Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Matias Johansen Vian

A Study of Transformers for Cross-Corpus Native Language Identification

Master's thesis in Computer Science
Supervisor: Björn Gambäck
June 2023

NTNU
Norwegian University of
Science and Technology

Matias Johansen Vian

A Study of Transformers for Cross-Corpus Native Language Identification

Master's thesis in Computer Science
Supervisor: Björn Gambäck
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Native Language Identification (NLI) aims to discover a person's first language based on something they have communicated in a second language. While first attempted in 2005, NLI has later on been the subject of numerous research papers, all of which have contributed to this task. However, in keeping up with the ever-changing world of Artifical Intelligence (AI), new research is needed. This thesis attempts to accurately identify native languages using the recently introduced *Transformer* in a cross-corpus setting. A focus of this report is to experiment with these transformers using different types of data, while attempting to prevent the model from incurring a bias toward any specific domains or genres, which is why cross-corpus evaluation is applied. The experiments in this work are divided into three parts: the first part focuses on determining suitable data subsets for this task, which includes gathering additional, novel data from Reddit, while the second part focuses on which transformer-based model will perform best. The third part uses the configuration found in the previous two, to explore how domain adaptation will affect the results. The best model in this work was a regular BERT classifier, which achieved a cross-corpus accuracy score of 52.0% when testing on TOEFL11, and training on a combination of three datasets (FCE, Italki-NLI, and Lang8). This thesis includes numerous, large experiments with multiple transformers. The emphasis on these experiments is how different data will affect different types of models. The other models in this report are inspired by Lotfi et al. (2020), and use BERT and GPT-2 in a multi-model setup, but these were shown to perform poorer in cross-corpus experiments. It was also shown that the particular method for domain adaptation attempted in this work did not improve the results.

Sammendrag

Førstespråksidentifisering har som mål å finne en persons førstespråk basert på noe de har kommunisert på sitt andrespråk. Dette er et fagfelt som først ble etablert i 2005, og har siden da vært et tema på en rekke forskningsartikler, som alle har bidratt til å framme utviklingen av feltet. Imidlertid krever den stadig skiftende verdenen innen kunstig intelligens ny forskning. Denne master-oppgaven bruker den nylig introduserte *transformeren* til formålet å identifisere førstespråk på tvers av korpus. Et fokus vil være å eksperimenter med disse transformerene ved hjelp av ulike typer data, og samtidig forhindre å utvikle partiske modeller som lener mot noe spesifikt domene eller sjanger, som er årsaken til bruken av krysskorpusevaluaring. Eksperimentene i denne rapporten er delt i tre deler: den første fokuserer på å avgjøre egnede datamengder, som blant annet inkluderer en introduksjon av ny Reddit-data, mens den andre delen fokuserer på hvilken transformer-basert modell vil være mest effektiv. Den tredje delen bruker konfigursjonen i de to forige til å utførske hvordan konseptet domenetilpasning vil påvirke resultatene. Den beste modellen i denne master-oppgaven var en standard BERT-klassifiserer, som oppnådde en krysskorpus-nøyaktighet på 52.0% når den ble testet på TOEFL11, og trent på en kombinasjon av tre datasett (FCE, Italki-NLI og Lang8). Dette studiet inkluderer et flertall store transformer-eksperimenter, med fokus på hvordan disse modellene oppfører seg i ulike omstendigheter. De andre modellene i disse eksperimentene er inspirert av Lotfi et al. (2020), og tar i bruk BERT og GPT-2 i et multimodelloppsett, men disse modellene presterte svakere i krysskorpusevaluering. Det ble også vist at den spesifikke metoden for domenetilpasning som ble forsøkt i dette arbeidet ikke forbedret resultatene.

Preface

This Master's thesis marks the completion of my education at the Norwegian University of Science and Technology (NTNU), and was undertaken my last semester during the spring of 2023.

The literary study for this thesis was performed as part of my specialisation project in the autumn of 2022, in which my task was to specialise in a narrow field within computer science. During the project, I gathered data and research about Native Language Identification, which constitutes my literary foundation and pre-study. Selected parts of the background for this thesis are adapted from this project. These parts will be signified in the relevant chapters.

Special thanks to my supervisor Björn Gambäck for all his guidance and feedback.

Furthermore, I would like to thank everyone who has made their data available: Yannakoudakis et al., Hudson and Jaf, Brooke and Hirst, Rabinovich et al., Blanchard et al., and Edvardsen. Special thanks to Thomas Hudson, who sent their data to me directly.

Matias Johansen Vian

Trondheim, 17th June 2023

Contents

List of Figures

List of Tables

Acronyms

AI Artifical Intelligence.

API Application Programming Interface.

BERT Bidirectional Encoder Representations from Transformers.

BPE Byte Pair Encodings.

E Experiment - followed by the experiment number and the sub-experiment number. E.g.: EII.I.

FCE First Certificate of English.

GPT Generative Pretrained Transformer.

ICANALE International Corpus Network of Asian Learners of English.

ICCI International Corpus of Crosslinguistic Interlanguage.

ICLE International Corpus of Learner English.

L1 Native language.

L2 Second language.

NLI Native Language Identification.

POS Part-of-Speech.

SoMe Social Media.

SVM Support Vector Machine.

TOEFL11 Test of English as a Foreign Language.

1 Introduction

Native Language Identification (NLI) aims to find a person's first language (L1) based on something they have communicated in a second language (L2), either written or verbally. It is a classification task that makes use of machine learning algorithms to identify subtle patterns and small differences between nationalities when they communicate in second languages. The task was first attempted in 2005 by Koppel et al., and has since been the subject of two shared tasks and several research papers. The number of L1s to distinguish between has ranged from 2 to 23 languages, and different researchers have focused their studies on different L2s, even though English is by far the most researched language.

NLI is first and foremost a useful tool within language education, where information about an author's native language can play an important role in the learning process. Distinguishing a learner's first language can enable targeted and more detailed feedback, leading to a better learning experience. NLI is also a useful tool for law enforcement aiming to deduce information about suspects, and for marketing entities who would like to classify their customers.

There are several challenges residing in NLI: the challenge of finding suitable data annotated with native language, and the challenge of discovering good language models, not to mention feature engineering, training, and evaluation. The current State of the Art utilises the relatively new Transformer (Vaswani et al., 2017) for NLI, which has achieved some of the best results in the field (Steinbakken and Gambäck, 2020; Lotfi et al., 2020). It is upon these results that this thesis is built, with the goal of optimising and familiarising with these models in order to correctly utilise them for this task.

In addition to the transformer, cross-corpus evaluation will also be central. Cross-corpus evaluation is used to identify independent features and to generalise across genres. The method uses a different corpus stemming from a different domain than the training set to test the language model. It is a further expansion of the task of NLI and may be considered a subtask within the field. The method presumably stands a better chance of discovering truly neutral NLI models able to function across domains and text types. Cross-corpus evaluation was first put into the spotlight in the first shared task within NLI in 2013 (Tetreault et al., 2013), and the subfield has been touched upon by several NLI researchers. However, a very small amount of research exists on applying transformers to this task, which is why this is the focus of this thesis. Different data and combinations of data will be tested, and different transformer architectures will be used. Additionally, domain adaptation methods will be added to these experiments, to further advance the results of cross-corpus evaluation.

1.1 Background and Motivation

The field of Artifical Intelligence (AI) is always in movement and has seemingly reached an unstoppable surge of innovation and advancement. AI has never been more in focus with recent discoveries of advanced chatbots, AI art, and deep fake content. Among these advancements is the Transformer, which in recent years has revolutionised the field of natural language processing. The transformer was introduced in 2017 and has already tackled language processing problems in a manner we have not seen before. Additionally, the transformer itself has evolved into everchanging shapes and structures, being optimised and specialised for several different problem spaces. There now exist hundreds of variations and descendants of the original transformer.

In the field of Native Language Identification, these transformers have made huge strides toward more accurate and more confident classification models. However, given the novelty and complexity of the model, relatively little research exists on this. Additionally, due to the heavily active field of transformers, it can be hard to keep up with the progress. Therefore, a proper study of different transformers should be conducted, which is the core motivation behind this thesis.

Another element of this study is cross-corpus evaluation. A large part of previous research has focused on within-corpus training, which makes for good, but narrow, models, but in order for NLI applications to truly be of use, the technology behind should be able to process all types of language, and not just the type seen in training. Thus, a seemingly natural next step for NLI is to evolve more general models, which means that cross-corpus experiments should be focused. Huge strides have been taken in other language processing applications, and the goal is that these same strides will be seen in NLI by developing more universal classifiers.

1.2 Goals and Research Questions

The goal of this Master's thesis is the following:

Goal *Achieve accurate NLI by discovering the optimal transformer model in combination with the optimal subset of training data when evaluating across corpora.*

The evaluation metric for this will be measured in the model's ability to correctly predict an unseen text, specifically given by the portion of correct predictions divided by the total samples in the test set. The goal will be separated into three research questions, which will constitute the basis for three experiments, all of which are related to experimenting with transformers in cross-corpus settings. All research questions will be tested twice on one learner corpus, in which the data stems from language learners writing small texts, and one corpus stemming from Social Media (SoMe).

The first research question, which is the basis for Experiment I, aims to find the best cross-corpus subset of training data for each of the testing sets. Brooke and Hirst

(2011) performed one of the earliest cross-corpus NLI experiments and found that their model performed best when training on the maximum amount of available data. It is also known that transformers generally need larger amounts of data to work properly. What is not known is whether some data will be more detrimental than helpful, for example when the training data is very different from the test data. An interesting factor here will be the divide between learner corpora and SoMe corpora, which differ in structures, topics, and language types, as well as other aspects. Will a transformer prefer large amounts of less-than-ideal data or a smaller amount of ideal data?

Research question 1 *What data will produce the best Native Language Identification model?*

The second research question will use the training circumstances and data found in Experiment 1 to determine the best NLI model for the task, given by its classification accuracy (see Section 2.5). Clearly, not all models can be tested, but a subset of high-performing models will be chosen, based on previous research. All of these are transformer-based, as these have shown to be the leading performers in recent years. Additionally, because of the novelty of the technology, these are bound to be the least explored ones, which might indicate unlocked potential.

Research question 2 *Which transformer-based State of the Art model will perform best?*

The third and last research question is to determine how domain adaptation will affect the classification accuracy using the ideal test setup found in the previous two experiments. Domain adaptation has shown to improve the accuracy of NLI models (Stehwien and Padó, 2016), however, it is not know how transformer models will behave when paired with this technique. The best training subset of Experiment I and the highest performing model of Experiment II will be used in combination with domain adaptation to further improve the model.

Research question 3 *How does domain adaptation affect a transformer-based cross-corpus model?*

1.3 Research Method

As has been mentioned in the preface, the chief portion of the literary study was conducted preceding this work during my specialisation project. The literary study in the specialisation project was further built upon in this work, which ultimately laid the groundwork for a goal and set of research questions. These comprise the foundation for this thesis.

The initial phase of the project portion of this work consisted of gathering and comparing suitable datasets. Additionally, new data was extracted from the internet to balance one of these datasets. After preprocessing the data, the experiments were designed with the research questions in mind, and subsequently conducted. The last phase involved an analysis and evaluation of the experimental results.

1.4 Contributions

1. *A thorough comparison of different data subsets, showing how different types of data might affect a cross-corpus evaluation of a Transformer.*
2. *A comparison of different transformer-based models, and how they perform in cross-corpus evaluation.*
3. *A cross-corpus classification score of 52.0% (see Section 2.5).*
4. *An addition to the dataset, Reddit-L2 (Rabinovich et al., 2018), which here will be named Reddit-L2.1.*

1.5 Thesis Structure

The following list describes the overall structure of this thesis:

- Chapter 2 provides the background theory for this report, which will cover the transformers, concepts, and tools used in the experimental setup.
- Chapter 3 will put this work in perspective of the field.
- Chapter 4 gives a brief overview of the different corpora, their sources, and their typical structures along with some discussion of their reliability and usability. This will be further discussed in chapter 7.
- Chapter 5 describes the specific experiments related to the three research goals, as well as some reasoning for the choices underlying them. This includes descriptions of the specific model architectures.
- Chapter 6 gives a plain overview of the results.
- Chapter 7 evaluates and discusses these results.
- Chapter 8 concludes the report with additional thoughts of future work.

2 Background Theory

This chapter will provide the theoretical foundation for understanding this thesis, starting with the fundamentals of neural networks and the transformer, which is a distinct architecture within the realm of neural networks. Further on, the specific implementations of the transformer used in the experiments are described in detail. Additionally, the concrete tools used in the experiments are presented, which include everything from the machine learning libraries to evaluation metrics to the tools related to data processing. Finally, some important key concepts needed to understand the chapter about related work are explained.

Note that all figures in this chapter, along with Section 2.6, were copied from my specialisation project in the autumn of 2022. See the Preface for information regarding the nature of that project.

2.1 Neural Networks

Neural Networks represent a fundament for deep learning. They consist of a network of nodes, divided into layers, which are connected to each other. Each node retains a value based on the inputs from upstream layers, and furthermore affects every node connected to it in downstream layers, and thus a signal is propagated through the network. This signal starts from a layer of input nodes, and ends with a layer of output nodes, with a certain number of hidden layers in between. A simple illustration of this is depicted in Figure 2.2 on the following page. The main task of a researcher is to distill a problem to fit a set of input nodes, and to interpret the output nodes as labeled information.

Supervised learning within neural networks consists of propagating a sample through the network, denoting the result and loss, and adjusting the weights accordingly through a process called backpropagation. The process of backpropagation and the mathematics behind it is described in detail in (Rosenblatt, 1962).

2.2 The Transformer

The Transformer is central to this thesis. It was first introduced by Vaswani et al. in 2017 and is a Deep Neural Network based upon a self-attention mechanism at its core, along with embedding layers, feed-forward layers, normalisation layers, and an output layer. The embedding layer is described in detail in Section 2.2.3. Transformers are

pre-trained on very large amounts of data and subsequently fine-tuned to fit a range of purposes, from text classification to chatbots. In abstract terms, the Transformer is divided into an encoder where representations capturing contextual relationships are produced, followed by a decoder which uses these representations to generate outputs to fit a specified purpose.

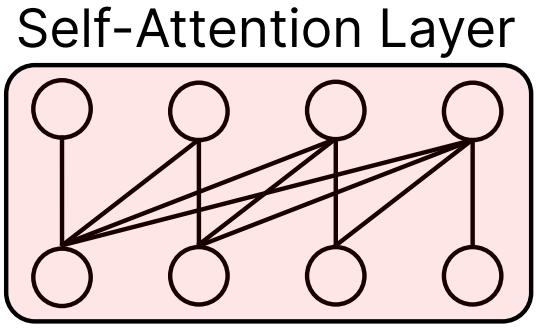


Figure 2.1: A self-attention layer.

The self-attention mechanism, which is depicted in Figure 2.1 inputs a sequence of words up until, and including, the current word, as opposed to a fully connected feedforward network where all nodes of a layer are connected to the nodes in the next, as seen in Figure 2.2. In attention layers, specific parts of a sequence are weighted depending on their impact, which provides every word with a specific context. In actual implementations of the transformer, multi-head self-attention components are used, meaning that every word keeps track of multiple sets of weights to capture multiple sets of contexts, reflecting different grammatical and syntactic relationships between them. Self-attention also allows every word to be processed in parallel, making the model both efficient and able to handle sequences of varying lengths (Jurafsky and Martin, 2022).

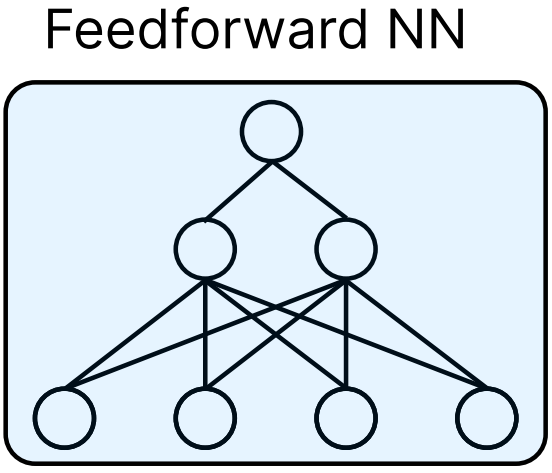


Figure 2.2: A three-layer feedforward Neural Network.

The other components of the encoding block of the transformer are depicted in Figure 2.3: a regular, fully connected feed-forward network, a normalisation layer, and an output layer, each of which provides different benefits for the model. The feed-forward layers provide complexity and memory, the normalisation layers keep the model from diverging and the output layer is designed to be interpretable. Multiple blocks of these, along with self-attention layers and decoding blocks, comprise the transformer. The decoding blocks are similar to the encoders, with the addition of multiple attention layers receiving input from the various encoder blocks. Regular inputs from the encoders ensure a closely linked relationship between the two blocks (Jurafsky and Martin, 2022).

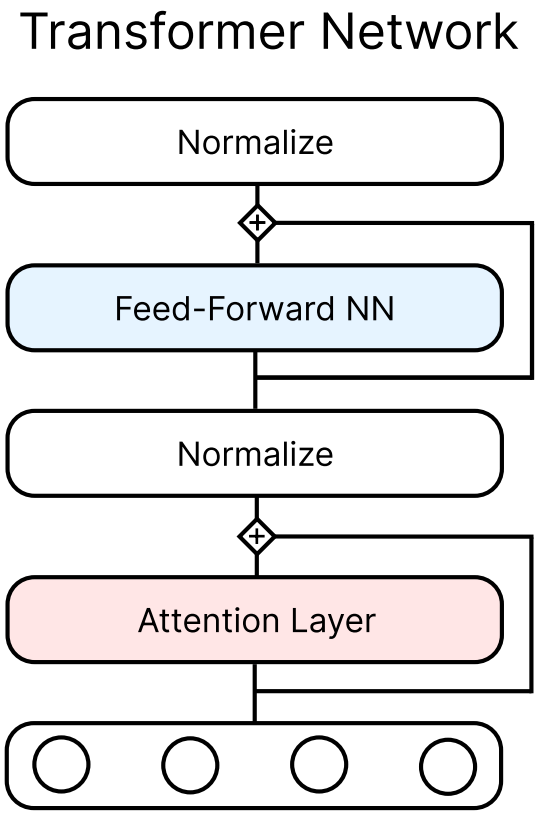


Figure 2.3: A simplification of the original transformer network. The input at the bottom is concatenated with itself after processed through an attention layer. This output is then normalised before the process is repeated with a feed-forward layer.

2.2.1 BERT

Bidirectional Encoder Representations from Transformers, or BERT, was developed by Google and is one of the closest implementations to the original transformer by Vaswani et al. (2017). It is specifically designed to draw out information about a sequence, using an

encoder-based architecture, meaning that the sequences are reduced to lower dimensional representations of themselves.

Like the transformer, BERT makes use of multihead self-attention layers, with the difference of inputting both the context to the left and to the right of a given word (hence 'bidirectional'). 12 of these self-attention layers comprise the 12 encoder layers of $BERT_{base}$, in addition to a fully connected layer, normalisation, and output layer which is built on top. $BERT_{large}$ has the same architecture but consists of 24 encoding layers.

BERT is pre-trained on about 3.3 billion words from books and Wikipedia. This pre-training is divided into two tasks: the first involves predicting masked tokens from a sequence, while the second focuses on predicting which sentence would be most likely to supersede a given input sentence. Following this, the model must be fine-tuned for specific tasks (Devlin et al., 2019).

2.2.2 GPT

Generative Pretrained Transformer (GPT) was developed by OpenAI and is often considered the leading counterpart to BERT. The two transformers, while similar in implementation of the original transformer, differ in general architecture, task type, training method, pre-training data, and masking strategy. While BERT is encoder-based, GPT is decoder-based, meaning that inputted text is used and decoded to generate outputs. It is therefore more suited for generating text sequences instead of extracting information. In the pre-training process, the model is trained on predicting the next word in a sequence, as well as predicting masked tokens (Radford et al., 2019). There exist four generations of GPT, the most significant difference being the number of parameters or nodes. GPT-2 is the second generation of GPT, and is the implementation that was used in (Lotfi et al., 2020) and that will also be used in this thesis. The reason for this is because the entire implementation is available to the public, whereas the two newer generations are available only through chat and Application Programming Interfaces. This will be further discussed in Chapter 5, explaining the details around the model choices.

2.2.3 Transformer Tokenising

All transformer inputs are tokenised before being fed into the model. This process begins by splitting the text into tokens, where words and punctuation are separated. These tokens are compressed using Byte Pair Encodings (BPE), where frequent subsets of characters are compressed into a single character. These BPE representations are subsequently mapped to a numerical value representation, based on a large, static vocabulary library. This vocabulary library is based on the most frequent words found in very large corpora. Finally, the tokenised text is padded to a fixed length and formatted in a way that can be processed by the Transformer model (Jurafsky and Martin, 2022).

2.3 Domain Adaptation

Domain adaptation (Farahani et al., 2020) is a technique within transfer learning where a model is trained to adapt a domain different than the learning data. These domains are called the source and target domains and are different, but related to each other. The technique is natural to take advantage of within cross-corpus learning.

There are multiple ways of applying domain adaptation to a model, and these are generally divided into three different types:

- Unsupervised domain adaptation, in which the target domain as well as a subset of the source domain are unlabelled.
- Semi-supervised domain adaptation, in which the target domain additionally contains a set of labeled samples.
- Supervised domain adaptation, in which all data is labeled.

Additionally, there are usually three different approaches:

- Divergence-based, where we search for features that are the least different between the source and target.
- Adversarial-based, where the goal is to train the model in both feature representation and domain representations, so that the model might learn which features are relevant for different domains.
- Reconstruction-based, where we search for representations that are capable of reconstructing the original input data accurately and effectively by learning latent space representations.

Some of these approaches can be done either unsupervised, semi-supervised or supervised, resulting in relatively many techniques. In this thesis, the focus will be adversarial, supervised domain adaptation.

In adversarial domain adaptation, the model is set to train on two different, or adverse, tasks, where the first task is the original one (which in this thesis is a classification problem), while the second is related to the domain itself. Thus, the model is simultaneously trained to both classify the original task and the domain, forcing it to rely on domain-invariant features, which are features that will not vary across text types, such as punctuation, spelling errors, etc.

In a practical implementation of adversarial domain adaptation, the domain classification is added as a new model head, where its own loss is calculated, which is subsequently added to the loss of the original task, both of which comprise the basis for the backpropagation. The domain-adaptive model used in this thesis is described in Section 5.3. In the training process, a source domain dataset and a target domain dataset are zipped together, where source samples are used to train the main classifier as well as the domain classifier, while target samples are used to train the domain classifier only (Farahani et al., 2020).

2.4 Tools

The code for this thesis was written in Python and makes use of a number of different libraries, which will be presented here.

2.4.1 Hugging Face Transformers

Hugging Face (Wolf et al., 2020) is an artificial intelligence hub for sharing AI model implementations and is notable for its library of Transformers, which are implemented using the Pytorch library described in Section 2.4.5. They offer a user-friendly Application Programming Interface (API), extensive documentation, and many additional tools for natural language processing. In this thesis, all transformer implementations and associated tokenisers are borrowed from the Hugging Face transformer library.

2.4.2 Optuna

Optuna (Akiba et al., 2019) is a Python library designed to automate the process of hyperparameter optimisation in machine learning model development. It uses a sequential model-based optimisation approach based on Bayesian inference principles to navigate the hyperparameter space of a neural network and find the best settings. Researchers define the range and type of hyperparameters to search over, and an objective function is defined to measure the performance of different configurations. Optuna suggests new values to evaluate based on the results of previous trials, continuously updating its probabilistic model of the objective function. The process also involves pruning mechanisms, which stop unpromising trials early on.

2.4.3 Pandas

Pandas (Wes McKinney, 2010) is the most widely used Python library for data handling and manipulation and provide the basis for all data processing in this project. It is built on top of the NumPy library, and provides several features to easily work with tabular data.

2.4.4 Pushshift API

The Pushshift API[1] is a tool for accessing large amounts of Reddit data. It uses a custom database that is built on top of Reddit's data, which in contrast to the Reddit database, is designed to be fast and efficient to easily retrieve data in bulk. Additionally, the API allows users to search for posts based on a wide range of criteria, including keywords, subreddit, author, and date.

[1] https://reddit-api.readthedocs.io/en/latest/

2.4.5 Pytorch

Pytorch is an AI library for Python, and contains everything required for running an AI model, such as data loaders, neural network base models, layers, optimisers, and all calculations needed to train and test a neural network. For more information about the library see (Paszke et al., 2019).

2.4.6 Polyglot

Polyglot (Al-Rfou et al., 2013) is a Python library for natural language processing and includes tools for everything from word embedding to named entity recognition. In this thesis, their language detector is utilised. The language detector works by analysing the character n-grams (sequences of n characters, see Section 2.7) present in a given text and comparing them with the n-grams present in a set of pre-defined language profiles. The language profile for each language consists of a set of n-grams and their frequencies, which have been calculated from a large corpus of texts in the particular language.

When given a new text, the language detector calculates the frequency of each n-gram and compares them with the n-gram frequencies in each language profile. The closest language profile is then selected.

2.5 Evaluation Metric: Accuracy

Accuracy is used to evaluate the models, and is given by the number of correctly predicted samples divided by the total number of samples:

$$\frac{\#\ Correct\ samples}{\#\ Total\ samples} \tag{2.1}$$

2.6 The Support Vector Machine

Support Vector Machine (SVM, Cortes and Vapnik, 1995) has, in the past, been one of the most broadly used models for Native Language Identification. It is a binary classification model that maps linearly separable data into a space that maximises the hyperplane gap between two classes. This hyperplane is used to classify unseen data, depending on which side specific observations are mapped to. When the data is not linearly separable, it is mapped into a higher dimensional feature space, in which a hyperplane capable of separating it can be found. This method is called the kernel method, and is illustrated in Figure 2.4. For SVMs to handle multi-class problems the model must be broken down into multiple binary classification problems in a one-vs-one or one-vs-all approach (Cortes and Vapnik, 1995).

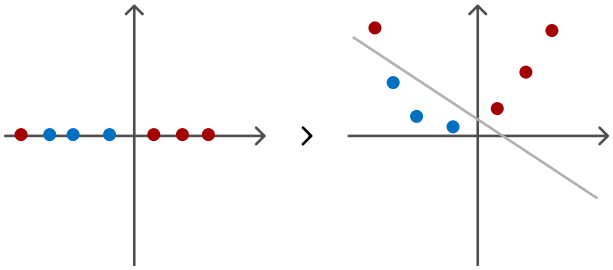


Figure 2.4: The kernel process in an SVM. The points are mapped from one to two dimensions, where they are separable.

2.7 N-grams

A very common lexical feature used in NLI is n-grams. N-grams can be both character- and word-based, and represent the probability of encountering a sequence of n characters or words. When calculating word n-grams, the probability of every word is given as $P(w_n|w_{n-N+1:n-1})$, where N is the n-gram size and n is the n-th word. The size of the n-gram thus represents the scope of the probability. These probabilities are based on a large corpus, and the same logic can be used to find character n-grams, which are based on individual character probabilities (Jurafsky and Martin, 2022).

3 Related Work

Early work within the field is characterised by rudimentary research and lack of data. Much of early research used the International Corpus of Learner English (ICLE), which had its limitations. Two milestones in the field are the shared tasks of 2013 and 2017, the first of which marked the introduction of the TOEFL11 corpus (Blanchard et al., 2013). See Chapter 4 for more information about ICLE and TOEFL11. The shared tasks both manifest in a surge of research and maximisation of accuracy using previous techniques while still exploring new pathways for future research. They have paved the way for recent work, which has focused on cross-corpus comparisons along with further analysis of features and methods. Additionally, the use of transformer-based language models has come to be of interest in recent years. The beginning of this chapter will be split into Early Work (Section 3.1), the First Shared Task of 2013 (Section 3.2), the Second Shared Task of 2017 (Section 3.3), and Recent Work (Section 3.4), all of which will focus on the general history and advancements in Native Language Identification (NLI), giving some context for this work. The section regarding recent work will be of particular interest. Separate from this, will be a section on cross-corpus NLI (Section 3.5), which is the focus of this thesis.

A selection of the material in this chapter was adapted from my specialisation project in the autumn of 2022. The four first sections contain adjustments and some additional content, while the fifth and last section is new.

3.1 Early Work

NLI was first explored by Koppel et al. (2005), who used a Support Vector Machine (SVM) trained on lexical and syntactic features of n-grams, Part-of-Speech (POS), function words and grammatical errors. They achieved an accuracy of 80.2% on a 7-way task. Working on the same dataset Wong and Dras (2011) achieved roughly the same accuracy using horizontal cross-sections of parse trees from the Stanford Parser (Klein and Manning, 2003) as inputs for a maximum entropy model. However, both research papers were based on the ICLE corpus, which as pointed out by Tetreault et al. in 2012, is flawed by topic bias, meaning that the corpus had an uneven topic distribution over the languages. This would result in models that relied on certain topic words instead of the actual linguistic traits of an author. For example, Tetreault et al. (2012) pointed out that only Chinese authors, and no other group, responded to a particular prompt regarding the use of credit cards, which lead to only Chinese samples containing the character "$",

making it artificially easy to correctly classify Chinese samples. They therefore used a subset of the dataset called ICLE-NLI with a more balanced topic distribution, which partially alleviated this problem. They additionally explored using an ensemble of n-gram language models, making them the first to use ensemble methods for NLI, which led to a record accuracy of 90.1%. The n-gram language models used perplexity scores from a language model trained on 5-grams. While the ICLE-NLI subset reduced some of the topic bias within the dataset, Tetreault et al. saw the need for a more balanced dataset tailored for NLI and subsequently introduced the TOEFL11 corpus (Blanchard et al., 2013).

3.2 The First Shared Task of 2013

Up until then, research within the field had used different datasets and different metrics for building and evaluating their NLI systems. To remedy this and to unite the field a shared task was organised, in which the newly established TOEFL11 was to be employed. The 2013 Shared Task was arranged as a part of The 8th Workshop on Innovative Use of NLP for Building Educational Applications. 27 teams participated and a multitude of techniques and feature combinations were applied to the task. An in-depth review of these techniques and their outcomes can be found in (Tetreault et al., 2013).

Notable from the shared task was that a great majority either exclusively or partially relied on SVMs, including the top performing group (Jarvis et al., 2013), which got an 83.6% accuracy. Their setup was rather simple, compared to for example the one of Tetreault et al. (2012) who used ensemble learning and higher order word n-grams, but the fact remains that the simpleness of the SVM seems to be highly suitable for Natural Language Processing applications. Another outcome from the shared task was an increased focus on cross-corpus Native Language Identification, because the competition included two separate tasks where teams could train on third-party data, either instead or in addition to TOEFL11. Little to no research existed on this topic preceding the shared task, and while not too many teams chose to compete in any of these categories, it still brought the problem of cross-corpus evaluation to light. A more detailed description of the outcome of these sub-tasks is given in Section 3.5 regarding cross-corpus NLI.

3.3 The Second Shared Task of 2017

The shared task of 2017 was similar to the one in 2013, with the addition of speech data, which was collected and annotated the year before. This speech data consisted not of the actual recordings, but of transcripts and i-vectors of fixed length (800) representing the data. I-vectors are, simply put, short vectorised representations of an audio recording. The shared task consisted of three sub-tasks: one using only this speech data, one essay-only sub-task using only the TOEFL11 corpus like the previous shared task, and a hybrid of the two. 19 teams participated and the highest accuracy achieved was 93.18%,

which was achieved in the hybrid sub-task using a string kernel model. A complete summary and comparison of the systems are given by Malmasi et al. (2017), and some of the highlights will be given here.

Ionescu and Popescu (2017) placed within the first bracket of the essay-only task, and was the only ones to utilise a kernel model. They were the first to use a string kernel for NLI three years before in the first shared task, where they placed third with an accuracy of 82.7% (Tetreault et al., 2013). The efficiency of string kernels has been confirmed by Franco-Salvador et al. (2017), who used them in combination with word embeddings, and found that the string kernel provided the most information gain. According to Ionescu and Popescu, string kernels are superior to other alternatives, because the method is topic independent. Topic independence might suggest the method performs well for cross-corpus evaluation. However, no such research has been performed, by neither Ionescu and Popescu nor any other researchers.

Another group of the shared task (Bjerva et al., 2017) found that using Convolutional Neural Networks did not perform better than conventional methods, which was also confirmed by Chen (2016) the year before, who compared various deep learning models and found that an SVM still outperformed them, suggesting that deep learning methods would need some improvement to yield results. On the other hand, ensemble methods have consistently yielded great results for NLI. For example, the winners of the essay-only portion of the shared task were Cimino and Dell'Orletta (2017), who used the outputs of a Logistic Regressor trained on sentences as input to an SVM to predict the entire document. This has also been shown to work by Malmasi and Dras (2018), who used a stacked model to achieve State of the Art results on a number of experiments, including cross-corpus evaluation and testing on other languages.

3.4 Recent Work

The world of AI has come a long way since 2005, and newfound models and computational power has spurred better and more complex models that require more data to train. The field has diverged from the classical Support Vector Machines, Maximum Entropy models and Regressors into exploring transformers, larger neural networks, and bidirectional encoders. Additionally, the information age has not only increased the amount of written, digital data, but also made it more available, making it possible to use data from Twitter, Reddit, and other databases.

The transformer was first applied to Native Language Identification by Steinbakken and Gambäck in 2020, who used BERT (see Section 2.2.1) to achieve an 85% accuracy on TOEFL11. This was further improved upon by Lotfi et al. (2020), who used multiple GPT-2 models, fine-tuned for each L1, and their language modeling loss to predict the native language. They currently sit on a record accuracy of 89% on TOEFL11. There is no doubt the approach yields great results, but it does require an enormous amount of processing power and storage space, as it requires one entire GPT-2 model for each native

language, all of which must be queried for each prediction. Uluslu and Schneider (2022) propose a transformer adapter that is 13 times faster, but also lowers the accuracy by 4.8% on the same experiment. This new approach keeps some of the original, pre-trained weight frozen, and focuses the training process on a selected set of weights for each layer.

Both BERT and GPT-2 require large amounts of data, which has become more and more available in recent years. The field experienced a shift from the traditional learner corpora to user-generated content from social media and other large databases. Rabinovich et al. (2018) assembled their dataset from Reddit (see Section 4.6), and used it to get a 68.97% accuracy on a 23-way classification task when only using textual features. This was further improved upon by Steinbakken and Gambäck (2020), who used BERT to get a 90.2% accuracy, which presumably is the highest score achieved on a 23-way task. They underline that these results could not have been achieved without the large quantity of data. Volkova et al. (2018) also used user-generated data, but based their research on tweets to determine which types of features had the most predictive power, and concluded that content-related features of word n-grams and word embeddings carried the most information about the user's native language. Edvardsen (2021) also used tweets, but in a cross-corpus setting, and using BERT-based language models.

3.5 Cross-Corpus Native Language Identification

Cross-Corpus Native Language Identification may be considered a subfield or subtask of NLI. The goal of cross-corpus NLI is to achieve an accurate model that is able to perform well across corpora. These corpora may stem from the same domain, they may be similar, or they may be entirely different. The task represents a greater challenge than within-corpus NLI, and typically results in poorer results.

While within-corpus NLI has been researched since 2005, the earliest known work in cross-corpus NLI is that of Brooke and Hirst (2011), who achieved an accuracy of 26.7% in a seven-way classification task. They were also the first to use the Lang8 corpus (see Section 4.3), which they dubbed a 'cheap' learner corpus, on the grounds of it being scraped from a somewhat unreliable source with noise and unregulated text. Brooke and Hirst continued their work in the 2013 Shared Task, competing in the third-party data track, in which they obtained the highest accuracy, with a score of 56.5%, which was a vast improvement on their previous results, especially considering that this was an eleven-way task. They also found that using more, and more varied, data yielded better results, as the model might be able to generalise better this way. The training data they used was taken from First Certificate of English (FCE), the International Corpus of Learner English (ICLE), the International Corpus Network of Asian Learners of English (ICANALE) and the International Corpus of Crosslinguistic Interlanguage (ICCI) corpora, and they found that using all of them proved most efficient.

Brooke and Hirst also experimented with adaptation techniques in 2012. In this research, they applied a Support Vector Machine on Lang8, and tested it on ICLE and

FCE, with a simple form of adaptation techniques using bias adaptation, where the model bias was iteratively shifted using the test set. They reported promising results using this technique. Adaptation techniques were also used by Stehwien and Padó (2016), who compared two different methods: feature augmentation, which is a form of adverse domain adaptation, and using marginalised stacked denoising autoencoders, which is a form of divergent adaptation technique. In the former, the sample features exist in three different versions depending on if it is the original representation, the source representation, or the target representation. In the latter, a marginalised stacked denoising autoencoder is used to extract features from both the source and the target. They found that the marginalised stacked denoising autoencoder yielded the best results.

Since then, the work by Malmasi and Dras (2017) is mentionable. They achieved a 43.7% accuracy on TOEFL11 by training on the EF Cambridge Open Language Database (Geertzen et al., 2014), however, only using 9 native languages. Edvardsen (2021) was the first to use transformer-based models for this task. They experimented with a novel Twitter corpus (see Section 4.2) and TOEFL11 on an 11-way cross-corpus NLI task. Their best accuracy when testing on TOEFL11 was 26.8%, which was achieved using CT-BERT, a domain-specific BERT model pre-trained on Covid-related tweets.

Finally, cross-corpus evaluation, while often a goal in itself, is also a useful tool to analyse whether a specific feature will transcend corpora, thus testing the robustness of the feature. This was the case for Bykh and Meurers (2014) and Markov et al. (2020), both of which focused their research on analysing feature types. Bykh and Meurers conducted a similar study to that of Brooke and Hirst (2013), using largely the same corpora, and found that lexicalised-based production rules worked better than stylistic-based ones, but did not compare a very large number of features. Markov et al. conducted a much larger study and centred their research on presumably topic-independent features. These features included punctuation usage, emotion expression, and similarities between the L1 and L2 in the form of anglicised words, cognates, and other misspellings, and were all deemed as useful for NLI, also when testing them across corpora.

4 Data

An important part of this thesis is to explore how text types and sources affect results. This section will present all corpora used in this research, with emphasis on how they differ from each other. The datasets are either learner corpora or gathered from Social Media (SoMe), and only the parts of the datasets with languages relevant to this thesis will be considered. Below is a table (4.1) containing the number of documents per language for the subsets of data that is used in this thesis. It is important to keep in mind that the total number of words might differ between datasets as the size of each document varies across corpora. Reddit posts and Twitter posts are generally shorter than the learner essays in TOEFL11, FCE, Italki-NLI, and Lang8. Additionally, most of the datasets in this thesis are designed to be used in combination with others, which partially compensates for the uneven language distribution, as the datasets will complete each other to some degree.

Most of the following datasets are borrowed from previous research, with the exception being Reddit-L2.1, which was gathered specifically for this work, and therefore constitutes a part of the foreground of this thesis. This data, and the process of retrieving it, is described in the last section of this chapter.

Note that Section 4.2 and 4.6 in this chapter were adapted from my specialisation project in the autumn of 2022.

4.1 TOEFL11

The Test of English as a Foreign Language (TOEFL11, Blanchard et al., 2013) dataset was created specifically for the task of Native Language Identification (NLI) as a result of previous datasets having too much topic bias affecting the results of experiments. Topic bias is a result of uneven distribution of topics over the native languages, meaning that some L1s contain more occurrences of a given topic. Models based on these uneven datasets tend to learn words from these topics, instead of the subtle influence of the author's native language which we are actually interested in. TOEFL11 consists of essays, but is limited to only 8 topics, which are almost perfectly distributed across the native languages, resulting in a cleaner, more controlled foundation for the NLI task. It has remained one of the most used corpora because of its size and structure, in addition to its key role in the two shared tasks (see Section 3.2 and 3.3).

L1	FCE	Italki	Lang8	Reddit	TOEFL	Twitter
ARA	0	13899	1049	161033	1008	20480
CHI	66	14323	0	258033	1022	18809
FRE	146	5789	613	239541	1017	14371
GER	69	1554	527	257450	1025	21856
HIN	0	3276	80	182544	1041	13416
ITA	76	9754	658	166347	1007	13845
JPN	81	13975	168082	217844	1011	17671
KOR	86	8879	19772	92649	1022	11546
SPA	200	12964	2276	199617	1025	19124
TEL	0	0	13	0	1027	15809
TUR	75	5130	145	138622	1021	8725

Table 4.1: Distribution of documents per language for all datasets. The datasets are partially balanced across language — see Section 5 for information about this process.

4.2 Twitter11

The Twitter dataset (hereafter denominated as "Twitter11") was collected, filtered, and labelled by Edvardsen (2021). The dataset contains the same languages as TOEFL11, but differs in language type, document size, context, and topics. For example, Tweets are more informal, shorter, and contain more abbreviations.

The process of producing the dataset involved scraping Twitter, filtering out unusable tweets, and annotating them based on the geographical position of the poster. Only tweets where all of the poster's tweets were from the same country were used, to ensure a certain confidence of the labels. This resulted in a relatively large corpus with the same L1s as in TOEFL11, making it particularly suitable for cross-corpus NLI.

4.3 Lang8

Lang8 (Brooke and Hirst, 2011) stems from a language learner site by the same name, where learners are given the opportunity to post arbitrary texts in a second language, and have them corrected by native speakers. These texts, along with the author's annotated native language, comprise the dataset. Additionally, the corrected texts, marked with the specific errors are provided. The dataset can be noisy, for example mixing the L2 and the L1 of the author. It is furthermore rather unstructured with regards to topics, which can concern anything from the author's everyday life to topics of music, business, or, very often, language itself. In comparison to the very strict guidelines of TOEFL11, this dataset is less suitable for Native Language Identification, but may satisfy the conditions of a cross-corpus experiment, which are less focused on topic and more concerned with

the language itself. The website Lang8 was based in Japan, which is why Japanese and other Asian languages are over-represented, as can be seen in Table 4.1.

4.4 Italki-NLI

Italki is a very similar dataset to Lang8, and was gathered in the same fashion from a language learner site, and therefore shares many traits with Lang8, with the exception of it being more international, and therefore more balanced across L1s. The entire Italki dataset encompasses 349 languages, but the subset called Italki-NLI was extracted by Hudson and Jaf (2018), and contains the same languages as in TOEFL11.

4.5 FCE

The smallest dataset in this thesis stems from the First Certificate of English portion of the Cambridge Learner Corpus (Yannakoudakis et al., 2011). The texts are made up of two short answers to a prompt in the form of a letter. Most texts start with the phrase "Dear . . . ". The samples are tagged with both evaluation scores and errors, but these features will not be used in this thesis.

4.6 Reddit-L2

Reddit-L2 is an enormous corpus and was assembled by Rabinovich et al. in 2018. It is based upon users belonging to European forums (called "subreddits") who have tagged themselves with their nationality. The posts of these users are divided into posts made in the European subreddits and posts made in other subreddits the users belong to, which ultimately comprise two parts of the dataset: the in-domain data from the European subreddits, and the out-of-domain data from the other subreddits. The exact subreddit the post was made to is a separate feature to optionally be used for classification. Reddit-L2 is by far the largest corpus ever assembled for NLI, as can be seen in Table 4.1, though it should be noted that the average document length ($\sim$50) is far smaller than other essay-based corpora.

As the data is based on self-annotated tags made by individual users, the credibility of the data can be argued, but both Rabinovich et al. (2018) and Goldin et al. (2018) have taken measures to ensure that these labels are indeed correct. Goldin et al. additionally note that Reddit users are not English learners as in previous datasets, but more proficient English speakers, making them harder to classify in NLI. However, all levels of proficiency are free to post to Reddit, making it difficult to say anything absolute regarding the proficiency level. This may be something that could be explored in future research.

4.7 Reddit-L2.1

Reddit-L2.1 is a novel dataset gathered for this project in the same manner as in (Rabinovich et al., 2018). Reddit-L2 was mostly focused on European languages, since the source was labeled data from European subreddits, however, most nationalities around the world are represented in these forums. The purpose of Reddit-L2.1 was to fill in some of the remaining TOEFL11 languages that were not present in Reddit-L2, to create a more complete dataset suitable for cross-corpus experiments.

As in (Rabinovich et al., 2018), the scraping process involved using the Pushshift API (see Section 2.4.4), and retrieving posts and comments from the previous 3 years. These posts and comments were used as a foundation for extracting more content from other subreddits, also in the same manner as Reddit-L2 (for the out-of-domain subset). Contrary to the Reddit-L2, this dataset does not separate between data from EU-related subreddits and data from non-EU-related subreddits, because there did not exist enough pure EU content for a complete dataset, as non-European citizens are naturally less active in European forums.

A big challenge with this dataset is the languages that are spoken in multiple countries, and in multiple dialects — as is the case with Arabic — and the languages which stem from countries with multiple national languages, which is the case for Hindi and Telugu. The reason these languages were excluded in the original Reddit-L2 was exactly because of these challenges. There is not very much to do about this, which is why the integrity of this dataset is not entirely to be trusted and will remain a source of error.

For the case of Arabic, data has been gathered from Egypt, Algeria, Sudan, Saudi Arabia, Morocco, and Iraq. All of these have different dialects of Arabic, which most certainly will affect the results, but presumably not dramatically, as the different dialects are closely related. It should be noted that other datasets, TOEFL11, Twitter11, and such, have not made any clear distinction of dialects, and thus it will not be made in this work either.

Hindi and Telugu are by far the most problematic languages in the dataset. Telugu is mainly spoken in two Indian states, Andhra Pradesh and Telangana. In an attempt to gather Telugu samples, a manual set of 27 areas and cities related to these states was compiled, in addition to increasing the time interval. However, no such samples were discovered, meaning that this label is unrepresented in the dataset, which is made clear in Table 4.1. Furthermore, the decision has been made to categorise all Indian samples as Hindi, even though the author might speak other languages either instead of or in addition to Hindi. The reason for this is simply because Hindi is the most spoken language in India (Wikipedia contributors, 2023).

After extracting texts, the corpus was preprocessed, which is described in detail in Chapter 5 about the experimental setup.

5 Experimental Setup

This chapter will present the three experiments relating to the three research questions of this thesis, in addition to describing the shared preprocessing step and an Optuna (see Section 2.4.2) study. For each setup, some reasoning will be given, discussing why the particular setup was chosen. A common denominator for each experiment is the use of TOEFL11 and Twitter11 as test sets (see Sections 4.1 and 4.2). The reason for this is the underlying goal of evaluating both learner corpora and Social Media (SoMe) corpora, and the fact that these datasets are the only ones that are complete with regards to L1s. TOEFL11 is also the de facto benchmark dataset for testing a Native Language Identification (NLI) model. Another common denominator for all experiments is the use of transformer-based architectures, all of which are presented under the relevant sections.

Preprocessing

All corpora are rather different, both in terms of supervision and level of pre-processing. TOEFL11 is, for example, a very controlled dataset, with equal distribution of languages and topics as well as having been through some pre-processing. Other corpora like Lang8 (see Section 4.3) are rather disorganised in comparison, containing everything from mixed language to URLs. The goal of the pre-processing in this thesis was to adapt the corpora towards a more similar format. Therefore, all of them have undergone the same processing functions. The objective of most processing steps was to clean up unwanted and unhelpful text, an example of which is given further down where a sample is printed before and after it was processed. Elements like URLs, non-English characters, and apostrophes are removed.

The following steps were made to remove unhelpful text:

1. Removing URLs
2. Removing linebreaks
3. Removing special characters
4. Removing multiple spaces
5. Removing short and empty texts

Unprocessed:
http://www.scientificamerican.com/article.cfm?id=mind-reviews-how-many-

Processed:
According to this article, the maximum number of human relationships that our

friends According to this article, the maximum number of human relationships that our brain can keep track of at once is about 150. This article reminds me of the Japanese kid's song "Ichinensei ni Nattara(一年生になったら)" words by Mado Michio. This song says "I wonder if I could make 100 new friends when I become a first grader of the elementary school." The number of "100" is the exaggeration of the feeling that kids want to have a lot of new friend. Nowadays, we are able to have more than 100 friends on SNS such as facebook, Lang8 etc. However the more friends we have, the more longing for new friend we feel. It's like a materialism of friendship I suppose. So how many friends do you really need? Anyhow, please enjoy the song. The children were so cute, though the many kids didn't remember the second verse well, fu fu fu„,. http://www.youtube.com/watch?v=JyP0Khl–eQfeature=related

brain can keep track of at once is about 150. This article reminds me of the Japanese kids song Ichinensei ni Nattara words by Mado Michio. This song says I wonder if I could make 100 new friends when I become a first grader of the elementary school. The number of 100 is the exaggeration of the feeling that kids want to have a lot of new friend. Nowadays, we are able to have more than 100 friends on SNS such as facebook, Lang8 etc. However the more friends we have, the more longing for new friend we feel. Its like a materialism of friendship I suppose. So how many friends do you really need? Anyhow, please enjoy the song. The children were so cute, though the many kids didnt remember the second verse well, fu fu fu„,.

Removing non-English language

Polyglot, the workings of which are described in Section 2.4.6, was used to remove non-English language. The library detects the top languages found in a sample, or whether there is too much doubt to determine the language. If English clearly was not the primary language, the sample was removed.

Balancing across L1s

Apart from TOEFL11 and Twitter11, none of the corpora were balanced across native language. For instance, Lang8 is partial towards Asian languages, especially Japanese, and Reddit-L2 mostly contains European languages. This makes for unstructured experiments, but it should be noted that the experiments are predestined to be somewhat unstructured in any case: Firstly, the datasets contain different subsets of languages, which is not possible to remedy without extracting new data from each of the sources, which in most cases is not possible. Secondly, it could be argued that language, by nature, is too fickle to be analysed evenly. Korean is much closer to Japanese, than German, and it has shown to be much harder to separate Korean and Japanese in previous research. This

could suggest that there should be introduced more training samples for these languages, to make up for this challenge. However, most languages in the experiment are similar to at least one other, making this principle difficult to uphold in practise.

As a result of this, it is decided to only perform partial balancing, removing unnatural spikes and very big differences. An important element to this thesis is to train on large amounts of data, and thus it makes little sense to remove large amounts of it in order to achieve a perfect balance of languages, which by nature is an unbalanced domain. Additionally, another task is to train on multiple sources of data, to create robustness. When multiple datasets are combined, each with different subsets of languages and with different magnitudes of samples, the task of balancing becomes difficult. In different combinations of datasets, the distribution is different, which is a source of error that simply has to be alloted for.

5.0 Experiment 0 | Finding Hyperparameters

Optuna was used to find suitable parameter combinations. The experiment was set up using a combination of First Certificate of English (FCE), Italki-NLI, and Lang8 (balanced) as training data and TOEFL11 as test data (only 1000 samples).

Hyperparameter	Ranges	Discovered Value
Learning Rate	[1e-06, 1e-02]	3.5242e-05
Epochs	[3, 100]	100
Batch Size	[2, 32]	23

Table 5.1: The Optuna study in Experiment 0, with the inputted ranges and the discovered values.

Based on this the official learning rate was set a little lower to 2e-05, to reduce the chance of over-shooting maxima. The batch size, while ideally would be set to around 23, had to be reduced to 8, to reduce memory usage for the larger tasks in this work. Additionally, it is clear that the maximum possible amount of training is preferred in this task, which accounts for the epoch number of 100. However, this experiment used a rather small training set (FCE, Italki-NLI, Lang8), which resulted in a large number of epochs as the optimal epochs, and the same amount cannot be expected to be used for tasks involving the Reddit dataset, where the number of samples is vastly higher. Therefore, the number of epochs will be set at 3 for these experiments, which still represents a significantly longer training time. 3 epochs were chosen because similar research also has chosen 3 epochs (Edvardsen, 2021; Steinbakken and Gambäck, 2020). Table 5.2 shows the final setup.

Learning Rate:	2e-05
Epochs:	3
Batch Size:	8 (1 for multi-transformers)

Table 5.2: Experiment 0 Final Hyperparameters, which will be used throughout the experiments.

5.1 Experiment I | Finding the Best Data

As mentioned in the introduction, the focus of Experiment I is the impact of different types of data, and the type and amount of data will be of special interest. Similar data will tend to correlate better with each other, but the opposite is not necessarily true for dissimilar data. The semantics and syntax of language are, after all, mostly categorical and common in most circumstances, even though the topics and structure vary. In all these experiments, BERT will be used, as this is the best-performing simple transformer from previous research (Steinbakken and Gambäck, 2020). The combinations of datasets are presented in the table below. Experiment I.I and I.II mark a separation of learner and SoMe corpora, which will figure out the presumed optimal baseline for each of the test sets, where Experiment I.I might produce better accuracy for TOEFL11, while Experiment I.II might do the same for Twitter11. Experiment I.III will test whether more data is always better. The hyperparameters for these experiments were established in Experiment 0.

Exp	**EI.I**	**EI.II**	**EI.III**
Train	FCE	Reddit-All	FCE
	Italki-NLI		Italki-NLI
	Lang8		Lang8
			Reddit-All
Test	Twitter11	Twitter11	Twitter11
	TOEFL11	TOEFL11	TOEFL11

Table 5.3: Experimental Setup of Experiment I (EI).

BERT Classifier

The BERT Classifier uses the output of BERT as input to a separate classification layer on top, consisting of 11 nodes. Figure 5.1 shows a simplified illustration of this architecture.

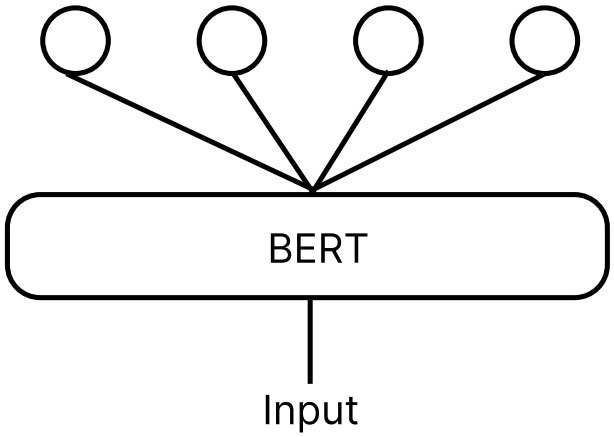


Figure 5.1: The BERT Classifier, here simplified with just four nodes instead of 11.

5.2 Experiment II | Finding the Best Model

The goal of Experiment II is to find the best performing cross-corpus model. The pool of models to be tested is based upon previous research papers. All transformer implementations are borrowed from the Huggingface Transformer library (see Section 2.4.1).

The models are:

1. BERT Classifier
2. Multi-BERT
3. Multi-GPT2

BERT was based on the results of Steinbakken and Gambäck (2020), while the multi-models are based on the results from (Lotfi et al., 2020). The BERT classifier is the same as in the previous experiment, while the multi-transformers are presented below.

Multi-BERT and Multi-GPT2

The multi-transformer, which in this work is nicknamed on the basis of consisting of multiple transformers, was devised by Lotfi et al. (2020). A general multi-transformer

with BERT is depicted in Figure 5.2, and the same setup is used for both BERT and GPT-2. As can be seen in the figure, 11 different fine-tuned models are used concurrently. The models are separately fine-tuned using only samples from one language. When making the prediction, each of the 11 transformer models is queried, and the most confident model is chosen. Lotfi et al. used GPT-2 for their experiments, which is part of reason why this particular generation of the transformer will be used. Additionally, GPT-2 is the latest generation that is available to the public. Some experimenting with other larger transformer models also showed them to prove problematic, because the memory size needed to handle them surpassed the available resources when 11 of these models had to be handled at the same time. Therefore the smaller GPT-2 had to be used. The multi-BERT model was inspired by a combination of the BERT classifier and multi-GPT2, with the hypothesis that this model would extract the best of the two techniques.

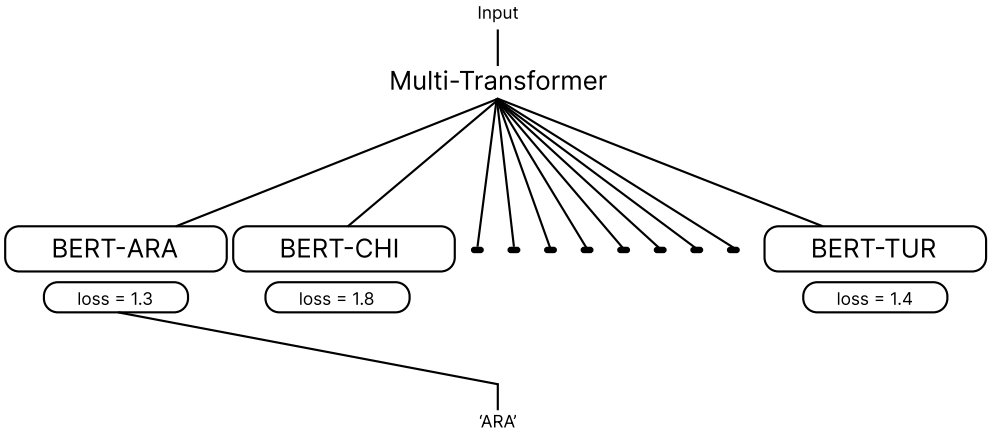


Figure 5.2: The Multi-transformer, here depicted with multiple BERT models.

5.3 Experiment III | Domain Adaptation

The different techniques of domain adaptation are explained in Chapter 2.3, where this is an adversarial-based supervised method. The source domain training set was chosen because of the results in Experiment I. Note that in this experiment, only TOEFL11 was used to test on, as the goal of this experiment was to reach the highest possible accuracy, and TOEFL11 provided generally higher scores, in addition to being closer to the domain of the optimal training subset, which was found in Experiment I. The following setup was used:

Source Domain:	FCE, Italki-NLI and Lang8
Target Domain:	TOEFL11
Test Set:	TOEFL11

Table 5.4: Experimental Setup of Experiment III.

Domain-Adaptive BERT

The domain-adaptive BERT model (see Figure 5.3) is very similar to the regular BERT classification model, with the addition of another domain classification head containing two nodes for the source domain and the target domain, respectively. The model is set to train on both these adverse tasks simultaneously.

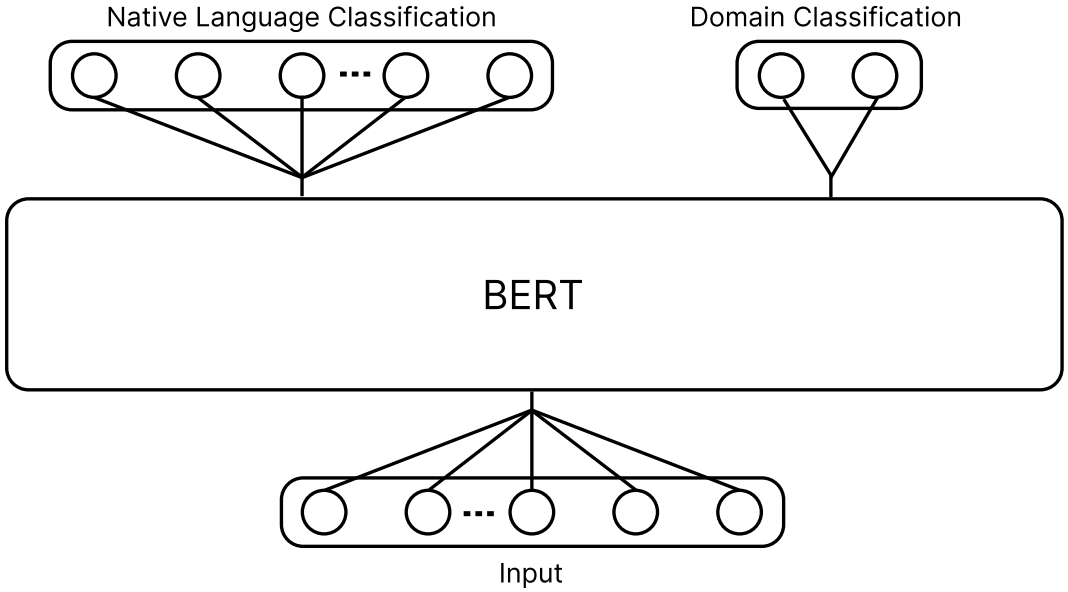


Figure 5.3: The Domaine-Adaptive BERT

6 Results

This chapter will give an overview of the results from the experiments described in the previous chapter. The results will be in the form of accuracy scores (see Section 2.5) and confusion matrices. While this work uses accuracy as evaluation for the reason of simplicity, interpretability and comparability, it is important to note that other research may utilise other means of evaluation, for example by using F1-score, cross-validation, etc. However, a large amount of related research uses accuracy either exclusively or in addition to these other methods, which further substantiates the choice of this particular evaluation metric.

6.1 Experiment I | Finding the Best Data

As can be seen from Table 6.1, the best subset of datasets was the learner corpora, which produced the topmost accuracy of 52.0%, using TOEFL11 for testing. Following this is the subset containing all available training data, which seems to follow the general pattern of a higher accuracy when testing on TOEFL11, and a lower accuracy for Twitter11. Another observation is that the Reddit data seems to be more detrimental than helpful when comparing Experiment I.I and I.III, as the only difference between the two is the addition of the Reddit data, which produces lower results. Additionally, the Twitter scores are generally much lower, except for when only training on Reddit data, which presumably is due to the Reddit data belonging to a genre closer to Twitter11 than the other training corpora.

6.2 Experiment II | Finding the Best Model

The, by far, best model is the BERT model. Note that the results of Experiment II.I are the same as Experiment I.I, as both the data and model overlap. It is also shown that the multi-transformers of this thesis did not perform as well as the one in (Lotfi et al., 2020), where the multi-BERT is particularly poor. Another observation is that the test scores for each of the test sets are far closer to each other for the multi-models, than with the BERT classifier.

Moving on to the confusion matrices, it can be seen that similar languages are generally hard to separate, where the model, for example, is a lot less confused between Hindi and German, than Hindi and Telugu. However, it can be seen in Figure 6.1 that

Experiment	Training Data	Accuracy	
		Twitter11	TOEFL11
EI.I (Learner Corpora)	FCE, Italki, Lang8	24.0%	52.0%
EI.II (SoMe Corpora)	Reddit-All	17.4%	15.9%
EI.III (All)	FCE, Italki, Lang8, Reddit-All	22.0%	49.9%

Table 6.1: Results of Experiment I (EI).

Chinese, for example, is one of the most accurate languages, in spite of being quite similar to both Japanese and Korean. In Figure 6.2 we can see that the multi-GPT2 model has major issues with German, Hindi, Telugu, and Turkish, which are generally underrepresented languages in the training set. Most of the predictions that should have been in these categories are incorrectly predicted as Chinese, Japanese, and Spanish, suggesting that the fault of the language model lies with the training data. In Figure 6.3 we can observe the obvious source of the multi-BERT's low accuracy score is the fact that the output prediction is almost always Japanese. All confusion matrices will be discussed in further detail in Chapter 7.

Experiment	Training Data	Accuracy	
		Twitter11	TOEFL11
EII.I (BERT)	FCE, Italki, Lang8	24.0%	52.0%
EII.II (Multi-BERT)	FCE, Italki, Lang8	9.9%	10.3%
EII.III (Multi-GPT2)	FCE, Italki, Lang8	21.5%	34.5%

Table 6.2: Results of Experiment II (EII).

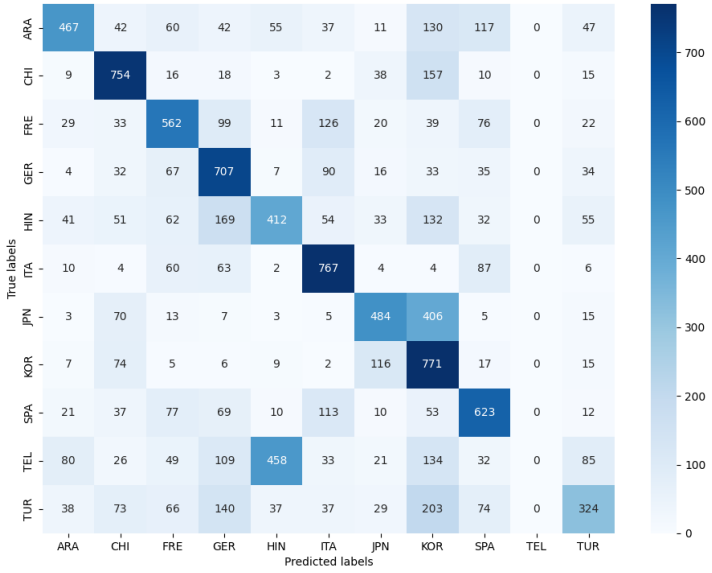


Figure 6.1: BERT Classifier confusion matrix when testing on TOEFL11.

Figure 6.2: Multi-GPT2 confusion matrix when testing on TOEFL11.

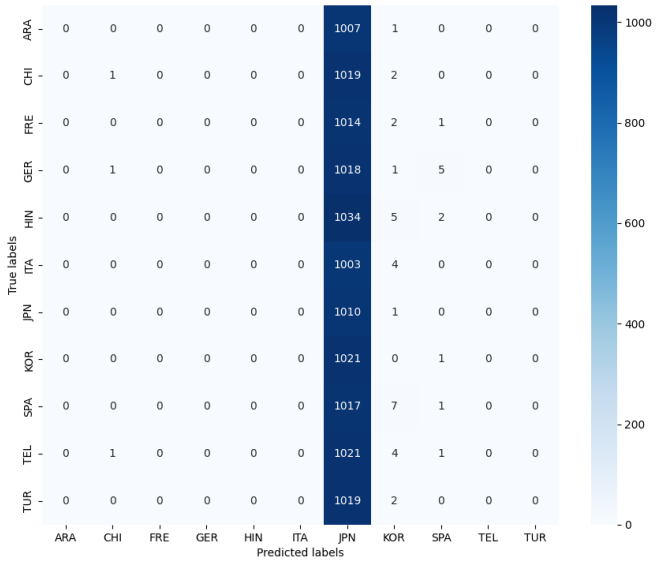


Figure 6.3: Multi-BERT confusion matrix when testing on TOEFL11.

6.3 Experiment III | Domain Adaptation

Experiment III is generally considered a failed experiment, as the model did not produce any higher results than the random baseline of 1/11th, as can be seen from Table 6.3. However, the reason why it failed will still be of interest, as this was generally thought to be the experiment that would produce the highest accuracy, and the opposite has been shown. In the search for good results, one would optimally know which direction to go, but it is useful still to know which direction not to take.

Experiment	**Training Data**	**Accuracy**
		TOEFL11
EIII (DA)	Source: FCE, Italki, Lang8 Target: TOEFL11	9.7%

Table 6.3: Results of Experiment III (EIII).

7 Evaluation and Discussion

This section will first present thoughts and evaluation of the results, discussing the underlying processes behind them, what might have affected them, and what impacts they have. The first three sections will take on each of the experiments, individually. Section 7.1 discusses Experiment I and focuses on how the nature of the data affects the general results, and how the specific subsets of training data affected the training process. Section 7.2 compares the different models in detail, while Section 7.3 shows some reasoning about why the domain-adaptive model failed. Following this, some general discussion of sources of error and challenges related to this work is given.

7.1 Evaluation of Experiment I | Finding the Best Data

Research question 1 *What data will produce the best Native Language Identification model?*

Experiment I was designed to find the most suitable combination of data for cross-corpus NLI when testing on different types of data. An overall observation of the work in this thesis shows that testing on Twitter11 provided much lower accuracies than testing on TOEFL11. This was to some degree expected, because, as pointed out before, tweets are harder to predict. The length alone makes for a smaller content foundation to build predictions upon, so this fact is to be expected. The Twitter dataset is also fundamentally different from all other datasets — including Reddit-L2 and Reddit-L2.1 — in this thesis. It was expected that the Reddit data would be closer to tweets, but the Reddit content still lacks the character limit that is set for Twitter content, and is thus generally longer. And so, perhaps the Reddit dataset is not as close to the Twitter data as first thought.

Another reason why tweets are harder to predict lies in the general level of proficiency of the writer. All datasets come from different sources and different demographics. Reddit posts and tweets are typically authored by relatively skilled English speakers, which are very often young adults. TOEFL11 and FCE are sourced by younger English learners, while Italki-NLI and Lang8 presumably stem from an older demographic who are trying to learn a second language sometime after completing their education. Older people are often less proficient in second languages, partially because they have greater difficulty absorbing new languages. On the other hand, the typical Reddit user is young and very proficient in English, which provides for language that is generally harder to predict. The reason why predicting this type of language is harder is because it contains fewer mistakes

and cultural-specific elements, on which the language model relies. As an author's English experience increases, so does the distance between their English communication abilities and their native language. Perfect English would be almost identical to that of a native English speaker, and almost impossible to detect the native language from.

It was expected that when trained on Reddit content, the Twitter tests would show better results. However, this is not the case when considering Experiment I.I and I.II, which respectively train on learner corpora and Reddit content. Experiment I.I produced a Twitter accuracy of 24.0%, while Experiment I.II produced an accuracy of 17.4%, as can be seen in Table 6.1 on Page 32. It should also be noted that the training time on Experiment I.II was significantly longer, because of the size of the dataset. Why then, when training on both a closer domain, for a longer time, are the results worse? This is most likely due to the general level of proficiency of Reddit users. The Reddit data also lack any samples for Telugu, which can be argued to decrease the accuracy by 1/11. Interesting still, is that Experiment I.II, while overall lower than Experiment I.I, still produces a Twitter accuracy better than TOEFL11. The reason might be because Reddit-L2 belongs to a genre closer to Twitter11, and as such, is not that detrimental to this dataset.

The best subset of data was found to be a combination of FCE, Italki-NLI, and Lang8. This may, as already pointed out, be because these datasets contain more mistakes and more typical language. Samples from Lang8 and Italki-NLI are not based on a specific prompt, so the author can discuss whatever they would like. Most often, they talk about their daily life – what they have done, the music they like, the street they grew up on, etc. – which are topics that might be more 'revealing' and contain more content-related words, which Volkova et al. (2018) showed to be one of the most important types of features.

7.2 Evaluation of Experiment II | Finding the Best Model

Research question 2 *Which transformer-based State of the Art model will perform best?*

Experiment II compares two of the best-performing models from previous research, and a novel combination of the two. The results show that the regular BERT model performs best in the test scenario, while the two multi-transformer models lack the ability to transfer information from the source domain to the target domain. This section will begin with some reasoning around why the single-transformer BERT performed better than the other multi-transformers, and subsequently move on to compare the two multi-transformers.

As neural networks are black boxes of information, it is difficult to pinpoint exactly why some models perform better. The expected behaviour for the multi-transformers was to focus their learning not on the classification of the training data, but on the language itself of each of the L1s, which one would think would make them better for generalising across other domains. It is likely that the models instead have overfitted

across the specific language of the training data and are not able to transfer this across to the test data. Perhaps the models require more, and more varied data. However, the only other available data was the Reddit-L2 and Reddit-L2.1, which as already seen, lead to poor results for cross-corpus evaluation. On the other hand, it would be interesting to observe how the different models behaved with all the subsets of training data, to see if the type and quantity of data would affect them differently.

It should also be noted that the multi-transformers performed more evenly for each of the two test sets. For example, the difference in test accuracy between TOEFL11 and Twitter11 is 13% for multi-GPT2, but 28% for BERT. This might suggest that the multi-transformer, while generally poor performing, is able to handle multiple genres and is, therefore, more general. On these premises, it could be thought that the multi-transformers are simply not trained enough or not properly tuned and optimised. After all, the same hyperparameters were used for all models, even though they were discovered using BERT experiments (from Experiment 0). The initial thought was for every model to have the exact same circumstances, but for the experiments to truly be equal, each of the models should receive its own hyperparameter optimisation. It should, however, be noted that the multi-models consist of not one but eleven transformers, and therefore require a particularly long time to train and tune, which is the reason why the same configuration and training time was allotted for them. Whether individual optimisation would close the TOEFL11 test accuracy gap of 17.5% between BERT and multi-GPT2 is not certain, but presumably, the more general properties of the multi-model would put it past the Twitter11 accuracy of BERT (24.0%).

When comparing the multi-GPT2 and multi-BERT models, it is obvious that the former is superior. It should be noted that the results from multi-BERT are so poor that the legitimacy should be questioned. However, the exact same method was applied to each of the models. A very large portion of the implementations is shared between the models, with the exception of the configuration schemes. Why then did the multi-BERT fail? First off, it can be seen from Figure 6.3 on Page 34 that the model only predicted Japanese. This is an obvious correlation with the fact that the training datasets, FCE, Italki-NLI, and Lang8, consist of very many samples from this category. As has been discussed in Section 5.0, the choice was made to not include too much balancing in the datasets to keep as much of the data as possible. This method worked perfectly fine for the other experiments but did not for the multi-BERT model. This might be due to the underlying design principles and capabilities of BERT: the model focuses on contextual language representation and is particularly well suited for extracting information and distilling texts into lower dimensional representations. This leads to the model extracting a lot of content, instead of language patterns, and the fact that BERT is so exceptional at this task might be the reason why it has overfitted to the Japanese samples. BERT has adapted so well to these samples, because of their large amount, and will consequently always give a lower perplexity score for them.

Another way of perceiving the outputs is that multi-GPT2 performed better because of its capabilities of capturing language patterns. It could be argued that the generative

nature of GPT-2 makes it more suitable for understanding language nuances and capturing the fine-grained distinctions that exist between different native languages. It is nevertheless of interest why the model did not overfit to the spiked Japanese category. The capabilities of GPT-2 is not only appropriate for capturing distinctions, but it could also be thought that the model will not overfit because of this: the model is more concerned with the language than the content, as opposed to BERT, and has consequently not been as affected by the content of the Japanese samples.

7.3 Evaluation of Experiment III | Domain Adaptation

Research question 3 *How does domain adaptation affect a transformer-based cross-corpus model?*

The domain-adaptive model was based on the BERT classification model, because it showed promising results in the previous experiments. The source datasets was chosen to be FCE, Italki-NLI, and Lang8, and the target was TOEFL11, because these belonged to the most similar domains, and because this was the best combination of data in previous runs. However, despite this, the model did only achieve a classification accuracy of 9.7%, which raises several points for discussion.

Adversarial domain adaption aims to bridge the gap between the source and target domain, through learning domain-invariant features. However, in the context of NLI, it appears that this approach was not effective. One reason for this might be poor hyperparameter settings. As the task is much more complex than the regular classification, it could be thought that the hyperparameterisation is much more sensitive. This might suggest that it was not advantageous to simply carry forward the hyperparameters from the previous experiments, and that a separate optimisation process should be conducted. Another reason is that there might not be too many common domain invariant features between the datasets. If the source and datasets are not sufficiently aligned, is could prompt the model to search for features that simply does not exist. However, this also raises the question of which types of features the other models in this thesis have learned, and whether they are all content based. Finally, it should be pointed out that the size of the source dataset should have been larger, because of the complexity of the task, as it requires more data to adapt to more difficult problems.

7.4 Discussion

Native Language Identification presents several challenges. Some are related to the nature of the task, while others relate to the data or the specific model implementations. This section will first and foremost give an overview of these challenges and sources of errors, and how they relate to the results.

The linguistic aspect of NLI is important. For example, the level of proficiency will

affect the results greatly. This was discussed in Section 7.1, where some insights were given on why more advanced English is more difficult to predict than erroneous and flawed English. To this, there is a cultural aspect as well: younger people are often more proficient than older people, and some countries and cultures have a greater tradition for learning English. For example, there is a greater tradition of learning English in Norway than in Korea. This might be the reason why the training subsets that only included datasets with a majority of Asian samples, like Lang8, gave generally better results. These results may be further emphasised by the fact that Reddit content is generally more confusing to predict. Another example of linguistic challenges is the fact that many people speak more than two languages, which further complicates the task. People from India, especially, know more than one language, and who is to say which language will affect their English the most?

This brings us to the problem of Telugu. Telugu represents many challenges of this thesis. Firstly, Telugu samples are difficult to come by: when gathering the data for Reddit-L2.1, no Reddit users had annotated their content with any of the 30 cities and regions that predominantly speak Telugu, leading to no Telugu samples among the Reddit data. In practice, this leads to Telugu never being the outcome of any predictions, even when using training data that at least included some Telugu samples. As can be seen from the confusion matrices on Page 33, the majority of Telugu samples are predicted as Hindi, which, while wrong, is not entirely wrong. As already pointed out, many Indians know more than one language, and most likely this language is Hindi. Additionally, Hindi and Telugu are somewhat related languages as well. While not belonging to the same language family, they still influence one another, through cultural similarities and geographical proximity. Most Indians learn the same English, which accounts for the confusion between the two languages. While this does not reflect on the accuracy scores, it is an important fact to consider when evaluating the results. Other L1s that are related are Chinese, Korean, and Japanese, which tend to be the most confusing languages to separate. Another language that poses a challenge is Arabic, mostly because Arabic is spoken in so many dialects. For example, when gathering the Reddit-L2.1 data and the Twitter11 data, multiple Arabic-speaking countries were gathered from, and while similar, the Arabic spoken in different countries can be very different, especially Arabic-speaking countries that are far apart, like Morocco and Egypt. This causes a very broad fundament to try and pinpoint accurate predictions. However, this is also a fact that is hard to work around – TOEFL11 does not disclose which Arabic country their samples were gathered from, and neither does any of the other corpora.

There are other challenges related to the datasets. As this is a cross-corpus study, content bias will be minimal, but not entirely diminished. The goal of cross-corpus evaluation is to reduce the content bias of the specific topics of the training corpora, for example, the topic words of the prompts made to generate learner essays. However, people from the same nationalities tend to talk about similar topics as well. Japanese people talk about Japanese politics and culture, while Spanish people do the same for Spain, etc. The language model is bound to rely on these topic words to some extent, which will affect the results. On the other hand, the subjects behind the data in this

research are free to use the words of their choosing, and to force them otherwise — for example, to not use words related to their country — would be a much greater source of error.

A challenge related to the model is that the input length is static, meaning that there are 512 input nodes. When a text is shorter than this, which both the Twitter and Reddit content very often is, only the left side of the model is trained. When encountering longer texts in the test set, the model will have weaker foundation to make predictions, because the right side is under-stimulated. It would be beneficial to investigate the performance across texts of varying length, and analyse the relationship between input length and accuracy. Additionally, a different padding strategy could be applied. In this project the texts have been padded to the right, but balanced padding, where the input text is padded on both sides resulting in centred inputs, would also remedy this problem. This is a technique that could be applied in future research, which will be further discussed in Chapter 8.

8 Conclusion and Future Work

This thesis has explored one of the most forthcoming Artifical Intelligence technologies in recent years, the Transformer, for the use of cross-corpus Native Language Identification (NLI). The study has focused on which transformer architectures have performed well under different circumstances. A divide between learner corpora and Social Media corpora has been explored and discussed. Additionally, some experimenting with domain adaptation has been conducted. The results show that the BERT classification model performed best with an accuracy of 52.0%, that the multi-transformers were lacking, and that the particular domain adaptation technique used in this research was unsuccessful. However, a large study has been conducted and discussed, and these contributions will be further summarised in the following section, along with some thoughts on how this study can be improved upon and how the field might evolve further in Section 8.2.

8.1 Contributions

In Experiment I, the aim was to find the best combination of data for cross-corpus NLI, which was found to be a combination of FCE, Italki, and Lang8, presumably because these contained more typical language and content-related words. The results additionally showed that testing on Twitter11 yielded lower accuracies compared to TOEFL11. This was expected due to the inherent challenges of predicting tweets, such as their shorter length and the higher proficiency of Twitter users. On the other hand, it was shown that the Reddit corpora, while generally resulting in low accuracies, proved beneficial for the Twitter corpora, because they belong to similar genres.

Experiment II compared different models, including a BERT classifier and multi-transformer models. The regular BERT model outperformed the multi-transformers, indicating that the latter struggled to transfer information from the source to the target domain. It was observed that the multi-transformers performed more evenly across different test sets, suggesting their ability to handle multiple genres. However, further investigation is needed to optimise and fine-tune the multi-transformer models for improved performance.

Experiment III focused on domain adaptation using the BERT classification model. Despite selecting the most similar datasets as the source and target domains, the domain-adaptive model achieved low classification accuracy. This could be attributed to suboptimal hyperparameter settings and the lack of common domain invariant features between the datasets. The size of the source dataset may have also affected the model's

adaptation to more complex problems.

These experiments represent a step towards exploring how transformers perform in NLI. While the overall goal has been achieved, not everything that was hypothesised at the beginning of the project came to be, and not all results were as favourable as one would like. It is crucial to point out that NLI is a challenging task, and accurately identifying a person's first language based on their communication in a second language remains a complex problem, a problem that is even harder when attempting to generalise across genres. And so, there still remains some work to be done, which will be discussed in the last section of this thesis.

8.2 Future Work

The findings of this thesis open up several avenues for future work. It is the hope that this thesis will operate as a stepping stone for further improvement within NLI. While some of the following points are very related to the particular experiments here, others will concern more general directions.

First off, there are numerous ways to improve the experiments conducted here. For example, multi-BERT would most certainly benefit from a more balanced training corpus, which might help overcome the bias towards just one native language. This could further be improved by proper optimisation, for example by adjusting the training time for each of the L1 models, as well as adjusting the hyperparameters. Tuning the models is something that would benefit most of the models in this thesis.

For Experiment III, it could potentially be useful to explore other techniques. In Section 2.3, explaining domain adaptation, other techniques were mentioned, all of which can be utilised in different manners. Future experiments could focus on applying these other methods to cross-corpus experiments, potentially in addition to attempting to optimise the domain adaptation experiment in this work.

Another improvement that could be done, is to address the data imbalances in this project. Some of the datasets were left somewhat unbalanced, to maintain a proper data amount. However, as has been seen with the multi-BERT model, this has greatly reduced the capabilities of the classifier. Balancing the training datasets could help mitigate biases and improve performance. The most pressing imbalance would be to gather more Telugu samples, as this is the obvious minority of the eleven native languages.

There are other ways of improving the data. For example, a few previous researchers Malmasi and Dras (2018) have constructed new datasets with fixed-length texts based on existing data. In this strategy sentences from different samples are concatenated to reach a specific length. Much of the data in this work uses quite short text, as is the case with Twitter11 and much of the Reddit data, which results in the models only receiving valuable input on the left of the input layer, resulting in an underdeveloped right side, which is detrimental for testing on longer texts. Potentially, an analysis on how the input length affects the accuracy could be conducted. Applying the fixed-length

technique would most likely lead to better results. Additionally, another solution is balanced padding, where the texts are padded on both sides, resulting in them being centred. On the other, this would still lead to weaker input nodes at either sides of the input layer, potentially making the solution of Malmasi and Dras the superior one.

Bibliography

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192.

Bjerva, J., Grigonyte, G., Östling, R., and Plank, B. (2017). Neural Networks and Spelling Features for Native Language Identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 235–239, Copenhagen, Denmark. Association for Computational Linguistics.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). TOEFL11: A Corpus of Non-native English. *ETS Research Report Series*, (2):i–15.

Brooke, J. and Hirst, G. (2011). Native Language Detection with 'Cheap' Learner Corpora. In *Proceedings of the First Learner Corpus Research Conference*, pages 37–47.

Brooke, J. and Hirst, G. (2012). Robust, Lexicalized Native Language Identification. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 391–408.

Brooke, J. and Hirst, G. (2013). Using Other Learner Corpora in the 2013 NLI Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–196, Atlanta, Georgia. Association for Computational Linguistics.

Bykh, S. and Meurers, D. (2014). Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1962–1973.

Chen, L. (2016). *Native Language Identification on Learner Corporas*. PhD thesis, University of Trento, Trento, Italy.

Cimino, A. and Dell'Orletta, F. (2017). Stacked Sentence-document Classifier Approach for Improving Native Language Identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 430–437, Copenhagen, Denmark.

Cortes, C. and Vapnik, V. (1995). Support-vector Networks. *Machine Learning*, 20(3):273–297.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Edvardsen, C. (2021). #NativeLanguageIdentification - Native Language Identification on a Novel Twitter Corpus Using Transformer-based Systems. Master's thesis, Norges Teknisk-Naturvitenskapelige Universitet, Trondheim, Norge.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2020). A Brief Review of Domain Adaptation. In *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pages 877–894.

Franco-Salvador, M., Kondrak, G., and Rosso, P. (2017). Bridging the Native Language and Language Variety Identification Tasks. *Procedia Computer Science*, 112:1554–1561.

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2014). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In *Proceedings of the 31st Second Language Research Forum*, pages 240–254.

Goldin, G., Rabinovich, E., and Wintner, S. (2018). Native Language Identification with User Generated Content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Brussels, Belgium. Association for Computational Linguistics.

Hudson, T. G. and Jaf, S. (2018). On the Development of a Large Scale Corpus for Native Language Identification. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories*, pages 115–129.

Ionescu, R. T. and Popescu, M. (2017). Can String Kernels Pass the Test of Time in Native Language Identification? In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–234, Copenhagen, Denmark. Association for Computational Linguistics.

Jarvis, S., Bestgen, Y., and Pepper, S. (2013). Maximizing Classification Accuracy in Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118, Atlanta, Georgia. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2022). *Speech and Language Processing (3rd Edition Draft)*. Prentice-Hall, Inc., USA.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.

Koppel, M., Schler, J., and Zigdon, K. (2005). Determining an Author's Native Language by Mining a Text for Errors. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, pages 624–629, Chicago, Illinois, USA. ACM Press.

Lotfi, E., Markov, I., and Daelemans, W. (2020). A Deep Generative Approach to Native Language Identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1778–1783, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Malmasi, S. and Dras, M. (2017). Native Language Identification using Stacked Generalization. arXiv:1703.06541.

Malmasi, S. and Dras, M. (2018). Native Language Identification with Classifier Stacking and Ensembles. *Computational Linguistics*, 44(3):403–446.

Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., Napolitano, D., and Qian, Y. (2017). A Report on the 2017 Native Language Identification Shared Task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark. Association for Computational Linguistics.

Markov, I., Nastase, V., and Strapparava, C. (2020). Exploiting Native Language Interference for Native Language Identification. *Natural Language Engineering*, 28(2):167–197.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Rabinovich, E., Tsvetkov, Y., and Wintner, S. (2018). Native Language Cognate Effects on Second Language Lexical Choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Rosenblatt, F. (1962). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. *Spartan Books, Washington DC*.

Stehwien, S. and Padó, S. (2016). Native Language Identification Across Text Types: How Special Are Scientists? *Italian Journal of Computational Linguistics*, 2(1):31–44.

Steinbakken, S. and Gambäck, B. (2020). Native-Language Identification with Attention. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 261–271, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A Report on the First Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use*

of NLP for Building Educational Applications, pages 48–57.

Tetreault, J., Blanchard, D., Cahill, A., and Chodorow, M. (2012). Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification. In *Proceedings of COLING 2012*, pages 2585–2602, Mumbai, India. The COLING 2012 Organizing Committee.

Uluslu, A. Y. and Schneider, G. (2022). Scaling Native Language Identification with Transformer Adapters. *International Conference on Natural Language and Speech Processing 2022*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Volkova, S., Ranshous, S., and Phillips, L. (2018). Predicting Foreign Language Usage from English-Only Social Media Posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 608–614, New Orleans, Louisiana. Association for Computational Linguistics.

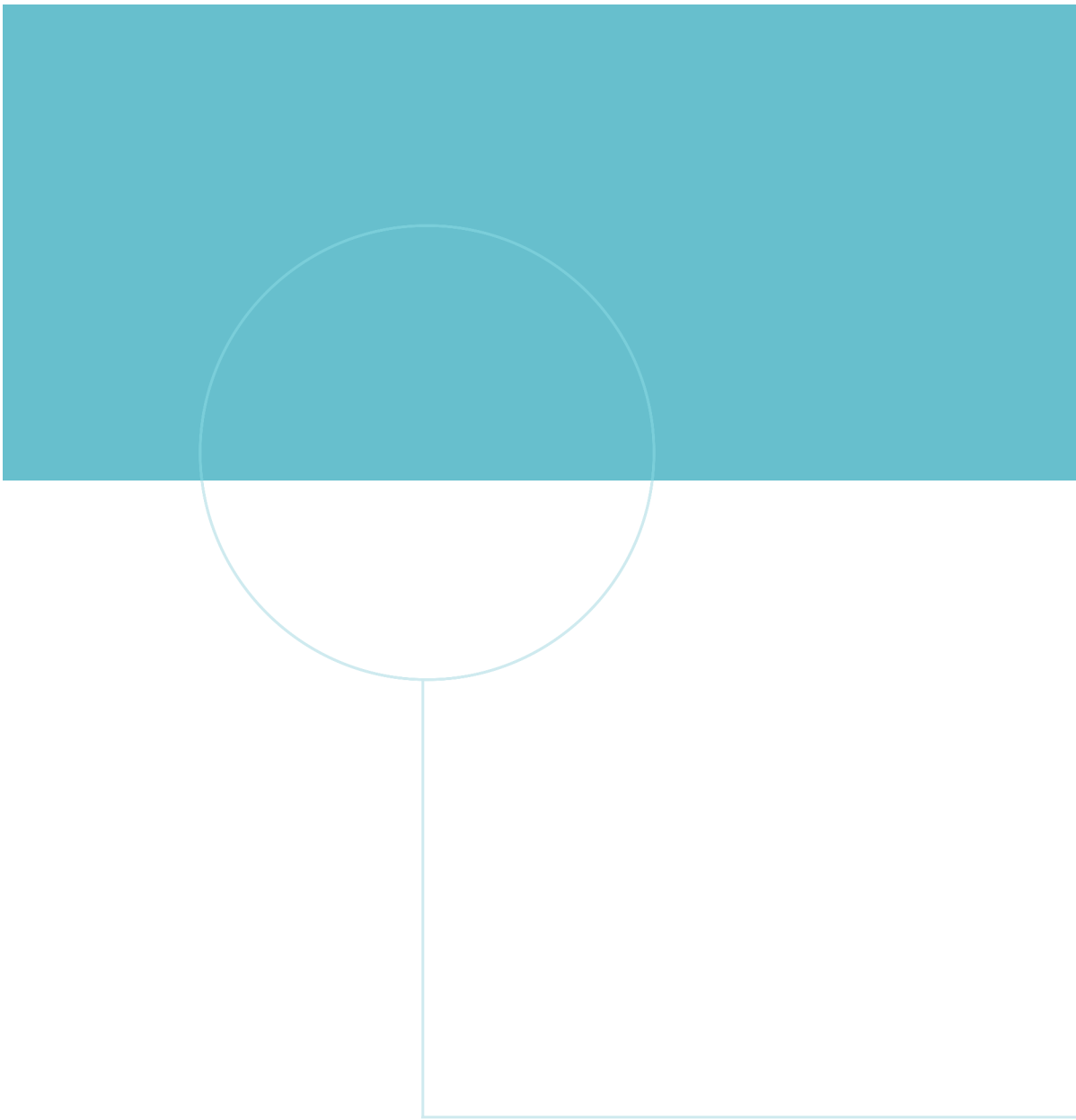
Wes McKinney (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Wikipedia contributors (2023). List of languages by number of native speakers in India — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_languages_by_number_of_native_speakers_in_India&oldid=1154270460. [Online; accessed 18-May-2023].

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Wong, S.-M. J. and Dras, M. (2011). Exploiting Parse Structures for Native Language Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.



NTNU

Norwegian University of
Science and Technology