Newspaper	42.251 sentences

Table 3.1: Parallel Corpora

For the evaluation of the models both on German and Dutch experiments the CoNLL annotated corpora are employed. CoNLL 2003 [Sang and De Meulder, 2003 [31]] and CoNLL 2002 are used for German and Dutch, respectively.[Tjong Kim Sang, 2002 [33]] CoNLL stands for Computational Natural Language Learning and is a conference that announces a shared task with the corresponding dataset every year. The years indicate when each dataset was presented. The Dutch Corpus contains sentences from the Belgian newspaper "De Morgen" and it consists of three columns: the word itsef, its part-of-speech and named-entity tags. The named-entity tag can indicate a person, an organisation, a location or a miscellaneous entity and it is possible to be composed of more than one token. These are represented with the use of the Inside-Outside-beginning scheme (IOB) as explained in Ramshaw and Marcus, 1995 [29] which uses prefixes "B" and "I" for the first and the trailing tokens of an entity, respectively. For non-entity tokens, the "O" prefix is used. For example in the sentence: "Eind 1998 telde de EU zo'n 360 nationale tv-stations.", "EU" is annotated as an organisation using the B-ORG tag while the rest of the tokens are assigned with an 'O' tag. In another example the name "Kristof Demasure" is classified by having its two tokens assigned with tags as follows: "Kristof" as B-PER and "Demasure" as I-PER. In the example sentence "Maar ook zij hebben door dat continentaal Europa achter loopt in de conjunctuurcyclus." the "Europa" token is labelled as a location using the "B-LOC" tag while the rest of the tokens are categorised as non-entities using the "O" tag. The German corpus has sentences extracted from the German newspaper Frankfurter Rundshau and it consists of four columns the word and its part-of-speech, chunk and named-entity tags. The named-entity tagging scheme (which is used for the purposes of this thesis) has the same type as the Dutch corpus. For the evaluation of the English named-entity recognition task the English CoNLL 2003 corpus is used.

To increase the performance of the model, alternative corpora with aligned

sentences from different domains were investigated. Tatoeba[1] is a collection of sentences and their corresponding translations which is mostly used for building translation systems. From tatoeba 44,445 sentences were retrieved. Global Voices Parallel Corpus[2] contains news articles from the Global Voices websites and consists of 42,251 aligned sentences.

Name	Language	Document Type	Scheme	Size
CoNLL 2003	English	Newspaper	CoNLL-IOB1 converted to IOB2	14,041 sentences
CoNLL 2003	German	Newspaper	CoNLL-IOB1 converted to IOB2	2,867 sentences
CoNLL 2002	Dutch	Newspaper	CoNLL-IOB2	15,806 sentences
GermanNER	German	Newspaper	CoNLL-IOB2	5,100 sentences
WikiNER	Dutch	Wikipedia	pipe-delimited converted to CoNLL-IOB1 converted to IOB2	184,331 sentences

Table 3.2: Evaluation Datasets

For the evaluation of Dutch and German models a silver-standard annotated corpus with text and structure of Wikipedia is additionally used. More information about the creation of this corpus can be found in the Nothman et *al.*, 2013 [22]. The Dutch models in particular, they are evaluated on Dutch Parliamentary Documents [3].

[1] https://tatoeba.org/eng/

[2] http://opus.nlpl.eu/GlobalVoices.php

[3] https://github.com/Poezedoez/Named-Entity-Recognition-on-Dutch-Parliamentary-Documents-using-Frog/blob/master/Code/data/lobby/golden_standard

Chapter 4

Experiments

This chapter provides an overview of all the experiments ran on English-Dutch and English-German corpora, the approaches followed and the machine learning models that are tested and reported.

4.1 Baseline

To create automatically annotated corpora in the Dutch language the European Parliament Proceedings corpus is used for the pair of English-Dutch and English-German language. As it is stated above the Dutch-English corpus consists of 1,997,775 aligned sentences while the German-English of 1,920,209. First we test the baseline, the English text is annotated with the named entities using a pre-trained model trained on the OntoNotes dataset. Then the named entities are projected in the aligned sentences in Dutch/German language using plain string matching without any preprocessing. Finally only sentences containing at least on tagged entity are kept and given as training data to a Conditional Random Field model.

To evaluate the performance of the system in the Dutch-English parallel corpora a sample of 100 sentences are chosen randomly and manually annotated appropriately from the English and the Dutch text. The annotations are used to evaluate the projection recall and the named-entity recognition in the Dutch corpus, and to evaluate the named-entity recognition step on the English corpus. The metrics of each step of the system are then calculated in comparison with the manually annotated samples. Then the most common mistakes in each step are identified, reported and categorized.

4.1.1 Evaluation on Gold Standard Corpora

First the English named-entity recognition performance is reported in the following table with a micro averaged F1 score of approximately 0.86. All the metrics

are calculated without the use of the none label. The support represents the number of items per entity.

	Precision	Recall	F1-score	Support
B-PERSON	0.90	0.64	0.75	28
I-PERSON	1.00	0.40	0.57	5
B-LOC	0.94	0.84	0.89	37
I-LOC	0.75	0.60	0.67	5
B-ORG	0.90	0.94	0.92	70
I-ORG	0.73	0.89	0.80	18
micro avg	0.88	0.83	0.86	163
macro avg	0.87	0.72	0.77	163
weighted avg	0.89	0.83	0.85	163

Table 4.1: Evaluation of an existing Convolutional Neural Network (CNN) pretrained on OntoNotes corpus for the purpose of English named-entity Recognition on 100 manually annotated sentences from the English European Parliament corpus.

After the second phase, the projections are evaluated on the manually annotated sentences on the target language. The observed projection recall is 19% with a per entity recall for B-PERSON 55.56 %, I-PERSON 40 %, B-ORG 6.76 %, I-ORG 0 %, B-LOC 19 % and I-LOC 29 %. The observed projection precision is 76% with a per entity precision for B-PERSON 88.24 %, I-PERSON 100 %, B-ORG 45.45 %, I-ORG 0 %, B-LOC 100 % and I-LOC 100 %. The model extracted from the final phase is also evaluated in a gold standard corpus. For Dutch this is CoNLL while for German, this is German NER corpus.

	precision	recall	f1-score	support
B-PERSON	0.62	0.23	0.34	4,716
I-PERSON	0.71	0.33	0.45	2,883
B-LOC	0.59	0.34	0.43	3,208
I-LOC	0.61	0.09	0.16	467
B-ORG	0.28	0.15	0.20	2,082
I-ORG	0.33	0.11	0.17	1,199
micro avg	0.55	0.25	0.34	14,555
macro avg	0.52	0.21	0.29	14,555
weighted avg	0.56	0.25	0.34	14,555

Table 4.2: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted following the baseline approach on English-Dutch European Parliament corpus and evaluated on the Dutch Target Evaluation Dataset(CoNLL)

The results of English-German on the Conditional Random Field of the baseline are reported. The final evaluation on the German NER file consisting

of 26,200 sentences is reported on 4.3.

	precision	recall	f1-score	support
B-PERSON	0.69	0.39	0.50	8,390
I-PERSON	0.77	0.52	0.62	4,932
B-LOC	0.77	0.41	0.54	9,044
I-LOC	0.69	0.12	0.20	1,313
B-ORG	0.43	0.27	0.33	5,751
I-ORG	0.59	0.14	0.23	4,136
micro avg	0.67	0.35	0.46	33,566
macro avg	0.66	0.31	0.40	33,566
weighted avg	0.67	0.35	0.45	33,566

Table 4.3: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted following the baseline approach on English-German European Parliament corpus and evaluated on the German Target Evaluation Dataset(German NER)

4.1.2 Evaluation on Test Set

The data extracted from phase 2 (meaning after the projection step) are split in train and test datasets. The reason for reporting also the metrics on never before seen data of the same distribution on table 4.4 is to access the reason of the bad performance. As it can be seen the model performs better after adding Lexical features as it gives more capacity. The model is still slightly overfitting but the overall performance is increased.

	Precision	Recall	F1-score	Support
B-PERSON	0,74	0,70	0,72	161,306
I-PERSON	0,72	0,79	0,75	63,526
B-LOC	0,73	0,70	0,71	90,091
I-LOC	0,66	0,50	0,57	4,814
B-ORG	0,66	0,54	0,59	201,416
I-ORG	0,62	0,49	0,55	37,783
micro avg	0,70	0,63	0,67	558,936
macro avg	0,69	0,62	0,65	558,936
weigthed avg	0,70	0,63	0,66	558,936

Table 4.4: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted following the baseline approach on English-Dutch European Parliament corpus and evaluated on test split for the English-Dutch pair

	precision	recall	f1-score	support
B-PERSON	0.81	0.82	0.81	61,839
I-PERSON	0.76	0.88	0.82	14,368
B-LOC	0.85	0.92	0.88	137,050
I-LOC	0.79	0.80	0.80	2,219
B-ORG	0.82	0.84	0.83	251,783
I-ORG	0.77	0.96	0.85	53,539
micro avg	0.82	0.87	0.85	520,798
macro avg	0.80	0.87	0.83	520,798
weighted avg	0.82	0.87	0.85	520,798

Table 4.5: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted following the baseline approach on English-German European Parliament corpus and evaluated on test split for the English-German pair

4.2 Error Analysis

An error analysis of the baseline system is conducted to find out the cause of the errors in each step of the baseline. The errors are currently metered token wise. The errors for the named-entity recognition systems are categorized in three main categories: found entities, not found entities, found misclassified entities. For the projection part, errors are considered if a token of an entity is not found in the target language corpus.

The English named-entity recognition system was evaluated in 100 manually annotated sentences of the English corpus. What is observed is that most of the mistakes are due to the predictions including stopwords like 'the', 'Mrs', 'Madam', 'on', 'of' etc or the "s' as part of the entities. This is due to the dataset used for training the Convolutional neural network model (Ontonotes) which is considering those words as part of the entities. In some cases the entities were not found and in even less cases the entities were misclassified.

For the projection phase most of the errors observed were due to transliteration errors, for example when trying to project the English word "European" to the Dutch "Europese" exact string matching fails to identify it. That is the reason that similarity metrics could be used alternatively. Furthermore, there are cases that words are not identical in two languages, one example of that is the word 'Group' in English which is translated as 'Fractie' in Dutch. That is the reason that the intermediate step of translation proves to be important. Some of the errors were also because the two texts the English and the target were not identic*al*. This is due to the way the dataset is built as it is agendas from the European Parliament kept in every language . Finally, some errors were due to the extracted english entities. More specifically, the stopwords mentioned above or the "s' could not be projected.

The last part of the model, the trained named-entity recognition system on the target language was evaluated on part of the same dataset excluded from the

training of the named-entity recognition system. This is of course not the ground truth but it can be compared with the successful projections. Furthermore the system was also evaluated on the CoNLL dataset with worse results. This is assumed to be due to the different distribution of the datasets and the different way the annotations are made. That is the motivation to experiment with datasets with different distributions that the European Parliament Corpus to investigate if the system evaluated in the CoNLL works better. It is observed that in the CoNLL dataset most of the errors are due to acronyms not being identified or misclassified. Also in CoNLL dataset there are a lot of entities that are classified differently than the named-entity recognition system is predicting it, for example Antwerpse was classified as location while the ground truth was non label.

The most common errors that are observed in the 100 manually annotated examples:

- names that are not recognised, this was observed in 6 out of 7 one word surnames
- person names misidentified as different entities, for example as Organisation
- in one case Madam was identified as location
- Location entities were not identified in 6 cases (in total 79 tokens)
- Organisation entities were not identified in 4 cases, in the last two bullets it was the case that in previous sentences the same entity was identified (183 tokens in total)

4.3 Enhancements on English-Dutch Parallel Corpora

The next step is to calculate the recall of the projection and compare it with different preprocess and/or projection methods. First the recall of exact string matching is calculated in plain text data and by normalizing first the input of the text data. Next the recall of the projection using similarities metrics instead is reported.

The errors observed in the exact string matching were reported ignoring the errors incorporated from the English named-entity recognition step. As it was expected, except of names and cities/countries that are exactly the same in both languages the rest of the entities was not possible to project. By using the normalization of removing accents and lower casing some countries that were using accents, for example België and Belgium was recognised but there was an increased error due to multiple Dutch words recognised as entities. In general this approach seem to be the worst of all as most of the entities were not recognised and the ones that were found were not correct.

Exact String Matching without any normalisation	26.62
Removal of accents and lowercasing	-16,88
Soundex	-1.3
Stemming	-14.28
Removal of accents, lowercasing and removal of double letters	-18.18
Consonant Signature	-9.74
Stemming and then consonant signature	-5.19

Table 4.6: Recall comparison on projection step between exact string matching with and without normalisation on 100 manually annotated sentences from the English-Dutch European Parliament corpus

By using the soundex of words the same behaviour is observed. Most words of the Dutch corpora are identified wrongly as entities. In this case there are only 3 examples of correctly identified entities. When using the stemming approach a similar behaviour is observed with only two named entities to be recognised correctly. When using the normalised form of the words, meaning that the accents and the doubles are removed and all words are lowercased there is none entity recognised. When the consonant signature of the word is used for the projection, it is observed that only two named entities are recognised correctly and there are a lot of misidentified entities from the Dutch corpora. Finally, when first stemming is applied before extract the consonant signature a total of only two correct named entities are identified with a lot of Dutch words being wrongly identified as named entities.

Exact String Matching without any normalisation	26.62
Levenshtein metric	+47.41
soft TF-IDF with Levenshtein metric	+46.76
soft TF-IDF with Jaro metric	+38.32
soft TF-IDF with Jaro-Winkler metric	+50.65

Table 4.7: Comparison of Recall of projection step using similarity functions on 100 manually annotated sentences from the English-Dutch European Parliament corpus

By using similarity metrics instead of exact string matching as it is observed the recall is increased. In general in Levenshtein metric more entities were recognised with only two false recognised entities from the Dutch corpora. In soft TF-IDF with Levenshtein less named entities were recognised but also there was no misclassification. In soft TF-IDF with Jaro, less entities were identified in comparison with the other options and there are again two misclassifications. By using the third similarity option there was none false recognised entities from the Dutch corpora with almost the same recall.

An alternative that is explored to increase the performance of the model is to put an intermediate step between entity recognition and projection. The

alternative that is explored is to use look up tables with the countries and cities and its different names that may exist in Dutch or organisations (European Union-Europese Unie) or common words(Commission-Commissie) that are used and it is not possible to identify just by projecting them. Of course the list is not complete but it was explored as a way to increase the performance.

Exact string Matching without any normalisation	31.82
Removal of accents and lowercasing	-18.83
Soundex	-7.14
Stemming	-18.83
Removal of accents, lowercasing and removal of double letters	-20.78
Consonant Signature	-14.94
Stemming and then consonant signature	-10.39

Table 4.8: Projection Recall comparison using exact string matching with and without normalisation after using a look-up table on 100 manually annotated sentences from the English-Dutch European Parliament corpus

For the first case a slight increase in the performance was observed because of the look up table. Mostly countries that were not possible to be identified without the lookup table. For the soundex approach more dutch words seemed to be recognised as entities even if they are not. For the rest of the approaches the same apply, almost all the recognised entities are not the correct entities and there are slight increase in the performance between this and the previous approach that was not using the look up table.

There is a slight increase in this approach and it had to do with the countries that was not possible to be identified even by using the similarity metrics, one example is: Oostenrijk which is the Austria in Dutch. By using the look up table it was possible to find this kind of entities.

Exact string Matching without any normalisation	31.82
Levenshtein metric	+50.65
soft TF-IDF with Levenshtein metric	+50
soft TF-IDF with Jaro metric	+51.3
soft TF-IDF with Jaro-Winkler metric	+52.6

Table 4.9: Comparison of recall of projection part using similarity functions after using a look-up table on 100 manually annotated sentences from the English-Dutch European Parliament corpus

Finally an overview of the performance on the annotated corpora is reported. Firstly the macro performance is calculated with comparison with the manual annotations in each step. Then by considering the output of the previous step in the pipeline as correct annotations the performance of the step is evaluated. The reason for the latter is to examine the errors incorporated in every step of the pipeline and to also have an overview of the total error. Following is the

table comparing every step with the manually annotated sentences.

	Precision	Recall	F1-score
English named-entity Recognition	0.87	0.72	0.77
Exact String Matching Projection	0.76	0.19	0.30
Projection using Similarity metrics	0.84	0.62	0.71
After adding the translation step (look-up table)			
Exact String Matching Projection	0.83	0.62	0.71
Projection using Similarity metrics	0.86	0.70	0.77

Table 4.10: Evaluation and comparison of the source language named-entity recognition, projection approaches before and after using a look-up table on 100 manually annotated sentences from the English-Dutch European Parliament corpus.

For each of the steps the best performing method was applied to find out the best possible performance of the pipeline. For example, exact string matching without any normalisation is applied in the first step, and for the projection using similarity metrics the soft TF-IDF with Jaro-Winlker distance metric is reported. Following is the table with the comparisons by considering the output of the previous step in the pipeline as correct annotations evaluated on target language.

	Precision	Recall	F1-score
English named-entity Recognition	0.87	0.72	0.77
Exact String Matching Projection	0.85	0.26	0.4
Projection using Similarity metrics	0.30	0.74	0.43
After adding the translation step (look-up table)			
Exact String Matching Projection	0.85	0.78	0.81
Projection using Similarity metrics	0.85	0.85	0.85

Table 4.11: Evaluation and comparison of the source language named-entity recognition, projection approaches before and after using a look-up table by considering the output of the previous step in the pipeline as correct annotations on 100 manually annotated sentences from the English-Dutch European Parliament corpus

Subsequently, the per entity tag performance of the model for each step is presented. Firstly the performance on the manually annotated data is compared with the performance when considering the output of the previous step in the pipeline as correct annotations. In both cases the projected entites are compared with the manual annotations in the target language. Again the best performing approaches for each step are reported, meaning exact string matching without any normalisation, exact string matching without any normalisation after using a in between step of a look-up table with the most usual entities and finally the soft TF-IDF with Jaro-Winkler distance metric.

Table 4.12: Evaluation and comparison of the source language named-entity recognition, projection approaches before and after using a look-up table between manual annotations and the projections by considering the output of the previous step in the pipeline as correct annotations

Per entity Precision/Recall-Evaluation on 100 manually annotated sentences on target language from the English-Dutch European Parliament corpus						
	B-ORG	I-ORG	B-PERSON	I-PERSON	B-LOC	I-LOC
Exact String matching	45.45/6.76	0.0/0.0	88.24/55.56	100/40	100/19	100/29
Exact String match after look-up table	77.27/62.96	71.43/31.25	88.24/55.56	100/40	96.15/67.57	100/29
soft TF-IDF after look-up table	85.07/75	73.33/61.11	88.24/55.56	100/40	89.3/67.57	100/29
Per entity Precision/Recall-Comparison with the predictions by considering the output of the previous step in the pipeline as correct annotations						
	B-ORG	I-ORG	B-PERSON	I-PERSON	B-LOC	I-LOC
Exact String matching	63.64/15.07	100/9.09	94.12/85	100/100	85.71/21.21	100/50
Exact String matching after look-up table	77.27/90.41	100/31.82	94.12/85	100/100	96.15/78.79	100/50
soft TF-IDF after look-up	76.12/91.78	100/68.18	94.12/85	100/100	89.3/84.85	100/50

The performance of the model on the CoNLL is increased. The reason that the latter is performing better is due to the use of the look-up tables and the similarity metrics which increase the number of projected labels and as a result the training data that are given to the Conditional Random Field are of better quality. The recall observed after this step is approximately 63.67 %, while the entity recall is for B-PERSON 85.7%, for I-PERSON 64.9%, for B-ORG 67.1%, for I-ORG 38.5%, for B-LOC 75.1% and for I-LOC 22.5%.

As it is observed in table 4.13 there is a slight increase in the performance of the "Translation & Similarity metrics projection" model on the target evaluation dataset but still the model hardly achieves a 50% F1-score. That is the main reason for investigating if Lexical features for the Conditional Random Field or more data from different domains will increase the performance of the model.

4.3.1 Basic features

In this section the experiments using basic features are presented. As basic features are defined if the word is lowercase, uppercase, is the title or is a digit

	precision	recall	f1-score	support
B-PERSON	0.76	0.47	0.58	4,716
I-PERSON	0.87	0.61	0.72	2,883
B-LOC	0.74	0.37	0.50	3,208
I-LOC	0.74	0.15	0.24	467
B-ORG	0.34	0.24	0.28	2,082
I-ORG	0.50	0.22	0.30	1,199
micro avg	0.69	0.41	0.52	14,555
macro avg	0.66	0.34	0.44	14,555
weighted avg	0.69	0.41	0.51	14,555

Table 4.13: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted following the "Translation & Similarity metrics projection" approach on English-Dutch European Parliament corpus and evaluated on target evaluation dataset (Dutch CoNLL)

and the two previous characters.

Baseline after adding multidisciplinary datasets

In this section the results of adding more data while keeping the same basic features of the Conditional Random Field of the baseline are reported. The achieved recall is 32.91 % while the per entity recall: for B-PERSON is 60.66 %, for I-PERSON is 42.18 %, for B-ORG is 26.81 %, for I-ORG is 15.12 %, for B-LOC is 32.20 % and for I-LOC is 17.32 %. The finally projected sentences after only keeping sentences containing at least one label are 805,032 where 483,019 are used for training the model and 322,012 for testing it. Additionally, a sample of 6,000 sentences are randomly chosen from the training data to evaluate the model on the before seen data. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.14.

The data extracted from phase 2(meaning after the projection step) are split in train and test datasets. The reason for reporting also the metrics on never before seen data of the same distribution on table 4.15 is to access the reason of the bad performance. As it can be seen the model does not work that well on the never before seen data with the same distribution. That means that the model slightly overfits and approaches to overpass this should be investigated.

"Translation & Similarity metrics projection" model after adding multidisciplinary datasets

In this section the results of adding more data while keeping the same basic features of the Conditional Random Field of the "Translation & Similarity metrics projection" model are reported. The achieved recall is 54.94 % while the per entity recall: for B-PERSON is 62.50 %, for I-PERSON is 44.29 %, for B-ORG is 60.98 %, for I-ORG is 35.67 %, for B-LOC is 62.45 % and for I-LOC is 24.82

	precision	recall	f1-score	support
B-PERSON	0.78	0.54	0.64	4,716
I-PERSON	0.89	0.65	0.75	2,883
B-LOC	0.63	0.16	0.26	3,208
I-LOC	0.83	0.10	0.18	467
B-ORG	0.37	0.24	0.29	2,082
I-ORG	0.46	0.15	0.23	1,199
micro avg	0.71	0.39	0.50	14,555
macro avg	0.66	0.31	0.39	14,555
weighted avg	0.69	0.39	0.48	14,555

Table 4.14: Target Language Named-Entity Recognition evaluation. CRF model with basic features trained on data extracted with the baseline pipeline after adding multidisciplinary datasets on target evaluation dataset Dutch CoNLL

	precision	recall	f1-score	support
B-PERSON	0.50	0.48	0.49	256,760
I-PERSON	0.49	0.69	0.57	112,295
B-LOC	0.56	0.47	0.51	177,069
I-LOC	0.55	0.45	0.49	29,294
B-ORG	0.46	0.39	0.42	197,083
I-ORG	0.52	0.60	0.56	69,340
micro avg	0.50	0.49	0.50	841,841
macro avg	0.51	0.51	0.51	841,841
weighted avg	0.50	0.49	0.49	841,841

Table 4.15: Target Language Named-Entity Recognition evaluation. CRF model with basic features trained on data extracted with the baseline pipeline after adding multidisciplinary datasets on a test split for the English - Dutch corpus

%. The finally projected sentences after only keeping sentences containing at least one label are 1,299,374 where 779,624 are used for training the model and 519,749 for testing it. Additionally, a sample of 6000 sentences are randomly chosen from the training data to evaluate the model on the before seen data. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.16.

The data extracted from phase 2(meaning after the projection step) are split in train and test datasets. The reason for reporting also the metrics on never before seen data of the same distribution on table 4.17 is to access the reason of the bad performance. As it can be seen the model performs well on the samples of the training data but it does not work that well on never before seen data with the same distribution. That means that the model slightly overfits and approaches to overpass this should be investigated.

	precision	recall	f1-score	support
B-PERSON	0.78	0.48	0.59	4,716
I-PERSON	0.88	0.60	0.71	2,883
B-LOC	0.74	0.39	0.51	3,208
I-LOC	0.83	0.15	0.26	467
B-ORG	0.35	0.23	0.27	2,082
I-ORG	0.50	0.21	0.29	1,199
micro avg	0.71	0.41	0.52	14,555
macro avg	0.68	0.34	0.44	14,555
weighted avg	0.71	0.41	0.52	14,555

Table 4.16: Target Language Named-Entity Recognition evaluation. CRF model with basic features trained on data extracted with the "Translation & Similarity metrics projection" model after adding multidisciplinary datasets on target evaluation dataset (Dutch CoNLL)

	precision	recall	f1-score	support
B-PERSON	0.45	0.44	0.45	375,516
I-PERSON	0.44	0.68	0.53	168,764
B-LOC	0.53	0.51	0.52	360,982
I-LOC	0.43	0.53	0.47	40,711
B-ORG	0.44	0.30	0.36	346,965
I-ORG	0.37	0.42	0.39	99,633
micro avg	0.46	0.46	0.46	1,392,571
macro avg	0.44	0.48	0.45	1,392,571
weighted avg	0.46	0.46	0.45	1,392,571

Table 4.17: Target Language Named-Entity Recognition evaluation. CRF model with basic features trained on data extracted with the "Translation & Similarity metrics projection" model after adding multidisciplinary datasets on test split

4.3.2 Lexical features

In this section the experiments using Lexical features are presented. The basic features are if the word is lowercase, uppercase, is the title of the text or is a digit and the two previous characters. As additional features are defined if the first character is capital, if it is alphanumeric, if the word has a hyphen, if it is a first or last word, and the previous and the next two words.

Baseline after adding multidisciplinary datasets

In this section the results of adding more data while adding Lexical features as explained in the previous chapter of the Conditional Random Field of the baseline model are reported. The achieved recall is 30.80 % while the per entity recall: for B-PERSON is 43.48 %, for I-PERSON is 34.80 %, for B-ORG is

29.21 %, for I-ORG is 18.53 %, for B-LOC is 28.40 % and for I-LOC is 12.52 %. The finally projected sentences after only keeping sentences containing at least one label are 10,112 where 6067 are used for training the model and 4,044 for testing it. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.19. Additionally, the Wikipedia already annotated corpus is used for evaluating the model as shown on table 4.18 and compare it with the CoNLL.

	precision	recall	f1-score	support
B-PERSON	0.65	0.32	0.43	86,370
I-PERSON	0.88	0.30	0.44	58,706
B-LOC	0.77	0.18	0.29	162,188
I-LOC	0.85	0.04	0.07	36,407
B-ORG	0.27	0.24	0.25	34,463
I-ORG	0.39	0.11	0.17	21,161
micro avg	0.62	0.21	0.32	399,295
macro avg	0.63	0.20	0.28	399,295
weighted avg	0.70	0.21	0.31	399,295

Table 4.18: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets on WikiNER dataset

	precision	recall	f1-score	support
B-PERSON	0.77	0.36	0.49	4,716
I-PERSON	0.91	0.43	0.58	2,883
B-LOC	0.70	0.13	0.22	3,208
I-LOC	0.93	0.06	0.11	467
B-ORG	0.28	0.14	0.18	2,082
I-ORG	0.42	0.08	0.14	1,199
micro avg	0.69	0.26	0.38	14,555
macro avg	0.67	0.20	0.29	14,555
weighted avg	0.69	0.26	0.36	14,555

Table 4.19: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets on CoNLL dataset

The data extracted from phase 2(meaning after the projection step) are split in train and test datasets. The reason for reporting also the metrics on never before seen data of the same distribution on table 4.20 is to access the reason of the bad performance. As it can be seen the model performs better after adding Lexical features as it gives more capacity. The model is still slightly overfitting but the overall performance is increased.

	precision	recall	f1-score	support
B-PERSON	0.81	0.74	0.77	2,264
I-PERSON	0.78	0.83	0.80	995
B-LOC	0.81	0.72	0.76	1,622
I-LOC	0.80	0.73	0.77	146
B-ORG	0.81	0.67	0.73	1,927
I-ORG	0.78	0.74	0.75	506
micro avg	0.80	0.73	0.76	7,460
macro avg	0.80	0.74	0.77	7,460
weighted avg	0.80	0.73	0.76	7,460

Table 4.20: Target Language Named-Entity Recognition evaluation. CRF model with basic lexical trained on data extracted with the baseline model after adding multidisciplinary datasets on test split

"Translation & Similarity metrics projection" model after adding multidisciplinary datasets

In this section the results of adding more data while adding Lexical features as explained in the previous chapter using Conditional Random Field and with the pipeline of the second model are reported. The achieved overall recall is 54.94 % while the per entity recall: for B-PERSON is 43.48 %, for I-PERSON is 34.80 %, for B-ORG is 29.21 %, for I-ORG is 18.53 %, for B-LOC is 28.40 % and for I-LOC is 12.52 %. The finally projected sentences after only keeping sentences containing at least one label are 1,299,374 where 779,624 are used for training the model and 519,749 for testing it. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.21.

	precision	recall	f1-score	support
B-PERSON	0.78	0.48	0.59	4,716
I-PERSON	0.88	0.60	0.71	2,883
B-LOC	0.74	0.39	0.51	3,208
I-LOC	0.83	0.15	0.26	467
B-ORG	0.35	0.23	0.27	2,082
I-ORG	0.50	0.21	0.29	1,199
micro avg	0.71	0.41	0.52	14,555
macro avg	0.68	0.34	0.44	14,555
weighted avg	0.71	0.41	0.52	14,555

Table 4.21: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the "Translation & Similarity metrics projection" model after adding multidisciplinary datasets evaluated on Dutch CoNLL dataset

The data extracted from phase 2(meaning after the projection step) are split in train and test datasets. The reason for reporting also the metrics on never

before seen data of the same distribution on table 4.22 is to access the reason of the bad performance. As it can be seen the model performs better after adding Lexical features as it gives more capacity. The model is still slightly overfitting but the overall performance is increased.

	precision	recall	f1-score	support
B-PERSON	0.45	0.44	0.45	375,516
I-PERSON	0.44	0.68	0.53	168,764
B-LOC	0.53	0.51	0.52	360,982
I-LOC	0.43	0.53	0.47	40,711
B-ORG	0.44	0.30	0.36	346,965
I-ORG	0.37	0.42	0.39	99,633
micro avg	0.46	0.46	0.46	1,392,571
macro avg	0.44	0.48	0.45	1,392,571
weighted avg	0.46	0.46	0.45	1,392,571

Table 4.22: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the "Translation & Similarity metrics projection" model after adding multidisciplinary datasets on split test

4.3.3 Evaluation Study

Next the performance of each added dataset is evaluated on the baseline model in order to report the impact that each of the datasets had on the model. The goal to that is to access if a specific dataset from the ones used actually increases by a important percentage the performance on CoNLL dataset.

With all datasets

In this section the results of using all available datasets are reported. The achieved overall recall is 32.43 % while the per entity recall: for B-PERSON is 47.72 %, for I-PERSON is 31.22 %, for B-ORG is 31.18 %, for I-ORG is 20.14 %, for B-LOC is 28.71 % and for I-LOC is 11.64 %. The finally projected sentences after only keeping sentences containing at least one label are 1126 where 675 are used for training the model and 450 for testing it. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.25. Additionally, the Wikipedia already annotated corpus is used for evaluating the model as shown on table 4.23 and compare it with the CoNLL.

All datasets except European Parliament Proceeding Corpus

In this section the results of using all available datasets except the European Parliament Proceeding Corpus are reported. The achieved overall recall is 32.43 % while the per entity recall: for B-PERSON is 47.72 %, for I-PERSON is 31.22

	precision	recall	f1-score	support
B-PERSON	0.65	0.31	0.42	86,370
I-PERSON	0.87	0.29	0.43	58,706
B-LOC	0.77	0.16	0.26	162,188
I-LOC	0.83	0.03	0.07	36,407
B-ORG	0.26	0.22	0.24	34,463
I-ORG	0.40	0.10	0.16	21,161
micro avg	0.62	0.20	0.30	399,295
macro avg	0.63	0.19	0.26	399,295
weighted avg	0.70	0.20	0.30	399,295

Table 4.23: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets evaluated on WikiNER dataset

	precision	recall	f1-score	support
B-PERSON	0.75	0.32	0.45	4,716
I-PERSON	0.91	0.39	0.54	2,883
B-LOC	0.71	0.11	0.19	3,208
I-LOC	0.71	0.05	0.10	467
B-ORG	0.31	0.13	0.18	2,082
I-ORG	0.40	0.07	0.12	1,199
micro avg	0.69	0.23	0.35	14,555
macro avg	0.63	0.18	0.26	14,555
weighted avg	0.68	0.23	0.33	14,555

Table 4.24: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets evaluated on the CoNLL dataset

%, for B-ORG is 31.18 %, for I-ORG is 20.14 %, for B-LOC is 28.71 % and for I-LOC is 11.64 %. The finally projected sentences after only keeping sentences containing at least one label are 1126 where 675 are used for training the model and 450 for testing it. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.26. Additionally, the Wikipedia already annotated corpus is used for evaluating the model as shown on table 4.25 and compare it with the CoNLL.

All datasets except a bigger copy of the Tatoeba dataset

In this section the results of using all available datasets except a bigger copy of the Tatoeba dataset are reported. The achieved overall recall is 32.43 % while the per entity recall: for B-PERSON is 47.72 %, for I-PERSON is 31.22 %, for B-ORG is 31.18 %, for I-ORG is 20.14 %, for B-LOC is 28.71 % and for I-LOC is 11.64 %. The finally projected sentences after only keeping sentences containing

	precision	recall	f1-score	support
B-PERSON	0.64	0.31	0.42	86,370
I-PERSON	0.87	0.29	0.43	58,706
B-LOC	0.77	0.17	0.28	162,188
I-LOC	0.83	0.04	0.07	36,407
B-ORG	0.27	0.23	0.25	34,463
I-ORG	0.42	0.11	0.17	21,161
micro avg	0.62	0.21	0.31	399,295
macro avg	0.63	0.19	0.27	399,295
weighted avg	0.70	0.21	0.30	399,295

Table 4.25: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except European Parliament Proceeding Corpus evaluated on WikiNER dataset

	precision	recall	f1-score	support
B-PERSON	0.75	0.34	0.47	4,716
I-PERSON	0.90	0.41	0.56	2,883
B-LOC	0.70	0.13	0.21	3,208
I-LOC	0.86	0.05	0.10	467
B-ORG	0.31	0.15	0.20	2,082
I-ORG	0.50	0.09	0.15	1,199
micro avg	0.69	0.25	0.37	14,555
macro avg	0.67	0.20	0.28	14,555
weighted avg	0.69	0.25	0.36	14,555

Table 4.26: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except European Parliament Proceeding Corpus evaluated on Dutch CoNLL dataset

at least one label are 1126 where 675 are used for training the model and 450 for testing it. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.28. Additionally, the Wikipedia already annotated corpus is used for evaluating the model as shown on table 4.27 and compare it with the CoNLL.

All datasets except Global Voices

In this section the results of using all available datasets except the Global Voices Corpus are reported. The achieved overall recall is 32.43 % while the per entity recall: for B-PERSON is 47.72 %, for I-PERSON is 31.22 %, for B-ORG is 31.18 %, for I-ORG is 20.14 %, for B-LOC is 28.71 % and for I-LOC is 11.64 %. The finally projected sentences after only keeping sentences containing at

	precision	recall	f1-score	support
B-PERSON	0.65	0.31	0.42	86,370
I-PERSON	0.87	0.28	0.43	58,706
B-LOC	0.76	0.17	0.27	162,188
I-LOC	0.86	0.04	0.07	36,407
B-ORG	0.27	0.23	0.25	34,463
I-ORG	0.40	0.10	0.17	21,161
micro avg	0.62	0.20	0.31	399,295
macro avg	0.64	0.19	0.27	399,295
weighted avg	0.70	0.20	0.30	399,295

Table 4.27: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except a bigger copy of the Tatoeba dataset evaluated on WikiNER dataset

	precision	recall	f1-score	support
B-PERSON	0.72	0.33	0.45	4,716
I-PERSON	0.90	0.41	0.56	2,883
B-LOC	0.68	0.12	0.20	3,208
I-LOC	0.79	0.05	0.09	467
B-ORG	0.27	0.12	0.17	2,082
I-ORG	0.40	0.07	0.11	1,199
micro avg	0.67	0.24	0.35	14,555
macro avg	0.63	0.18	0.27	14,555
weighted avg	0.66	0.24	0.34	14,555

Table 4.28: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except a bigger copy of the Tatoeba dataset evaluated on Dutch CoNLL dataset

least one label are 1,126 where 675 are used for training the model and 450 for testing it. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.30. Additionally, the Wikipedia already annotated corpus is used for evaluating the model as shown on table 4.29 and compare it with the CoNLL.

All datasets except the Wikipedia articles dataset

In this section the results of using all available datasets except the the Wikipedia articles dataset are reported. The achieved overall recall is 32.43 % while the per entity recall: for B-PERSON is 47.72 %, for I-PERSON is 31.22 %, for B-ORG is 31.18 %, for I-ORG is 20.14 %, for B-LOC is 28.71 % and for I-LOC is 11.64 %. The finally projected sentences after only keeping sentences containing at

	precision	recall	f1-score	support
B-PERSON	0.60	0.35	0.44	86,370
I-PERSON	0.84	0.32	0.47	58,706
B-LOC	0.72	0.15	0.25	162,188
I-LOC	0.89	0.04	0.08	36,407
B-ORG	0.26	0.22	0.24	34,463
I-ORG	0.37	0.10	0.16	21,161
micro avg	0.59	0.21	0.31	399,295
macro avg	0.61	0.20	0.27	399,295
weighted avg	0.67	0.21	0.30	399,295

Table 4.29: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except Global Voices evaluated on WikiNER dataset

	precision	recall	f1-score	support
B-PERSON	0.68	0.36	0.47	4,716
I-PERSON	0.88	0.41	0.56	2,883
B-LOC	0.60	0.13	0.21	3,208
I-LOC	0.78	0.06	0.12	467
B-ORG	0.21	0.15	0.18	2,082
I-ORG	0.41	0.12	0.19	1,199
micro avg	0.59	0.26	0.36	14,555
macro avg	0.60	0.21	0.29	14,555
weighted avg	0.62	0.26	0.35	14,555

Table 4.30: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except Global Voices evaluated on Dutch CoNLL dataset

least one label are 1126 where 675 are used for training the model and 450 for testing it. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.32. Additionally, the Wikipedia already annotated corpus is used for evaluating the model as shown on table 4.31 and compare it with the CoNLL.

All datasets except a smaller copy of the Tatoeba dataset

In this section the results of using all available datasets except a smaller copy of the Tatoeba dataset are reported. The achieved overall recall is 32.43 % while the per entity recall: for B-PERSON is 47.72 %, for I-PERSON is 31.22 %, for B-ORG is 31.18 %, for I-ORG is 20.14 %, for B-LOC is 28.71 % and for I-LOC is 11.64 %. The finally projected sentences after only keeping sentences containing

	precision	recall	f1-score	support
B-PERSON	0.64	0.31	0.42	86,370
I-PERSON	0.87	0.28	0.43	58,706
B-LOC	0.77	0.17	0.28	162,188
I-LOC	0.85	0.04	0.07	36,407
B-ORG	0.27	0.24	0.25	34,463
I-ORG	0.39	0.11	0.17	21,161
micro avg	0.62	0.21	0.31	399,295
macro avg	0.63	0.19	0.27	399,295
weighted avg	0.70	0.21	0.31	399,295

Table 4.31: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except the Wikipedia articles dataset evaluated on WikiNER dataset

	precision	recall	f1-score	support
B-PERSON	0.75	0.34	0.47	4,716
I-PERSON	0.91	0.41	0.57	2,883
B-LOC	0.68	0.13	0.21	3,208
I-LOC	0.76	0.06	0.10	467
B-ORG	0.27	0.14	0.18	2,082
I-ORG	0.42	0.07	0.12	1,199
micro avg	0.68	0.25	0.36	14,555
macro avg	0.63	0.19	0.28	14,555
weighted avg	0.67	0.25	0.35	14,555

Table 4.32: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except the Wikipedia articles dataset evaluated on Dutch CoNLL dataset

at least one label are 1126 where 675 are used for training the model and 450 for testing it. The final evaluation on the CoNLL file consisting of 15,806 sentences is reported on 4.34. Additionally, the Wikipedia already annotated corpus is used for evaluating the model as shown on table 4.33 and compare it with the CoNLL.

4.3.4 Stacked Embeddings with BiLSTM

For this architecture Stacked embeddings and a bidirectional Long Short Term Memory network with Conditional Random Field as decoding layer is used. The stacked embeddings include contextual string embeddings stacked with BERT embeddings. For the NER system a sequence tagger was used. Contextual string embeddings are powerful embeddings that capture latent syntactic-semantic in-

	precision	recall	f1-score	support
B-PERSON	0.65	0.32	0.43	86,370
I-PERSON	0.88	0.30	0.44	58,706
B-LOC	0.77	0.17	0.28	162,188
I-LOC	0.84	0.03	0.07	36,407
B-ORG	0.27	0.24	0.26	34,463
I-ORG	0.39	0.10	0.16	21,161
micro avg	0.62	0.21	0.32	399,295
macro avg	0.63	0.20	0.27	399,295
weighted avg	0.70	0.21	0.31	399,295

Table 4.33: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except the Tatoeba dataset evaluated on WikiNER dataset

	precision	recall	f1-score	support
B-PERSON	0.74	0.34	0.47	4,716
I-PERSON	0.91	0.41	0.56	2,883
B-LOC	0.69	0.13	0.22	3,208
I-LOC	0.84	0.06	0.10	467
B-ORG	0.28	0.14	0.19	2,082
I-ORG	0.42	0.08	0.13	1,199
micro avg	0.67	0.25	0.36	14,555
macro avg	0.65	0.19	0.28	14,555
weighted avg	0.67	0.25	0.35	14,555

Table 4.34: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model after adding multidisciplinary datasets except the Tatoeba dataset evaluated on CoNLL dataset

formation that goes beyond standard word embedding, they are trained on a mix of corpora (i.e. Web, Wikipedia, Subtitles, News) as explained on Akbik et *al.*, 2018b; M.Peters et *al.*, 2017; M.Peters et *al.*, 2018 [27, 26, 3]. BERT embeddings were developed by Devlin et *al.*, 2019 [6] and are based on a bidirectional transformer architecture. For the purpose of this experiment a pre-trained BERT model is used which is trained in 104 languages has 12-layer, 768-hidden, 12-heads and 110M parameters. The sequence labelling use a LSTM variant of bidirectional recurrent neural networks (BiLSTMs), and a subsequent conditional random field (CRF) decoding layer (Akbik et *al.*, 2018b; Huang et *al.*, 2015 [12, 3]).

After following the procedure used on the "Translation & Similarity metrics projection" model to create a training dataset, the model was trained and then evaluated on the CoNLL dataset. The overall recall obtained from the

	precision	recall	f1-score	support
B-PERSON	0.77	0.78	0.78	4,492
I-PERSON	0.89	0.77	0.82	2,868
B-LOC	0.82	0.49	0.62	2,991
I-LOC	0.71	0.18	0.28	467
B-ORG	0.42	0.46	0.44	2,082
I-ORG	0.48	0.31	0.38	1,199
micro avg	0.71	0.61	0.66	14,099
macro avg	0.68	0.50	0.55	14,099
weighted avg	0.72	0.61	0.65	14,099

Table 4.35: Target Language Named-Entity Recognition evaluation. Stacked Embeddings with BiLSTM Model on data extracted with the "Translation & Similarity metrics projection" model evaluated on the Target Evaluation Dataset (Dutch CoNLL)

"Translation & Similarity metrics projection" model Recall is: 63.91 %, while the per entity recall is for B-PERSON 86.53 %, for I-PERSON 60.31 %, for B-ORG 67.23 %, for I-ORG 39.17 %, for B-LOC 75.68% and for I-LOC 22.05%. The size of the training data is 4993 sentences, the size of the test dataset is 1664 while the CoNLL contains 15,806 sentences. On table **??** on page **??** are the results of the model and a comparison with the Conditional Random Field model on table 4.37 on the same training data. As it is observed this model outperforms the Conditional Random Field model.

	precision	recall	f1-score	support
B-PERSON	0.62	0.23	0.34	4,716
I-PERSON	0.71	0.33	0.45	2,883
B-LOC	0.59	0.34	0.43	3,208
I-LOC	0.61	0.09	0.16	467
B-ORG	0.28	0.15	0.20	2,082
I-ORG	0.33	0.11	0.17	1,199
micro avg	0.55	0.25	0.34	14,555
macro avg	0.52	0.21	0.29	14,555
weighted avg	0.56	0.25	0.34	14,555

Table 4.36: Target Language Named-Entity Recognition evaluation. CRF on data extracted with the "Translation & Similarity metrics projection" model evaluated on the Target Evaluation Dataset(CoNLL)

	precision	recall	f1-score	support
B-PERSON	0.62	0.78	0.69	133
I-PERSON	0.19	0.89	0.31	18
B-LOC	0.66	0.89	0.76	61
I-LOC	0.50	1.00	0.67	2
B-ORG	0.64	0.47	0.55	354
I-ORG	0.15	0.40	0.21	47
micro avg	0.50	0.59	0.54	615
macro avg	0.46	0.74	0.53	615
weighted avg	0.59	0.59	0.57	615

Table 4.37: Target Language Named-Entity Recognition evaluation. Stacked Embeddings with BiLSTM Model on data extracted with the "Translation & Similarity metrics projection" model evaluated on the Dutch Parliamentary Documents

4.4 Enhancements on English-German Parallel Corpora

To evaluate the performance of the system in the German-English parallel corpora, a sample of 100 sentences is chosen randomly and annotated appropriately from the English and the German text. The metrics of each step of the system were then calculated in comparison with the manual annotations as defined from the annotated samples. Then a grouping of the most common mistakes in each step was identified and reported.

First is the evaluation of the projection recall using exact string matching and the comparison of different preprocessing methods on table 4.38 and next the similarity metrics comparison follows on table 4.39. All the metrics are calculated without the use of the "none" label. The support represents the number of items per entity.

Exact String Matching	34.30
Removal of accents and lowercasing	-15.46
Soundex	-11.59
Stemming	-14.49
Removal of accents lowercasing and removal of double letters	-19.81
Consonant Signature	-12.08
Stemming and then consonant signature	-11.11

Table 4.38: Projection Recall using exact string matching with and without normalisation on 100 manually annotated sentences from the English-German European Parliament Corpus

The same evaluation metrics are reported after adding the translation step. For the German experiment a pre-trained system is used to translate each en-

Exact String Matching	34.30
Levenshtein metric	+22.22
soft TF-IDF with Levenshtein	+20.29
soft TF-IDF with Jaro	+31.4
soft TF-IDF with Jaro-Winkler	+38.16

Table 4.39: Projection Recall using similarity functions on 100 manually annotated sentences from English-German European Parliament Corpus

tity from English to German and then the recall of the projection using string matching with various pre-process options as well as the recall by using similarity functions is reported.

Exact String Matching	47.34
Removal of accents and lowercasing	-25.12
Soundex	-25.6
Stemming	-26.08
Removal of accents lowercasing and remove of double letters	-29.95
Consonant Signature	-27.05
Stemming and then consonant signature	-26.08

Table 4.40: Projection Recall using exact string matching with and without normalisation after using a translation model on 100 manually annotated sentences from the English-German European Parliament Corpus

Exact String Matching	47.34
Levenshtein metric	+3.87
soft TF-IDF with Levenshtein	+1.94
soft TF-IDF with Jaro	+14.98
soft TF-IDF with Jaro-Winkler	+17.88

Table 4.41: Projection Recall using similarity functions after using a translation system on 100 manually annotated sentences from the English-German European Parliament Corpus

4.4.1 Basic features

In this section the results of the basic features on the Conditional Random Field of the baseline are reported. The final evaluation on the German NER file consisting of 26,200 sentences is reported on 4.42.

The data extracted from phase 2(meaning after the projection step) are split in train and test datasets. The reason for reporting also the metrics on never before seen data of the same distribution on table 4.43 is to access the reason of the bad performance. As it can be seen the model performs better after adding

	precision	recall	f1-score	support
B-PERSON	0.69	0.39	0.50	8,390
I-PERSON	0.77	0.52	0.62	4,932
B-LOC	0.77	0.41	0.54	9,044
I-LOC	0.69	0.12	0.20	1,313
B-ORG	0.43	0.27	0.33	5,751
I-ORG	0.59	0.14	0.23	4,136
micro avg	0.67	0.35	0.46	33,566
macro avg	0.66	0.31	0.40	33,566
weighted avg	0.67	0.35	0.45	33,566

Table 4.42: Target Language Named-Entity Recognition evaluation. CRF model with basic features trained on data extracted with the baseline model as explained on the previous section evaluated on German NER

Lexical features as it gives more capacity. The model is still slightly overfitting but the overall performance is increased.

	precision	recall	f1-score	support
B-PERSON	0.81	0.82	0.81	61,839
I-PERSON	0.76	0.88	0.82	14,368
B-LOC	0.85	0.92	0.88	137,050
I-LOC	0.79	0.80	0.80	2,219
B-ORG	0.82	0.84	0.83	251,783
I-ORG	0.77	0.96	0.85	53,539
micro avg	0.82	0.87	0.85	520,798
macro avg	0.80	0.87	0.83	520,798
weighted avg	0.82	0.87	0.85	520,798

Table 4.43: Target Language Named-Entity Recognition evaluation. CRF model with basic features trained on data extracted with the baseline model as explained on the previous section evaluated evaluated on test split

4.4.2 Lexical features

In this section the results of adding Lexical features on the Conditional Random Field of the baseline are reported. The achieved recall is 57.42 % while the per entity recall: for B-PERSON is 88.41 %, for I-PERSON is 69.18 %, for B-ORG is 58.17 %, for I-ORG is 33.18 %, for B-LOC is 70.96 % and for I-LOC is 7.69 %. The finally projected sentences after only keeping sentences containing at least one label are 767,839 where 460,703 are used for training the model and 307,135 for testing it. Additionally, a sample of 6000 sentences are randomly chosen from the training data to evaluate the model on the before seen data. The final evaluation on the German NER file consisting of 26,200 sentences is

reported on 4.44.

	precision	recall	f1-score	support
B-PERSON	0.64	0.40	0.49	8,390
I-PERSON	0.75	0.50	0.60	4,932
B-LOC	0.74	0.42	0.54	9,044
I-LOC	0.60	0.11	0.19	1,313
B-ORG	0.35	0.28	0.31	5,751
I-ORG	0.52	0.16	0.24	4,136
micro avg	0.61	0.36	0.45	33,566
macro avg	0.60	0.31	0.40	33,566
weighted avg	0.62	0.36	0.45	33,566

Table 4.44: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model as explained on the previous section evaluated on German NER dataset

The data extracted from phase 2(meaning after the projection step) are split in train and test datasets. The reason for reporting also the metrics on never before seen data of the same distribution on table 4.45 as aforementioned is to access the reason of the bad performance. The model is slightly overfitting but the overall performance is increased.

	precision	recall	f1-score	support
B-PERSON	0.76	0.70	0.73	61,452
I-PERSON	0.72	0.79	0.75	14,196
B-LOC	0.83	0.89	0.86	131,005
I-LOC	0.72	0.68	0.70	1,832
B-ORG	0.80	0.82	0.81	257,987
I-ORG	0.75	0.92	0.83	54,606
micro avg	0.80	0.83	0.81	521,078
macro avg	0.76	0.80	0.78	521,078
weighted avg	0.80	0.83	0.81	521,078

Table 4.45: Target Language Named-Entity Recognition evaluation. CRF model with lexical features trained on data extracted with the baseline model as explained on the previous section evaluated on test split

4.4.3 Comparison between model trained on data from the pipeline and on manually annotated data

In this section the results of the "Translation & Similarity metrics projection" model using the Conditional Random Field are reported on manually annotated data. The finally projected sentences after only keeping sentences containing at least one label are 767,839 where 3,538 are used for training the model and 858

are left out for testing it. The final evaluation on the 90 sentences is reported on 4.46. The reason for this experiment is to prove that due to different distributions of the two datasets it is not possible to have better performance with the Conditional Random Field model even if there exist manually annotated data.

	precision	recall	f1-score	support
B-PERSON	1.00	0.11	0.19	19
I-PERSON	1.00	0.50	0.67	4
B-LOC	0.92	0.73	0.81	15
B-ORG	0.33	0.33	0.33	3
I-ORG	0.00	0.00	0.00	1
micro avg	0.84	0.38	0.52	42
macro avg	0.65	0.33	0.40	42
weighted avg	0.90	0.38	0.46	42

Table 4.46: Target Language Named-Entity Recognition evaluation. CRF model trained on manually annotated data evaluated on target language evaluation dataset (German CoNLL)

	precision	recall	f1-score	support
B-PERSON	0.00	0.00	0.00	19
I-PERSON	0.00	0.00	0.00	4
B-LOC	0.92	0.73	0.81	15
B-ORG	1.00	0.33	0.50	3
I-ORG	1.00	1.00	1.00	1
micro avg	0.87	0.31	0.46	42
macro avg	0.58	0.41	0.46	42
weighted avg	0.42	0.31	0.35	42

Table 4.47: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted with the "Translation & Similarity metrics projection" model as explained on the previous section evaluated on target language evaluation dataset (German CoNLL)

4.4.4 Stacked Embeddings with BiLSTM

For this architecture Stacked embeddings were used. Those include contextual string embeddings stacked with BERT embeddings. For the NER system a sequence tagger was used. Contextual string embeddings are powerful embeddings that capture latent syntactic-semantic information that goes beyond standard word embedding, they are trained on a mix of corpora (Web, Wikipedia, Subtitles, News) [Peters et *al.*, 2017; Peters et *al.*, 2018; Akbik et *al.*, 2018b [27, 26, 3]] BERT embeddings were developed by Devlin et *al.*, 2019 [6] and are based on a bidirectional transformer architecture. For the purpose of this experiment

a pre-trained BERT model is used which is trained in 104 languages has 12-layer, 768-hidden, 12-heads and 110M parameters. The sequence labelling use a LSTM variant of bidirectional recurrent neural networks (BiLSTMs), and a subsequent conditional random field (CRF) decoding layer (Huang et *al.*, 2015; Akbik et *al.*, 2018b [12, 3]).

After following the procedure used on the "Translation & Similarity metrics projection" model to create a training dataset, the flair model was trained and then evaluated on the CoNLL dataset. Following are the results of the model and a comparison with the CRF model. As it is observed the flair model outperforms the CRF.

	precision	recall	f1-score	support
B-PERSON	0.73	0.82	0.77	1,639
I-PERSON	0.81	0.75	0.78	912
B-LOC	0.79	0.43	0.55	1,706
I-LOC	0.47	0.03	0.05	303
B-ORG	0.33	0.30	0.32	1,150
I-ORG	0.57	0.14	0.23	698
micro avg	0.66	0.50	0.57	6,408
macro avg	0.62	0.41	0.45	6,408
weighted avg	0.66	0.50	0.54	6,408

Table 4.48: Stacked Embeddings with BiLSTM using the "Translation & Similarity metrics projection" model and the German English European corpus and evaluated on the target language evaluation dataset-GermaNER

After following the procedure used on the "Translation & Similarity metrics projection" model to create a training dataset, the model was trained and then evaluated on the German NER dataset. The overall recall obtained from the "Translation & Similarity metrics projection" model Recall is: 62.94 %, while the per entity recall is for B-PERSON 88.61 %, for I-PERSON 69.82 %, for B-ORG 73.46 %, for I-ORG 19.29 %, for B-LOC 73.24% and for I-LOC 10.44%. The size of the training data is 5,079 sentences, the size of the test dataset is 3,387.

4.5 Overall Results

In this section the overall results of both language models are presented. The English-Dutch models are firstly presented and then the English-German are presented.

4.5.1 Dutch Results

In this section the overall results for the Dutch-English model are presented. Firstly, the baseline model and the "Translation & Similarity metrics projection"

	precision	recall	f1-score	support
B-PERSON	0.47	0.12	0.19	1,696
I-PERSON	0.64	0.17	0.27	915
B-LOC	0.63	0.16	0.26	2,376
I-LOC	1.00	0.01	0.03	307
B-ORG	0.24	0.08	0.12	1,331
I-ORG	0.39	0.02	0.03	705
micro avg	0.49	0.12	0.19	7,330
macro avg	0.56	0.09	0.15	7,330
weighted avg	0.51	0.12	0.19	7,330

Table 4.49: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted with the "Translation & Similarity metrics projection" model evaluated on the target language evaluation dataset-GermaNER

model results are presented. The "Translation & Similarity metrics projection" model is defined as the model using a look-up table in between and similarity metrics. Both baseline and "Translation & Similarity metrics projection" model are using a Conditional Random Field for training a named-entity recognition model. Then a baseline and a "Translation & Similarity metrics projection" model with basic and additional features are compared. Finally the "Translation & Similarity metrics projection" Model that uses stacked embeddings with Bidirectional Long Short Term Memory.

	Precision	Recall	F1-score
Baseline	0.67	0.38	0.49
"Translation & Similarity metrics projection" Model	0.69	0.41	0.52
Baseline after adding multidisciplinary datasets with basic features	0.71	0.39	0.50
"Translation & Similarity metrics projection" Model after adding multidisciplinary datasets with basic features	0.71	0.41	0.52
Baseline after adding multidisciplinary datasets with Lexical features	0.69	0.26	0.38
"Translation & Similarity metrics projection" Model after adding multidisciplinary datasets with Lexical features	0.71	0.41	0.52
"Translation & Similarity metrics projection" Model with Stacked Embeddings with BiLSTM	0.71	0.61	0.66

Table 4.50: Overall evaluation of the English-Dutch models on CoNLL dataset

The next table shows the result by removing additional datasets. In the first results the European Parliament Proceeding Corpus part is removed, in the "Translation & Similarity metrics projection" results a bigger copy of the Tatoeba is removed, in the third results Global Voiced is removed, in the fourth

results Wikipedia is removed, in the fifth results Tatoeba is removed.

	Precision	Recall	F1-score
Except European Parliament Proceeding Corpus	0.69	0.25	0.37
Except a bigger copy of the Tatoeba	0.62	0.20	0.31
Except Global Voices	0.59	0.21	0.31
Except the Wikipedia	0.68	0.25	0.36
Except the Tatoeba	0.67	0.25	0.36

Table 4.51: Baseline after adding multidisciplinary datasets with basic features and evaluation on CoNLL dataset

4.5.2 German Results

In this section the overall results for the German-English model are presented. Firstly, the baseline model with basic features and added features is compared. Finally the "Translation & Similarity metrics projection" model with stacked embeddings with bidirectional Long Short Term Memory is reported.

	Precision	Recall	F1-score
Baseline with basic features	0.67	0.35	0.46
Baseline with Lexical features	0.61	0.36	0.45
"Translation & Similarity metrics projection" Model vs model trained on manually annotated data	0.84	0.38	0.52
"Translation & Similarity metrics projection" Model with Stacked Embeddings with BiLSTM	0.66	0.50	0.57

Table 4.52: Overall evaluation of the English-German models on German NER dataset

The next table shows the result by removing additional datasets. In the first results the European Parliament Proceeding Corpus part is removed, in the "Translation & Similarity metrics projection" results a bigger copy of the Tatoeba is removed, in the third results Global Voiced is removed, in the fourth results Wikipedia is removed, in the fifth results Tatoeba is removed.

	Precision	Recall	F1-score
Except European Parliament Proceeding Corpus	0.69	0.25	0.37
Except a bigger copy of the Tatoeba	0.62	0.20	0.31
Except Global Voices	0.59	0.21	0.31
Except the Wikipedia	0.68	0.25	0.36
Except the Tatoeba	0.67	0.25	0.36

Table 4.53: Baseline after adding multidisciplinary datasets with basic features and evaluation on German NER dataset

Chapter 5

Conclusion

In this thesis, we present multiple models to automatically create annotated training data in other languages than English. Multiple projection methods are tested and preprocessing and normalization methods before the projection. Furthermore, the option of using a translation model or a look-up table is explored. For the named-entity recognition a Conditional Random Field model and a Bi-directional Long Short Term Memory Neural Network are evaluated using the training data extracted from the previous step of the model. For the Conditional Random Field model two set of features are reported. The research questions that were presented on the introduction are replied.

Which models perform better when it comes to the automatic creation of annotated datasets for other languages?

The model that used a look-up table and similarity metrics proved to be able to identify and project more named entities with a recall of 63.91% of identified entities and as a result to create more complete and better quality training data. This is expected as the words/named-entities are easier to project between two different languages due to not checking every character of the two words in source and target language but to evaluate how similar these two words are. The Bi-directional Long Short Term Memory Neural Network with Embeddings using the previously extracted training data outperformed the Conditional Random Field by achieving F1-score of 66% in comparison with the CRF that achieved 34% when trained in the same amount of data. The first is considered the state-of-art on named entity recognition with BERT embeddings being the state-of-art for embeddings and encoding a language and consequently it is expected to perform better than the latter even with training data of lower quality.

What is the quality of a classification task like named-entity recognition in such datasets compared to classification results of the original dataset?

The task of named-entity recognition model on the Dutch training data achieves a F1-score of 66% when the "Translation & Similarity metrics projection" is used to create the model and the Bi-directional Long Short Term Memory Neural Network with Embeddings model is used. The state-of-the art

in CoNLL achieves 90.44% based on Akbik et *al.*, 2019 [1]. For the German named-entity recognition the achieved F1-score is 57% while the state of the art is 88.27% evaluated on the German CoNLL. The probable cause for that is the different distributions between the data used for training the model and the data used for evaluating the model.

What is the exact quality improvement of such methods measured in different well established metrics (e.g. Accuracy, Recall, Precision etc)?

Those results are reported on the chapter Experiments. The "Translation & Similarity metrics projection" model proved to outperform the baseline meaning that using a look-up table and a soft TF-IDF based on Jaro-Winkler metric without normalization for the projection outperformed the rest of the similarity metrics and normalization approaches. The bidirectional LSTM outperformed the rest when trained on the training data extracted from the previously described step. This model worked the best for the Dutch - English corpus and achieved 71% precision, 61% recall with a F1 score of 66%.

What are limitations of the approach, what are the main causes of the final error and how could this be improved?

The main limitation of this approach is that it can not perform as well as the models trained on manually annotated corpora. The main reason for that is that each step of the procedure adds some amount of error in the final creation of the training dataset. For example in the baseline model errors are due to the errors incorporated on the English named-entity recognition system, the projection error (which in this case is the biggest amount of error) and the error from the trained CRF model. In the "Translation & Similarity metrics projection" the projection error is way smaller but there are errors incorporated from the look-up tables or the pre-trained translation model. The error from the pre-trained translation model is bigger than the one from the look up table. When using a BiLSTM instead of a CRF for the final step of training a named-entity recognition system the errors incorporated in this step become fewer. Finally other approaches and models could investigate to improve more the overall performance.

Which pre- and post-processing techniques can be applied to improve the performance?

To improve the performance of the model the removal of stop-words and words like "Mr" and "Mrs" from the named entities proved to be most beneficial post processing technique. This is due to be able to project the entities on the target language without having to deal with the translation of these words increasing like that the recall. For the pre-processing techniques the use of similarity metrics instead of string matching with any kind of normalization proved to outperform the rest.

Chapter 6

Future Work

The process of automatically creating annotated corpora in languages other than English is explored in this thesis. The assignment offers an overview of basic approaches to handle it as a problem but there are a lot of prospects to investigate and improvise on. This section gives a small overview of the possible approaches and differentiations of the explored methods.

Firstly, different pre-trained models could be used to extract the English named entities or even train a named-entity recognition system from scratch after first annotating the English part of the an aligned multilingual corpus. In the final step, different machine learning models for the named-entity recognition could be tested and different types of pre-trained embeddings. Furthermore, training of embeddings in the same corpus could be explored. Specifically for the CRF case multiple feature sets could be used and researched.

To increase the recall of identified named entities from English to a target language a more concise and complete look-up table could be created involving all identified entities for the English Corpus. Alternatively, a machine translation model trained on the same corpus could be used for this purpose. Moreover, diverse similarity metrics could be explored for the cause.

A non token based evaluation function could be used when evaluating how well the proposed system is working. In this assignment mainly the token based evaluation function is used, meaning that a correctly identified and classified entity is compared with the manually annotated entities by comparing the tokenswords belonging to a class.

Finally, the exploration of more aligned corpora and the extension of the experiments in other classification problems can result in interesting results.

Appendices

Appendix A

Evaluation of the Dutch-English model

The data extracted using the baseline model are split in to training data and test data. The target language named-entity recognition using a CRF is evaluated on randomly sampled examples with size 40% of the initial dataset that was left out as test set is reported in the next table.

	Precision	Recall	F1-score	Support
B-PERSON	0,74	0,70	0,72	161,306
I-PERSON	0,72	0,79	0,75	63,526
B-LOC	0,73	0,70	0,71	90,091
I-LOC	0,66	0,50	0,57	4,814
B-ORG	0,66	0,54	0,59	201,416
I-ORG	0,62	0,49	0,55	37,783
micro avg	0,70	0,63	0,67	558,936
macro avg	0,69	0,62	0,65	558,936
weigthed avg	0,70	0,63	0,66	558,936

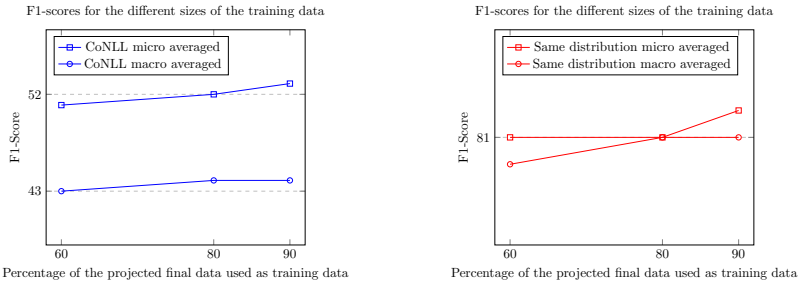
Table A.1: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the baseline approach on 60% of English-Dutch European Parliament corpus and evaluated on the 40% split used as test set

The performance of the "Translation & Similarity metrics projection" model on randomly sampled examples on 10% of the initial dataset that was left out as test set is reported in the next table. The size of this dataset is 149,321 sentences.

	precision	recall	f1-score	support
B-PERSON	0.78	0.80	0.79	47,058
I-PERSON	0.76	0.84	0.80	13,163
B-LOC	0.79	0.86	0.83	76,279
I-LOC	0.78	0.87	0.82	5,195
B-ORG	0.78	0.82	0.80	179,323
I-ORG	0.75	0.91	0.82	44,887
micro avg	0.78	0.84	0.81	365,905
macro avg	0.77	0.85	0.81	365,905
weighted avg	0.78	0.84	0.81	365,905

Table A.2: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 90% of English-Dutch European Parliament corpus and evaluated on the 10% split used as test set

A.1 Results with different sizes of training data



The results of CRF with basic features using different sizes of training data is reported. The initial dataset in source language is 1,997,775 sentences the same as the target parallel. The evaluation dataset is 15,806 sentences. After the preprocessing and the projection 985,014 sentences are exported. Only sentences with at least one named-entity are kept, that correspond to 753,679 sentences and are used for the training and testing of the model. Three different models are reported, first using the 90% of the projected data and 83,742 sentences are used for the test data (meaning 10% of the final projected data). Secondly a model using 669,937 sentences for the training (meaning the 80% of the final projected data) and 167,484 sentences are used for testing the model (corresponding to 20% of the final data). Finally, 502,453 sentences are used for the training (meaning the 60% of the final data) and 315,077 sentences are used for testing the model (40% of the final data).

	precision	recall	f1-score	support
B-PERSON	0.78	0.46	0.58	4,716
I-PERSON	0.87	0.59	0.70	2,883
B-LOC	0.73	0.37	0.49	3,208
I-LOC	0.80	0.14	0.24	467
B-ORG	0.34	0.22	0.27	2,082
I-ORG	0.50	0.21	0.29	1,199
micro avg	0.70	0.40	0.51	14,555
macro avg	0.67	0.33	0.43	14,555
weighted avg	0.70	0.40	0.50	14,555

Table A.3: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 60%(502,453 sentences) of English-Dutch European Parliament corpus and evaluated on CoNLL (15,806 sentences)

	precision	recall	f1-score	support
B-PERSON	0.76	0.73	0.75	62,271
I-PERSON	0.71	0.80	0.75	13,797
B-LOC	0.82	0.88	0.85	143,948
I-LOC	0.75	0.87	0.81	5,684
B-ORG	0.79	0.82	0.81	304,131
I-ORG	0.73	0.93	0.82	65,035
micro avg	0.78	0.84	0.81	594,866
macro avg	0.76	0.84	0.80	594,866
weighted avg	0.78	0.84	0.81	594,866

Table A.4: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 60%(502,453 sentences) of English-Dutch European Parliament corpus and evaluated on the 40% of the initial dataset that was left out as an evaluation dataset(334,968 sentences)

	precision	recall	f1-score	support
B-PERSON	0.79	0.47	0.59	4,716
I-PERSON	0.88	0.61	0.72	2,883
B-LOC	0.74	0.38	0.50	3,208
I-LOC	0.83	0.15	0.26	467
B-ORG	0.35	0.22	0.27	2,082
I-ORG	0.49	0.20	0.28	1,199
micro avg	0.71	0.41	0.52	14,555
macro avg	0.68	0.34	0.44	14,555
weighted avg	0.71	0.41	0.51	14,555

Table A.5: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 80%(669,937 sentences) of English-Dutch European Parliament corpus and evaluated on CoNLL (15,806 sentences)

	precision	recall	f1-score	support
B-PERSON	0.78	0.79	0.78	34,943
I-PERSON	0.73	0.83	0.78	6,654
B-LOC	0.80	0.86	0.83	62,885
I-LOC	0.76	0.89	0.82	2,879
B-ORG	0.80	0.84	0.82	153,722
I-ORG	0.73	0.92	0.81	29,458
micro avg	0.79	0.84	0.81	290,541
macro avg	0.77	0.85	0.81	290,541
weighted avg	0.79	0.84	0.81	290,541

Table A.6: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 80%(669,937 sentences) of English-Dutch European Parliament corpus and evaluated on the 20% of the initial dataset that was left out as an evaluation dataset(167,484 sentences)

	precision	recall	f1-score	support
B-PERSON	0.78	0.48	0.60	4,716
I-PERSON	0.88	0.62	0.73	2,883
B-LOC	0.75	0.38	0.51	3,208
I-LOC	0.86	0.15	0.26	467
B-ORG	0.35	0.23	0.28	2,082
I-ORG	0.49	0.20	0.29	1,199
micro avg	0.71	0.42	0.53	14,555
macro avg	0.68	0.35	0.44	14,555
weighted avg	0.71	0.42	0.52	14,555

Table A.7: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 90%(753,679 sentences) of English-Dutch European Parliament corpus and evaluated on CoNLL (15,806 sentences)

	precision	recall	f1-score	support
B-PERSON	0.79	0.81	0.80	17,209
I-PERSON	0.75	0.85	0.80	3,312
B-LOC	0.80	0.86	0.83	31,280
I-LOC	0.75	0.88	0.81	1,403
B-ORG	0.80	0.84	0.82	76,662
I-ORG	0.73	0.92	0.82	14,398
micro avg	0.79	0.85	0.82	144,264
macro avg	0.77	0.86	0.81	144,264
weighted avg	0.79	0.85	0.82	144,264

Table A.8: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 90%(753,679 sentences) of English-Dutch European Parliament corpus and evaluated on the 10% of the initial dataset that was left out as an evaluation dataset(83,742 sentences)

Appendix B

Evalutation of the German-English model

B.1 Comparison between "Translation & Similarity metrics projection" pipeline and a model trained on manually annotated data

In this section the comparison between the "Translation & Similarity metrics projection" pipeline model and a model trained on manually annotated data using CRF with basic features is presented. The experiments are conducted using training data with size of 3,538 or 1,000 sentences, and the evaluation is on the CoNLL gold corpus with size of 2,867 sentences and on data held out from the training dataset with size of 858 sentences.

	precision	recall	f1-score	support
B-PERSON	0.77	0.34	0.47	1,385
I-PERSON	0.85	0.60	0.70	605
B-LOC	0.68	0.37	0.48	1,178
I-LOC	0.61	0.07	0.12	162
B-ORG	0.37	0.15	0.21	1,240
I-ORG	0.67	0.06	0.11	877
micro avg	0.67	0.28	0.39	5,447
macro avg	0.66	0.26	0.35	5,447
weighted avg	0.65	0.28	0.37	5,447

Table B.1: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 3,538 sentences of English-German European Parliament corpus and evaluated on the CoNLL dataset

	precision	recall	f1-score	support
B-PERSON	0.88	0.68	0.77	75
I-PERSON	1.00	0.69	0.82	29
B-LOC	0.90	0.61	0.73	147
I-LOC	0.00	0.00	0.00	34
B-ORG	0.21	0.87	0.34	84
I-ORG	0.82	0.83	0.83	60
micro avg	0.48	0.66	0.56	429
macro avg	0.63	0.61	0.58	429
weighted avg	0.69	0.66	0.62	429

Table B.2: Target Language Named-Entity Recognition evaluation. CRF model trained on data extracted using the "Translation & Similarity metrics projection" approach on 3,538 sentences of English-German European Parliament corpus and evaluated on test set (858 sentences)

	precision	recall	f1-score	support
B-PERSON	0.71	0.10	0.17	1,385
I-PERSON	0.76	0.22	0.34	605
B-LOC	0.73	0.15	0.25	1,178
I-LOC	0.21	0.07	0.10	162
B-ORG	0.29	0.02	0.03	1,240
I-ORG	0.24	0.03	0.05	877
micro avg	0.60	0.09	0.16	5,447
macro avg	0.49	0.10	0.16	5,447
weighted avg	0.54	0.09	0.15	5,447

Table B.3: Target Language Named-Entity Recognition evaluation. CRF model trained on manually annotated data and evaluated on the CoNLL dataset

	precision	recall	f1-score	support
B-PERSON	1.00	0.79	0.88	75
I-PERSON	1.00	0.10	0.19	29
B-LOC	0.95	0.52	0.67	147
I-LOC	0.95	0.62	0.75	34
B-ORG	0.97	0.79	0.87	84
I-ORG	0.89	0.93	0.91	60
micro avg	0.95	0.66	0.78	429
macro avg	0.96	0.62	0.71	429
weighted avg	0.96	0.66	0.75	429

Table B.4: Target Language Named-Entity Recognition evaluation. CRF model trained on manually annotated data evaluated on test split

Bibliography

[1] Alan Akbik, Tanja Bergmann, and Roland Vollgraf. "Pooled Contextualized Embeddings for Named Entity Recognition". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 724–728. DOI: 10.18653/v1/N19-1078. URL: https://www.aclweb.org/anthology/N19-1078.

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1638–1649. URL: https://www.aclweb.org/anthology/C18-1139.

[3] Alan Akbik, Duncan Blythe, and Roland Vollgraf. "Contextual String Embeddings for Sequence Labeling". In: *COLING 2018, 27th International Conference on Computational Linguistics*. 2018, pp. 1638–1649.

[4] Luisa Bentivogli and Emanuele Pianta. "Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus". In: *Natural Language Engineering* 11.3 (2005), pp. 247–261.

[5] Upasana Biswas. "Normalization of Extracted Named-Entities in Text Mining". MA thesis. 2019.

[6] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[7] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[8] Maud Ehrmann, Marco Turchi, and Ralf Steinberger. "Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar, Bulgaria: Association for Computational Linguistics, Sept. 2011, pp. 118–124. URL: https://www.aclweb.org/anthology/R11-1017.

[9] Abbas Ghaddar and Philippe Langlais. "Winer: A wikipedia annotated corpus for named entity recognition". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2017, pp. 413–422.

[10] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[11] Eduard Hovy et al. "OntoNotes: The 90% Solution". In: Association for Computational Linguistics, 2006.

[12] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv preprint arXiv:1508.01991* (2015).

[13] Fredrick Edward Kitoogo, Venansius Baryamureeba, and Guy De Pauw. "Towards domain independent named entity recognition". In: *Strengthening the Role of ICT in Development* 4 (2008), pp. 84–95.

[14] Alexandre Klementiev and Dan Roth. "Named entity transliteration and discovery in multilingual corpora". In: *Learning Machine Translation* (2008).

[15] Philipp Koehn and Christof Monz. "Shared task: Statistical machine translation between European languages". In: *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. 2005, pp. 119–124.

[16] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: (2001).

[17] Oliver Mason. "Roger Garside, Geoffrey Leech, Anthony McEnery (eds). Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman. 1997. ISBN 0-582-29837-7 (Paperback).£ 26.00. 281 pages". In: *Natural Language Engineering* 5.3 (1999), pp. 301–307.

[18] Sepideh Mesbah et al. "Training Data Augmentation for Detecting Adverse Drug Reactions in User-Generated Content". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 2349–2359.

[19] Sepideh Mesbah et al. "Tse-ner: An iterative approach for long-tail entity extraction in scientific publications". In: *International Semantic Web Conference*. Springer. 2018, pp. 127–143.

[20] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[21] Joel Nothman, James R Curran, and Tara Murphy. "Transforming Wikipedia into named entity training data". In: *Proceedings of the Australasian Language Technology Association Workshop 2008*. 2008, pp. 124–132.

[22] Joel Nothman et al. "Learning multilingual named entity recognition from Wikipedia". In: *Artificial Intelligence* 194 (2013), pp. 151–175.

[23] Yaser Al-Onaizan et al. "Statistical machine translation". In: *Final Report, JHU Summer Workshop*. Vol. 30. 1999.

[24] Sebastian Padó and Mirella Lapata. "Cross-linguistic projection of role-semantic information". In: *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics. 2005, pp. 859–866.

[25] Matthew E Peters et al. "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365* (2018).

[26] Matthew Peters et al. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://www.aclweb.org/anthology/N18-1202.

[27] Matthew Peters et al. "Semi-supervised sequence tagging with bidirectional language models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1756–1765. DOI: 10.18653/v1/P17-1161. URL: https://www.aclweb.org/anthology/P17-1161.

[28] Telmo Pires, Eva Schlinger, and Dan Garrette. "How multilingual is Multilingual BERT?" In: *arXiv preprint arXiv:1906.01502* (2019).

[29] Lance Ramshaw and Mitch Marcus. "Text Chunking using Transformation-Based Learning". In: *Third Workshop on Very Large Corpora*. 1995. URL: https://www.aclweb.org/anthology/W95-0107.

[30] Alexander E Richman and Patrick Schone. "Mining wiki resources for multilingual named entity recognition". In: *Proceedings of ACL-08: HLT*. 2008, pp. 1–9.

[31] Erik F Sang and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". In: *arXiv preprint cs/0306050* (2003).

[32] Rushin Shah et al. "SYNERGY: a named entity recognition system for resource-scarce languages such as Swahili using online machine translation". In: *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*. 2010, pp. 21–26.

[33] Erik F. Tjong Kim Sang. "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of CoNLL-2002*. Taipei, Taiwan, 2002, pp. 155–158.

[34] Daniel Vliegenthart et al. “Coner: A Collaborative Approach for Long-Tail Named Entity Recognition in Scientific Publications”. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2019, pp. 3–17.

[35] Martin Volk, Anne Göhring, and Torsten Marek. “Combining parallel treebanks and geo-tagging”. In: (2010).

[36] David Yarowsky, Grace Ngai, and Richard Wicentowski. “Inducing multilingual text analysis tools via robust projection across aligned corpora”. In: *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics. 2001, pp. 1–8.