HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildningsprogram – Degree Programme	
Humanistiska fakulteten	Magisterprogrammet i språklig diversitet och digitala metoder	
Opintosuunta – Studieinriktning – Study Track Språkteknologi		
Tekijä – Författare – Author Sabine Nyholm		
Työn nimi – Arbetets titel – Title Feature-based Transfer of Multilingual Sentence Representations to Cross-lingual Tasks		
Työn laji – Arbetets art – Level Magisteravhandling	Aika – Datum – Month and year 5 / 2020	Sivumäärä– Sidoantal – Number of pages 55

Tiivistelmä – Referat – Abstract

Universella meningsrepresentationer och flerspråkig språkmodellering är heta ämnen inom språkteknologi, specifikt området som berör förståelse för naturligt språk (*natural language understanding*). En meningsinbäddning (*sentence embedding*) är en numerisk skildring av en följd ord som motsvaras av en hel fras eller mening, speficikt som ett resultat av en omkodare (*encoder*) inom maskininlärning. Dessa representationer behövs för automatiska uppgifter inom språkteknologi som kräver förståelse för betydelsen av en hel mening, till skillnad från kombinationer av enskilda ords betydelser. Till sådana uppgifter kan räknas till exempel inferens (huruvida ett par satser är logiskt anknutna, *natural language inference*) samt åsiktsanalys (*sentiment analysis*). Med universalitet avses kodad betydelse som är tillräckligt allmän för att gynna andra relaterade uppgifter, som till exempel klassificering. Det efterfrågas tydligare samförstånd kring strategier som används för att bedöma kvaliteten på dessa inbäddningar, antingen genom att direkt undersöka deras lingvistiska egenskaper eller genom att använda dem som oberoende variabler (*features*) i relaterade modeller.

På grund av att det är kostsamt att skapa resurser av hög kvalitet och upprätthålla sofistikerade system på alla språk som används i världen finns det även ett stort intresse för uppskalering av moderna system till språk med knappa resurser. Tanken med detta är så kallad överföring (*transfer*) av kunskap inte bara mellan olika uppgifter, utan även mellan olika språk. Trots att behovet av tvärspråkiga överföringsmetoder erkänns i forskningssamhället är utvärderingsverktyg och riktmärken fortfarande i ett tidigt skede.

SentEval är ett existerande verktyg för utvärdering av meningsinbäddningar med speciell betoning på deras universalitet. Syftet med detta avhandlingsprojekt är ett försök att utvidga detta verktyg att stödja samtidig bedömning på nya uppgifter som omfattar flera olika språk. Bedömningssättet bygger på strategin att låta kodade meningar fungera som variabler i så kallade *downstream*-uppgifter och observera huruvida resultaten förbättras. En modern mångspråkig modell baserad på så kallad *transformers*-arkitektur utvärderas på en etablerad inferensuppgift såväl som en ny känsloanalyssuppgift (*emotion detection*), av vilka båda omfattar data på en mängd olika språk. Även om det praktiska genomförandet i stor utsträckning förblev experimentellt rapporteras vissa tentativa resultat i denna avhandling.

Avainsanat – Nyckelord – Keywords
sentence representation, cross-lingual transfer, feature-based transfer, natural language inference, emotion detection, transformers, natural language understanding, language technology, evaluation

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

University of Helsinki
Linguistic Diversity in the Digital Age
Language Technology

Master's Thesis

Feature-based Transfer of Multilingual Sentence Representations to Cross-lingual Tasks

Sabine Nyholm
013883984

Supervisor: Jörg Tiedemann 11.5.2020

Contents

1 Introduction

1.1 Focus and objectives

Natural Language Understanding has matured to the point where state-of-the-art models employ deep learning over large amounts of data, yielding highly generalised representations of language. There is a growing interest in scaling such systems to benefit not only various natural language processing tasks, but also handle such tasks in low-resource languages with comparably little data to contribute to the training stage.

While so-called distributed *embeddings* on word level have been around since the beginning of the millennium, sentence representation obtained with neural network architectures only began to emerge in the last few years. The ability for a natural language processing system to learn from an entire sequence of words is strongly motivated by the linguistic notion of connecting phrases to human "common sense" (Norman, 1972). As such, the extraction of meaningful linguistic information from phrases and sentences is a critical component of quality language understanding systems. As the generalisation capabilities of systems have accelerated, so has the interest in extending efforts to cross-lingual natural language understanding, or "XNLU". Noble humanitarian goals such as overcoming language barriers and enabling global information access underpin the motivations of scalability to all the world's languages, as noted by Hu et al. (2020) and others.

The emergence of multilingual approaches calls for suitable evaluation benchmarks. The practical component of this project is to integrate cross-lingual tasks into a new framework based on SentEval – an evaluation suite for assessing the universality of sentence embeddings, published in 2018 by Conneau

and Kiela. The successful integration would allow for evaluation of multilingual sentence encoders with a special focus on their language-independence. The transfer of the generalised knowledge encoded in the embeddings to other tasks or languages can be accomplished either by (a) using the representations as feature extractors in the task, or (b) by fine-tuning a pre-trained encoder on the task. These approaches are explained further in section 2.2.

The central questions this thesis aims to explore are:

- Can a feature-based approach to cross-lingual transfer learning feasibly be implemented as a SentEval adaptation?
- How do the results compare to an approach based on fine-tuning on similar data?
- How does a well-established multilingual model perform on a novel task within this framework?

The aim is to create an evaluation library encompassing cross-lingual NLP tasks with multilingual data-sets that multilingual encoders may be tested on. The chosen model for this purpose is the Hugging Face implementation of Multilingual BERT (Wolf et al., 2019), which is combined with the framework provided by the SentEval toolkit. With the practical framework in place, the hope is to succeed in extracting quality sentence embeddings from state-of-the-art models in a feature-based manner, and apply them to existing cross-lingual evaluation benchmarks as well as produce results for a novel cross-lingual task.

Originally a practical goal of this thesis project was to provide new cross-lingual benchmarks as a result; however, towards the end of development,

a much larger, more extensive multilingual multi-task benchmark was published by Hu et al. (2020). The scope of the thesis was narrowed over time and the main focus of the project shifted to the technical implementation of this extended toolkit, as as well as the integration of a novel task not yet included in existing multilingual benchmarks: fine-grained Cross-lingual Emotion Detection (XED) (Kajava, 2018).

This project received guidance and contributions as follows:

- Supervision by Jörg Tiedemann, Professor of Language Technology (University of Helsinki).
- Guidance by Aarne Talman (Doctoral Programme in Language Studies) and Alessandro Raganato (postdoctoral researcher of Language Technology), University of Helsinki.
- XED contributions by Emily Öhman and Kaisla Kajava (Doctoral Programme in Language Studies, University of Helsinki).

1.2 Outline of the thesis

This thesis has seven chapters in total.

Chapter 1 introduces sentence representation, motivates multilingual language modelling, and broaches the topic of evaluation benchmarks.

In chapter 2, the background and evolution of distributed representations in natural language processing is briefly described. The chapter gradually moves from embeddings of smaller units to embeddings that encode higher-level linguistic information.

Chapter 3 outlines the ways in which the quality of embeddings have traditionally been evaluated, and how the SentEval toolkit fits into the picture.

The multi-language data sets that were used for evaluation within MultiSent, the modified SentEval library, are presented in chapter 4.

The methods and modification work of SentEval is described in detail in chapter 5.

Results of the project are presented in chapter 6.

Lastly, chapter 7 is dedicated to closing discussion and conclusions.

2 Sentence representation

2.1 From word vectors to sentence embeddings

Distributed representations of words have been used extensively in the field of natural language processing (NLP) for decades. These representations, also called word embeddings, are arrays of real values that act as numerical representations of some linguistic unit, such as a word, sentence, or even document. In representation learning in general, an embedding can be seen as point in a n-dimensional space, where n is the number of aspects used to describe a particular data set. Each unit of the data can then be described with n different real values, effectively converting the unit to a mathematical representation that can be processed by machine learning systems such as neural networks. With distributed word representations, the idea is that each position in the vector denotes denotes some characteristic of the data. The properties and concepts associated with the data, however, end up diffused over the dimensions of the array in a way that is not clear or understood from a human perspective. That is, linguistic and other characteristics that inform the feature vectors are *distributed* over multiple dimensions, and each dimension connects to various concepts (Al-Rfou' et al., 2013). It is these values that allows for relating and connecting the linguistic units to each other on some semantic level, in accordance with features that have been specified or automatically learned, depending on the algorithm used.

The underlying assumption of the *distributional hypothesis*[1] is that units of language with similar semantic or syntactic meaning generally appear in

[1]The distributional methodology of linguistics can be traced back to Zellig S. Harris and his 1954 work, "Distributional Structure".

similar contexts, which corresponds to the resulting vector representations projecting the linguistic units to similar locations in the embedding space. Thus, given the assumption of shared contexts, the more meaning two units share, the closer their vector embeddings are to each other. This allows for computing similarity in meaning through mathematical means, such as measuring the Euclidian distance or cosine similarity between the vectors. Such vector space models have come to be used in various NLP and information retrieval applications, as well as parsing and other classification tasks (Pennington et al., 2014).

Automatically learning a distributed representation of words in the context of language modeling was first proposed by Bengio et al. in 2003. A standard approach today, leveraging neural networks to learn a language model emerged chiefly from this work. The authors used their models experimentally to formulate a probability function for sequences of words based on the representations they had obtained. This allowed the language model to take into account a wider context than previously popular n-gram models, which do not capture relationships between units and are thus ill-suited to generalisations. Notably, a departure from task-specific engineering was first suggested by Collobert et al. (2011), with their proposal that vector representation spaces be utilised to improve the performance of other language processing systems.

Word embeddings have since evolved to be more fine-grained, giving rise to evaluation schemes beyond simply measuring distances or angles between vectors. Building on the notion of distributional semantics, *distributional representations* of words specifically use their surrounding context to inform the resulting embedding. Mikolov et al. (2013) examine in their work the finer structures of the word vector space, observing semantic and syntactic

regularities through constant vector offsets. They show that subtracting the word vectors for *apple* and *apples* yields a result similar to that of subtracting the vectors of other words and their corresponding plural forms. The popular example equation of *King - Man + Woman = Queen* to illustrate the dimensions of meaning originates from this work. Mikolov et al. also point out the possibility of using semantically and syntactically meaningful representations in other, potentially unrelated tasks, highlighting their general-purpose potential.

Present-day word embeddings have generally been trained on corpora spanning billions of tokens using unsupervised machine learning methods. The task of embedding can be viewed as finding representations that predict a word based on its context, or the focus can be on leveraging corpus word occurrence statistics as the primary source of information. An example of the latter is GloVe (Global Vectors for Word Representation), a log-bilinear regression model proposed by Pennington et al. (2014) which captures global corpus statistics directly through the original co-occurrence matrix. The authors demonstrated the meaningful substructure of the resulting vector space and also analysed the model properties that give rise to such geometry. While still not entirely clear, the origin of regularities in the word representation space was generally not very well understood until this point.

Subsequent efforts brought about various neural word embedding approaches; Schnabel et al. examine existing evaluation schemes for them in their 2015 paper. The evaluation approaches are divided into one of two categories: *intrinsic* and *extrinsic*. The latter, which will be discussed further in the context of universal representations, involves using the embeddings as features in downstream tasks and measuring changes in the performance of some model. The downside of extrinsic evaluation is the specificity to a task, and

that it is unclear how this singular way of measuring the quality would connect to other measures. Intrinsic evaluations, on the other hand, used by for example Mikolov et al. (2013), involve queries that directly test the word representations for syntactic and semantic relationships. Such a query can be an analogy of the form

$$a \text{ is to } b \text{ what } c \text{ is to } \_?,$$

thus extracting a relation that can be either syntactic and semantic in nature, depending on how the terms are selected. Schnabel et al. drew attention to current problems in limited or biased query inventories due to re-purposing them from other fields, such as psycho-linguistics, and propose instead their own inventory better calibrated to corpus study.

With the evolution of neural network architectures, the earliest claims of meaningful phrase representations emerged approximately a decade later, chiefly within statistical machine translation. In the work by Cho et al. (2014), for example, the RNN Encoder-Decoder neural network model was introduced to compute the probability of a target sequence given a source sequence. Furthermore, the representations were shown to capture linguistic meaning: through qualitative analysis of embedding visualisations, Cho et al. noted clear clustering of syntactically similar phrases in some parts of the space and semantically similar in others. However, while capturing the relationship between words and phrases has by now certainly been shown to be possible, the connection between the representations and human semantic intuition has only been informally noted. As Hill et al. point out in their 2016 survey of state-of-the-art distributed sentence representation learning models, there is no obvious path from this observation to concrete strategies

of how to obtain the highest-quality or most useful sentence representations. That is, while the signs are clear, the question of what type of architecture and data to use remains a significant missing link in natural language understanding systems at large.

While there should no longer be any doubt that obtaining meaningful sentence embeddings is possible, some questions can certainly be raised about the nature of the representations. Because the models have become rather complex from a human perspective, it is difficult to tell what sort of linguistic information is present in the embeddings. Naturally, some of the difficulty lies in the fact that it is decidedly not always straightforward for humans to pinpoint what exactly makes a statement subjective or objective. One problem with previous techniques designed to find out what input sentence properties carry over to the embeddings has been that they are tailored to specific encoders. In the example by Conneau et al. (2018a), one such task might be to attempt to train a tense classifier on the embeddings produced by a pre-trained task, such as a long-short term memory (so-called LSTM) encoder for machine translation. If the classifier succeeds, the implication is that there is evidence of tense having been captured by the encoder to a certain degree. Thus, the ten probing tasks introduced by Conneau et al. do not plug into any specific encoder architecture - to maximise generality and facilitate the interpretation of results, they only require vector representations of sentences to be used. Existing sentence embeddings have already been shown to capture for example tense and number, and Conneau et al. indeed suggest that deeper investigation reveals a bigger and more detailed picture than previously assumed.

2.2 Transfer learning and universal representations

Machine learning has traditionally been rather narrowly focused, with data and methods being tailored and confined to a particular prediction task, such as classification or parsing. Over the past few years, however, efforts have been devoted to so-called *transfer learning*, in which the fruits of some system can be leveraged to improve a separate domain or task (Ruder et al., 2019). As sentence encoding has become increasingly feasible, some of the scientific community's interests have therefore shifted to so-called *universal* representations. These universal embeddings are meant to encode phrases on a higher level which are not only useful for a specific task or purpose, but generic enough to be able to improve any system that makes use of sentence understanding. A key point to universality is that it facilitates comparison of different encoders and approaches, which helps streamline the greater evaluation process and further development in the scientific field. Another central benefit is the mitigation of resource-scarcity for many languages: there is plenty of data and existing research available for English, especially, but such is not the case for most languages in use in the world today. With the development of transfer learning comes the possibility of being able to train high-quality models on high-resource languages and then use the knowledge gained to process languages with little or almost no data available whatsoever. The extension of transfer learning to these so-called *cross-lingual* approaches will be discussed further in the next section.

As mentioned above, task-specificity has traditionally been a practical requirement for obtaining high-quality representations, but as computing power increases and encoding schemes become more sophisticated, it seems natural that the direction should be to pursue representations that are generic enough

to transcend task borders and be useful in a wider variety of settings. In one of the early notable attempts at moving away from merely constructing sentence representations on the basis of word embeddings, Kiros et al. (2015) introduced an approach to obtain highly generic sentence representations that are not designed for any particular task. Drawing on the skip-gram vector learning model that uses a word to predict its surrounding context, Kiros et al. extended this approach to encoding sentences instead of words, calling the resulting representations *skip-thought vectors*. They also proposed in this work the idea of evaluating vectors by "freezing" the model and then using the encoder as a feature extractor in other tasks, which has since established itself as a standard method to measure universality.

Another factor that has fluctuated in the best-performing approaches of a given time is the degree of supervision to be used. With the evolution of machine learning and increased availability of language data on the web, unsupervised methods have prevailed over supervised ones in natural language processing systems for some time. Unsupervised learning typically makes use of large amounts of raw data to learn patterns automatically on the basis of the sheer amount of information. There has been some success in obtaining robust sentence embeddings through unsupervised approaches, although their wider applicability has been called into question (Conneau et al., 2017). In their recent work, Conneau et al. highlight the need for the scientific community to reach a consensus regarding standards and best practices for universal sentence encoding. The two main issues they raise concern firstly the type of neural network architecture that should be used and, secondly, what type of task and data the model should ideally be trained on. Conneau et al. explore sentence embeddings trained on the task of Natural Language Inference (NLI), which is concerned with determining whether a hypothesis

sentence is an entailment or contradiction (or neither) given a premise sentence. The motivation for circling back to a more supervised approach is the call for "deeper" language understanding to reach a higher level of generality, which the authors hypothesise is gained from the logical-semantic nature of the inference task. They also highlighted the usefulness of human-annotated quality data by demonstrating the lower volume of training data and less computing needed to achieve competitive results compared to unsupervised methods such as SkipThought.

Jointly with this work Conneau and Kiela also introduce the SentEval toolkit to assess the transfer learning and measure the universality of their representations, which will be discussed in greater detail in the chapter 3.

A common thread in various efforts at universality, however, is the avoidance of repeatedly reinventing the wheel: instead of starting the process from scratch with each language modeling task, it seems clear that the ability to share and build on knowledge already obtained is preferable. Transfer learning based on supervised learning is one way of realising to this notion, although success has also been achieved through unsupervised approaches.

As we have seen, pre-trained representations may be used as additional features in downstream tasks, but more recently the fine-tuning based approach has emerged in earnest. These techniques leverage a large amount of unlabelled data to obtain general language representations that may then be fine-tuned on specific NLP tasks. The Bidirectional Encoder Representations from Transformers (BERT) approach expanded on this two-fold method of pre-training and fine-tuning, in particular through the use of bidirectional encoding, as opposed to previous unidirectional techniques (Devlin et al., 2019). Their language modelling strategy is based on the masked language

model (MLM) objective, in which some units of the input are masked and the model learns to predict the identity of the masked tokens based on their context. A central concept of this method is thus to have both the left and right contexts inform the model simultaneously. Due to the multi-layered context and bidirectional processing, the token masking is introduced to prevent each word from seeing itself and therefore "cheating" out of the prediction step. The masking was done for 15% of input tokens, where a token chosen for masking would be replaced with [MASK] 80% of the time, a random word 10% of the time, and have a 10% chance of being unchanged.

Along with proposing BERT, a key contribution of their 2019 paper was demonstrating the importance of this bidirectionality. Notably, BERT achieved state-of-the-art performance on both token and sentence level tasks. The BERT approach to fine-tuning is to have each task initialised with the same pre-trained parameters, which are then adjusted using labelled data from the specific downstream task.

From the perspective of sentence representation, in addition to MLM, crucially the BERT model is trained on the sequence-level task of "next sentence prediction" (NSP). As BERT is designed to handle a variety of downstream tasks, it is built to process inputs that may consist of:

- a single sequence, e.g. sentiment classification, or
- a pair of sequences, e.g. the inference task's premise and hypothesis sentences.

Because dual input objectives look beyond the boundaries of individual sentences, including NSP in the training process is designed to allow the model capture the relationship not only between words, but also between sentences.

In this training procedure, given sentences A and B, 50% of the time sentence B follows sentence A in the corpus, whereas the other 50% of the time sentence B is a randomly chosen sentence from elsewhere in the corpus. Therefore, it is essential that the training data consist of large chunks of contiguous text, where the sentences are meaningfully arranged as a whole. Conveniently, any monolingual corpus with this property can be leveraged for this seemingly trivial language understanding objective; however, it demonstrably improves performance for downstream tasks such as inference and question answering. The model proposed in Devlin et al. (2019) was trained on English Wikipedia data as well as a books corpus, amounting to a total estimated 3,300 million words.

While BERT's general mode of application is through fine-tuning, it can also be used in a feature-based way. Fine-tuning is designed to specialise the model on particular features of the data and task at hand, but as noted by the authors, not all tasks are equally well suited to Transformer encoder architectures. Similarly to the approach proposed by Kiros et al. (2015), the parameters are fixed, and the encoder output can instead be extracted as features for the downstream task.

A multilingual version of BERT has since been released, which is the subject of evaluation in the MultiSent framework of thesis project. It is discussed further in the sections below.

2.3 The pursuit for multilingual sentence encoding

As with single-language representation, the work in multilingual language encoding started and has as of yet mainly stayed on sub-phrase level. In keeping with the idea that meaning should be mirrored in the vector repre-

sentation, the idea of a multilingual neural language model would be for the word equivalents of the different languages to appear in similar locations in the embedding space.

As multilingual word embeddings began to emerge through unsupervised methods, the free Wikipedia encyclopedia[2] became a popular source. Still a widely used resource in NLP today, benefits of Wikipedia include its language variety, quality text, accessibility, and continued growth (Al-Rfou' et al., 2013). Marking a notable effort for its time, Al-Rfou' et al. trained distributed word representations for 117 languages[3] using data from Wikipedia. They achieved competitive results when using their embeddings as sole features in a part-of-speech tagging task. Wikipedia is considered small from a present-day perspective, although it is still used for its suitability to multilingual applications (Grave et al., 2018). State-of-the-art *fastText* multilingual word embeddings spanning 157 languages have been obtained with data from Wikipedia as well as the Common Crawl web crawl data repository[4].

In their 2016 work on "Massively Multilingual Word Embeddings", for example, Ammar et al. project words of as many as fifty languages into a single shared vector space. They use some amount of supervision in the form of dictionary look-up and monolingual corpora, but do not rely on parallel data for the alignment of projections. In terms of transfer learning, such large multi-language resources can facilitate the handling of unseen words in cross-lingual tasks, such as machine translation, by seeking out the nearest vector-space neighbors of the unknown word.

Because sufficiently large amounts of data is not available for nearly all lan-

[2]www.wikipedia.org

[3]At the time, 117 languages had Wikipedia versions with at least 10,000 articles.

[4]commoncrawl.org

guages or their pairings - much less vast dictionaries or properly aligned corpora - reliance on parallel data can be a limiting factor. Mappings, i.e. typically linear transformations, can be learned between independently trained embedding spaces, but this method has been unavailable to source and target languages for which no bilingual dictionary exists. To address this problem, there have been attempts at creating shared word embedding spaces using little or no parallel data: For example, with their proposed self-learning framework, Artetxe et al. (2017) demonstrate that it is possible to obtain a competitive model with simple dictionary-mapping methods and very limited bilingual evidence in general.

In the case of sentence encoding, as discussed in the previous section, in recent efforts the applicability of the embeddings has largely been restricted to individual tasks - or at least to a single language at a time. It is generally recognised that to advance natural language understanding at large, models should be able to learn and generalise from as many languages as possible, including low-resource ones. There is also a practical need for language technology applications in as many languages as there are in use, but availability of language input data is highly biased toward a small number of high-resource languages, such as English.

Expanding on the borders of universal sentence representation and task-independent language understanding to cover multiple languages at once, *cross-lingual language understanding* (XLU) can still be said to be an emerging field. Because collecting sufficiently large amounts of data for all required languages is, if not impossible, at least infeasible, there is a growing interest in the scalability of models. If language systems can be tuned to handle unseen languages in some way, the need to directly rely on the available volumes of language data is diminished or even eliminated. In the ideal scenario, a

proper XLU model could indeed process languages not seen in training at all (so-called *zero-shot* approaches). The hope is that a multilingual model enriched with additional languages would ultimately outperform even highly fine-tuned, monolingual encoders.

There are, however, multiple open questions surrounding how this work should proceed, with some initial proposals guiding current efforts in the field. The following subsections discuss two major factors: the utilisation of machine translation and the degree of supervision.

2.3.1 Usage of machine translation

In state-of-the-art language modelling, one consideration is the degree of machine translation to be employed in training multilingual systems. While machine translation has evolved rapidly along with machine learning and NLP in general, and human translation labor is unarguably relatively costly, it has been argued that having automatic translation in the pipeline adds both computational intensity and potential inaccuracy (Conneau et al., 2018b). It is clear that on larger scales, it becomes impossible to assess whether the performance of a model suffers from poor machine translation quality, and to which extent. As such, overall accuracy is not only a function of the model's effectiveness, but also the machine translation system's performance.

Causes for translation errors are not limited to differences in syntactic structure or messy data. In their cross-lingual sentiment analysis study, Smith et al. discussed the impact of using machine translated data to classify subjective well-being in the social media domain. They found that source-language models significantly outperform machine-translated versions, in which an English model would be used to classify Spanish data translated into En-

glish, and vice versa. They conclude that cultural differences rather than language differences led to the translation errors and poorly captured sentiment in this case. They speculate that similar results can be expected for other tasks that require deep semantic understanding, such as emotion detection.

Apart from the concern of linguistic information lost in translation, including machine translation at the development stage is clearly more costly and resource-intensive than not.

2.3.2 Degree of supervision

For the purpose of modeling deeper language understanding, Conneau et al. advocate leveraging a feasible amount of supervision consisting of human annotation and translation. In what they describe as an alternative that is more elegant than machine translation approaches, Conneau et al. extend their natural language inference task encompassing a multitude of domains (MultiNLI)Conneau et al. (2017) to multiple languages by employing human-translated examples for higher-quality training data and making their encoders cross-lingual instead. The inference classifier is trained on English examples, and encoders in 14 additional languages are *aligned* with the English one using parallel corpora. Each encoder was then made to mimic the English encoder well enough that the classifier could still successfully process sentence pairs in any of the system's languages, thus bypassing the need for machine translation. The data-set created for the purpose of evaluating this model, dubbed XNLI (Cross-lingual Natural Language Inference), has since been used as an evaluation benchmark for further research in this vein. For the evaluation of transfer learning on the XNLI task, Conneau et al. released

sets of machine-translated training data for each of the 15 languages.

Another example of multilingual sentence representation that utilises supervision is the multi-language paraphrase detection system used for *Opusparcus* (Open Subtitles Paraphrase Corpus for Six Languages, Creutz (2018)). In this approach, only complete phrases - no sub-phrase units or structures - are used to determine the paraphrase likelihood of two sentences, and partially supervised annotation is used in the training process.

2.3.3 Deep pre-training

Some efforts at learning joint multilingual sentence embeddings have been made in recent years, such as the language-agnostic sentence representations ("LASER") by Artetxe and Schwenk (2019), who use a single BiLSTM encoder and shared byte-pair encoding vocabulary for all its 93 languages. With all the languages mapped to the same space, the idea is that a classifier can be trained on top with labelled data in only one language. The classifier can then be transferred to any of the represented languages without modification.

As mentioned in section 2.2, however, the BERT model by Devlin et al. (2019) has dominated the field of language understanding models since its release. An unsupervised, contextualised system, Multilingual BERT[5] (M-BERT) is pre-trained on large monolingual corpora similar to original BERT, but in 104 different languages as of this writing. To account for relatively well-represented languages like English, smoothed weighting was performed on the data to under-sample high-resource languages and over-sample low-resource ones.

Because multilingual representations have traditionally relied on parallel data

[5]https://github.com/google-research/bert/blob/master/multilingual.md

or bilingual dictionaries to successfully map inputs from multiple different languages into the same representational space, M-BERT drew a great deal of scientific attention upon its release. In particular, the M-BERT's degree of multilinguality has been studied by e.g. Pires et al. (2019), which is described further in the next chapter.

3 Evaluation

This chapter covers the evolution of assessment methods for sentence embeddings, starting from early evaluation methods and moving on to modern strategies of measuring deeper linguistic knowledge and universality. Lastly, the SentEval evaluation suite developed by Conneau and Kiela (2018) is described.

It should be noted that the notion of "quality" or "usefulness" associated with representations are somewhat nebulous terms. It is necessary to define the specific criteria that are used to measure the utility or quality of learned embeddings in a given setting. For example, they can be evaluated based on applicability to related NLP tasks, as discussed in section 2.2, which would be a measure of their generalisation capability. Alternatively, as discussed in section 2.1, the embeddings themselves can be targeted with techniques designed to probe for characteristics and concepts, shedding light on the ways in which linguistic properties may be encoded in the representations (Conneau et al., 2018a).

3.1 Traditional approaches

As pointed out by Conneau and Kiela (2018), there has been a lack of consensus on how to go about evaluating the quality of general-purpose word and sentence embeddings. The approaches can be roughly divided into two types: *intrinsic* and *extrinsic*.

Historically, intrinsic evaluation has meant human similarity judgements. As an example, Hill et al. (2015) especially argued for the importance of distinguishing similarity from mere association or relatedness when measuring

semantic similarity. By employing the efforts of 500 native speakers, they constructed the SimLex-999 resource, which is based on pairs of concrete concepts covering a range of grammatical functions, together with human similarity ratings. The resource was designed to highlight problematic areas in the performance of distributional models and guide the improvement of their architecture.

Extrinsic evaluation refers to indirectly assessing the quality by putting the representations to use and seeing how they perform. Some have argued that this is a preferable approach to intrinsic evaluation: Instead of focusing too much on lexical similarity scores, Nayak et al. (2016) suggested that word embedding performance on downstream tasks is more relevant and a better way forward. They propose evaluation a set of standardised downstream tasks to (1) increase real-world applicability and (2) get a more fine-grained view of a model's strengths and weaknesses, such as whether they're syntactic or semantic in nature.

Whichever the approach, the literature in the field of representation learning in NLP was abundant with calls for a unified evaluation methodology by this time.

3.2 Measuring universality

As touched upon in various points of this thesis, the idea of transfer learning is that a model is able to capture some general-purpose knowledge that can be applied elsewhere, in a different but related task. For example, if the embeddings obtained through the inference task (Conneau et al., 2017) capture high-level linguistic information that it encodes in the representations, they should improve the performance of a task that benefits from deep

sentence-level understanding, such as sentiment classification.

Other examples of downstream tasks include:

- **inference** – are two sentences logically entailed?
- **paraphrase detection** – are two sentences paraphrases of each other?
- **subjectivity detection** – is a sentence subjective or objective?
- **bitext mining** – identifying parallel sentences in comparable corpora

The goal of *universality* in transfer learning is that the obtained representations are generalized enough to be applicable to a wide range of tasks. As mentioned in section 2.2, in state-of-the-art language modelling, there are two main approaches to transferring knowledge from one system to another: *feature-based* and *fine-tuning–based*. Peters et al. (2019) compare these two methods for BERT and ELMo (Peters et al., 2018), finding that in practice, the best-performing approach seems to mainly depend on the similarity of the pre-training and target task.

3.3 Assessing multilinguality

Due to its relatively recent emergence, the rigorous study of deep cross-lingual language modelling is still in its infancy. *Cross-language transfer* generally refers to the ability of a system to apply learning from one language to another. In the work by Conneau et al. (2018b), for example, an inference classifier is trained on English representations, and the transfer is applied by aligning encoders in various other languages with the English one (using parallel corpora) in order to represent the foreign languages in a way that can

be successfully handled by the English classifier. In this specific study, this multilingual encoding approach did not outperform techniques that involved directly translating the samples before encoding them, although the results were deemed promising.

The BERT model introduced in section 2.2 has been the subject of the scientific community's interest also with regard to cross-lingual transfer. Following the zero-shot idea of testing a model on languages not used at the training stage, Pires et al. (2019) found that the model is clearly capable of transfer across languages, but with certain restrictions. For example, zero-shot transfer worked to some extent even between languages with different scripts and thus zero lexical overlap, but there was no sign of M-BERT learning to systematically accommodate for typological differences such as word order. It should, however, be noted that these properties were present in the model without training objectives specifically encouraging multilinguality. Pires et al. theorise that words from different languages end up close in the representation space indirectly due to globally present tokens such as digits and URLs, which spread this effect to other nearby words as well.

While zero-shot capability was initially an interest guiding this thesis' aims, extensive testing for multilinguality as well as accompanying hypothesising around language-independence was ultimately beyond the scope of this project.

3.4 The SentEval toolkit

Because of the lack of a central framework, evaluation pipelines had been created separately for each project, often with small sets of data and individually tuned hyperparameters (Conneau and Kiela, 2018). The work by Hill

et al. (2016), cited in the background section 2.1, is one example of setting up a particular framework for systematically comparing different encoders on a multitude of tasks. As noted in the SentEval paper, not only is this unnecessarily cumbersome, but results obtained from largely similar experiments are not necessarily wholly comparable. To address these issues and facilitate evaluation, they introduce the SentEval toolkit[6]. It should be noted that in the later stages of this project, such a cross-lingual benchmark and generalisation evaluation framework was released by Hu et al. (2020). This XTREME benchmark[7] includes sentence classification on the XNLI task, among others.

The purpose of SentEval is to provide a user-friendly means of assessing the generalisation capability of sentence representations by using them as features in other, so-called *transfer tasks*. As outlined in the section discussing the measuring of universality, the two main approaches to transfer learning today are (1) feature-based and (2) fine-tuning. The SentEval suite in itself is mainly a tool to evaluate the quality of the sentence embeddings themselves, regardless of how they were obtained. The authors express hope that their toolkit will be used for centralised and comparable evaluation by the community, especially in the vein of probing for a clearer picture of how linguistic information is encoded.

In the experiment related to this project, the M-BERT embeddings were obtained through a feature-based procedure, which is described in further detail in chapter 5. The cross-lingual transfer tasks integrated into the framework are presented in the next chapter.

The basic use case of SentEval consists of the following:

[6]https://github.com/facebookresearch/SentEval

[7]https://github.com/google-research/xtreme

1. Implementing a function that encodes sentences from the task data into embeddings, one batch at a time

2. Optionally adjusting SentEval and classification parameters

3. Optionally pre-processing the samples by constructing their word vectors (needed for bag-of-word-vector techniques)

4. Specifying with transfer tasks to use for evaluation

For the first step, the user can either implement a sentence encoder in any Python framework of their own choice, or make use of the example scripts provided in the repository. These include the InferSent[8] inference-based encoder released along with SentEval (Conneau et al., 2017); GenSen[9] based on large-scale multi-task learning; and Google's Universal Sentence Encoder.[10] There is also an option to simply evaluate the quality of the average of word embeddings. Included by default is the option to construct the word vector vocabulary with either GloVe or fastText.

Once the sentence representations are obtained, SentEval uses the vectors as input to classifiers trained on top. The user may choose between logistic regression and a Multi-Layer Perceptron for the classifier type. In the case of the Semantic Textual Similarity task, however, the cosine distance between two sentence representations are evaluated against human similarity judgements for a correlation score. (Both of the cross-lingual tasks used in this thesis project are classification-based.)

SentEval also provides a script for automatically obtaining the task data from their known locations and applying pre-processing such as tokenisation

[8]https://github.com/facebookresearch/InferSent

[9]https://github.com/Maluuba/gensen

[10]https://tfhub.dev/google/universal-sentence-encoder/4

to it.

In terms of additional Python libraries, SentEval's functionality requires NumPy, SciPy, PyTorch and scikit-learn.

4 Data

Because the original SentEval version included data and tasks in English only, the main objective was to find suitable cross-lingual tasks and build their data into the SentEval framework, resulting in a new, multilingual library of tasks. The chosen tasks – XNLI (Conneau et al., 2018b) and XED (Kajava, 2018) – represent the objectives of inference and emotion detection, respectively. These can be seen as a more advanced type of sentence classification that require relatively deep natural language understanding for a model to successfully accomplish. They also represent both single and dual sentence input: for emotion detection, one sentence at a time is processed, whereas entailment classification comprises a pair of sentences and the semantic-logical relationship between them.

4.1 XNLI: Cross-lingual Natural Language Inference

The task of Natural Language Inference involves taking two sentences as input and deciding whether they are *entailed*, *contradictory*, or neither (*neutral*). The pair of sentences is referred to as the *premise* and the *hypothesis*. The Stanford Natural Language Inference corpus was first introduced as a sufficiently large high-quality NLI corpus for modern language understanding systems by Bowman et al. in 2015. The SNLI corpus consists of over 570 000 sentence pairs written and labeled by humans. A crowd-sourced corpus modelled on SNLI was later released which spans multiple genres of both spoken and written text. The resulting corpus comprises approximately 433 000 sentence pairs and is dubbed MultiNLI.[11] For each premise sentence,

[11]https://cims.nyu.edu/ sbowman/multinli/

there is a corresponding hypothesis representing each of the inference labels. As an example from the XNLI corpus, given the premise of `He didn 't get to go .`, the hypothesis sentences are as follows:

- Entailment: `He wasn 't allowed to attend .`
- Contradiction: `He was the first to be invited and enjoyed the experience .`
- Neutral: `He wasn 't allowed to go to the museum 's opening .`

The multi-language XNLI set was similarly constructed by creating 7,500 human-labelled development and test samples in 15 different languages, making for a total of 112,500 sentence pairs. It is specifically designed to function as a benchmark set for cross-lingual language understanding, and has since been actively utilised in related research. For example, it is included as one of the nine tasks in the recently published XTREME benchmark (Hu et al., 2020) mentioned in section 2.3. Notably, the English sets are not copies or derivatives of the MultiNLI or SNLI data, but newly created for the XNLI task and professionally translated into each of the languages. Because the labels were copied over from English as such, the preservation of meaning in translation is acknowledged as a potential issue. Measures were taken to ensure that such discrepancies were rare enough to be negligible.

The XNLI languages are: Arabic (ar), Bulgarian (bg), German (de), Greek (el), English (en), Spanish (es), French (fr), Hindi (hi), Russian (ru), Swahili (sw), Thai (th), Turkish (tr), Urdu (ur), Vietnamese (vi) and Simplified Chinese (zh). The motivation behind the particular set of languages is to invite diversity over a range of typological metrics, such as language family,

syntactic structure, and writing system. Among these, Swahili and Urdu are cited as low-resource languages (Conneau et al., 2018b).

4.2 XED: Cross-lingual Emotion Detection

Emotion detection can be seen as a subgroup of sentiment analysis, which is a buzzing area of classification in NLP. Extracting opinions and predicting behaviors of people based on the data they generate is valuable from various commercial and societal perspectives. Because sentiment analysis is generally done on phrase level, as a task it is well suited to testing language models for their semantic understanding. The emotion-annotated data used for this project was adapted from Kaisla Kajava's Master's Thesis titled "Cross-Lingual Sentiment Preservation and Transfer Learning in Binary and Multi-Class Classification" (University of Helsinki, 2018). Some of the central findings include that sentiment is generally preserved in translation well enough at least for binary classification, although reliable multi-dimensional classification requires a larger amount of data. Only the fine-grained multi-class component was considered for this project. The data set was dubbed Cross-lingual Emotion Detection (or "XED" for short) for the purpose of its usage as a downstream task.

The data originates from the Open Parallel Corpus of automatically aligned movie subtitles (Tiedemann, 2012). English was chosen as the source language; the data in the other languages are translations thereof. The data was manually filtered for misalignments, rendering mistakes, and other noise. The original target languages are Finnish, French and Italian. Only the latter two were included in this experiment due to difficulties with obtaining enough data in a suitable format. Each sample has been human-labelled

XED: Class distribution

	ang	ant	dis	fea	joy	sad	sur	tru
Frequency (%)	14	13	13	11	15	11	10	13

Table 1: Proportion of samples within each emotion class.

as expressing one of eight emotions based on Plutchik's theory of emotion: *anger*, *disgust*, *fear*, *sadness*, *anticipation*, *joy*, *trust*, and *surprise* (Kajava, 2018).

The distribution of classes is not even, but it is stratified across the splits of the data set, meaning that the proportions of samples belonging to each class is the same within each split. See table 1 for the approximate frequency distribution. In the sets used for this project, there is minor variation (± 0.5 points) in the frequencies due to the random extraction of the development sets from the training sets.

5 Methodology

The following sections describe the components of my experiment: the environment I used, and the data sets which the models were trained and tested on. Detailed are also the modifications made to the SentEval fork, which was entitled *MultiSent*.[12]

5.1 Objectives

The chief goal of the project has been to evaluate multilingual encoders on cross-lingual tasks by setting up a reusable evaluation framework modelled on the SentEval toolkit, which is designed to evaluate sentence embeddings for universality Conneau and Kiela (2018). Originally the tasks considered for transfer in MultiSent included Opusparcus (Creutz, 2018), although they were ultimately narrowed down to XNLI and XED. The practical integration of these tasks is detailed in sections 5.4. Another initial goal was to train a multilingual model from scratch, or fine-tune a model on multilingual data; however, due to time constraints, the pre-trained Multilingual BERT model was evaluated in a feature-based manner (outlined in section 5.3).

The work started with studying current language understanding tasks and their feasible extension to multiple languages, with SentEval's architecture as the template. The practical work consisted of the following main steps:

- Adjusting the MultiSent interface to allow selection of languages to include in the transfer evaluation
- Setting up a data structure for each transfer task and adjusting the

[12]https://github.com/Helsinki-NLP/MultiSent-Benchmark

SentEval engine accordingly

- Combining a suitable Multilingual BERT implementation with MultiSent
- Implementing a multi-language CBOW baseline encoder

5.2 Computing setup

MultiSent is part of the Language Technology at the University of Helsinki research project.[13] The storage, processing and computing was conducted on the IT Center For Science (CSC) Puhti supercomputer.[14] Running the models and evaluation was generally done on Puhti's GPU nodes, which feature NVIDIA Volta V100 GPUs.

The BERT implementation used for this project is provided by Hugging Face Transformers[15], which is an extensive library of natural language understanding architectures. Hugging Face Transformers library contains PyTorch and Tensorflow implementations, usage scripts, and a host of state-of-the-art models that can be trained from scratch or imported with pre-trained weights.

5.3 Model: Multilingual BERT

Preexisting SentEval examples use monolingual encoders and tasks to obtain and evaluate sentence embeddings. Because the goal was set on cross-lingual evaluation, an implementation of a pre-trained multilingual model was fused

[13]https://github.com/Helsinki-NLP
[14]https://docs.csc.fi/computing/system/#puhti
[15]https://huggingface.co/transformers/index.html

with SentEval for the purpose of this thesis. Per the recommendation of Devlin et al. (2019), the latest cased version of multilingual BERT-base was used.

- `BertTokenizer` – tokenises the sentence and converts the tokens to IDs
- `BertModel` – encodes the sentence and outputs the embedding

The idea for obtaining a sentence embedding is to process the sequence with a BERT model and extract the hidden state of the last layer as a representation for the sentence. There are multiple ways to approach this objective depending on task at hand. For example, in the work by Pires et al. (2019) dedicated to probing BERT, the hidden feature activations at each layer are extracted, and the representations for the input tokens are then averaged to obtain a vector representing the entire sentence. While there is no clear-cut method for obtaining the embedding that best encompasses the semantic information, there appears to be a rough community consensus that much of the sentence's content is efficiently summarised in the output of the initial token, especially in the case of classification.

Obtaining the embeddings was first attempted according to the Hugging Face Quickstart tutorial[16] and by averaging over the word representations using the `torch.mean` function. Due to unknown causes, the encoding of the sentences proved to be impractically slow for proper testing purposes with this method. With more time to debug, the issue could likely be resolved, but for the the sake of this thesis, a workaround was attempted instead. The encoding scheme (`batcher()` function) is shown in figure 1. The strategy was mostly adapted from the instructions by Jay Alammar.[17]

[16]https://huggingface.co/transformers/quickstart.html

[17]https://github.com/jalammar/jalammar.github.io/blob/master/notebooks/bert

Figure 1: The M-BERT encoding implementation in MultiSent

```
def batcher(params, batch):
    batch = [sent if sent != [] else ['.'] for sent in batch]
    tokenized = []
    max_len = 0

    # Tokenize and prepare the sentences
    for sent in batch:
        tokenized_sent = tokenizer.encode(sent)
        tokenized.append(tokenized_sent)

    max_len = 0
    for i in tokenized:
        if len(i) > max_len:
            max_len = len(i)

    # Make all inputs same length and compile into single batch
    padded = np.array([i + [0]*(max_len-len(i)) for i in tokenized])
    mask_padded = np.where(padded != 0, 1, 0)

    # If model is put on CUDA, put tensors there as well
    input_ids = torch.tensor(padded).to('cuda')
    mask = torch.tensor(mask_padded).to('cuda')

    with torch.no_grad():
        encoded_layers = model(input_ids, attention_mask=mask)

    # Get the hidden state of the initial [CLS] token
    embeddings = encoded_layers[0][:,0,:].cpu().detach().numpy()

    return embeddings
```

To speed up computation, the text input to the model is given as a single batch rather then looping over each sentence separately to encode it. Once the sentences are tokenised and converted to their vocabulary IDs, the sequences are batched together into an array before being fed as input to the model. As per the Transformers documentation, because the sequences within the array might be of different lengths, each sequence in the array should be padded to the length of the longest sequence in the batch. Additionally, an attention mask[18] may be passed to the model to indicate the positions of actual tokens it should pay attention to. As such, an array mirroring the sentence array is created, where each position that corresponds to a token ID is denoted by the value 1, and otherwise the value 0. The arrays are then converted to tensors and given as input to the model, after which the pooled hidden state described in the previous paragraph is extracted. The array of embeddings can then be passed further on to the MultiSent downstream tasks.

5.4 Downstream tasks

The transfer tasks make use of the data described in detail in chapter 4. This section describes their specific integration into the MultiSent suite. For the most part, the data was structured so that minimal modifications to existing SentEval files would have to be made. The aim was also to keep the data as modular as possible in order to (1) make it straightforward for the user to include or exclude languages based on their specific needs, (2) later extend the downstream tasks with data in additional languages in a straightforward way.

[18]https://huggingface.co/transformers/glossary.html#attention-mask

XNLI The XNLI training, development, and test sets are included as provided by Conneau et al. (2018b). As mentioned in chapter 4, the dev and test sets are not based on MultiNLI, but newly created and professionally translated from the English XNLI sets to the other 14 languages, for a total of 7,500 samples per language. The *training* sets released to the public are, however, machine-translated. The "Translate-Train" baseline in the XNLI paper was obtained using these sets, which were automatically translated from the MultiNLI training data into each of the other 14 languages using neural machine translation. As such, some of the classification accuracy necessarily depends on the quality of the machine translation. Because the translation is done at the training stage, a classifier is trained and tested separately for each language as part of the evaluation pipeline.

The evaluation script designed for the monolingual NLI task, `snli.py`, was then adapted to simply use the XNLI data instead. For this purpose, the XNLI data was divided into separate folders on a per-language basis. In the `data/downstream/XNLI` directory, for each language `$lg`, there is a subdirectory `XNLI-$lg` with identical train, dev and test file structures mimicking the SNLI transfer task structure. The premise sentences, hypothesis sentences, and gold labels are simply split into their own files, with lines aligned accordingly. That is, the first line of `labels` contains the label for the entailment relationship between the first line of `s1` (premise) and the first line of `s2` (hypothesis). The total number of samples is outlined in table 2.

Regarding the files, the entailment labels included in the machine-translated training data deviated from the dev and test data in that the term "contradictory" was used instead of "contradiction". This was corrected as part of the pre-processing.

XNLI data	train	dev	test
number of samples	392,702	2,490	5,010

Table 2: Number of lines per file in the XNLI data.

XED Because SentEval already has a fine-grained sentiment classification task built in, it was used as a template for integrating the Cross-lingual Emotion Detection task. The main difference is the number of classes used: the Stanford Sentiment Treebank (Socher et al., 2013) makes use of five classes ranging from very negative to very positive, whereas XED's fine-grained scheme spans eight classes. However, it should be noted that emotion detection represents *multi-dimensional* classification, which comes with its own host of challenges (Öhman et al., 2016). Although SST-5 makes use of several classes, they are segments of the same one-dimensional positive/negative scale. Given the scope and time-frame of the project, the SST template was still deemed suitable enough for rudimentary testing.

The emotion detection data and its pre-processing into SST-like format was provided by Emily Öhman and Kaisla Kajava (University of Helsinki). Similar to the XNLI scheme, for each language `$lg`, there is a subdirectory `data/downstream/XED-$lg` with language-specific train, dev, and test files. Each line is a simple text string: The first character of the line denotes the emotion class assigned to the sentence, which follows the emotion ID.

Because of the data deficiency noted in Kajava's work, an effort was made to use related resources to extend the data sets where possible. In particular, larger validation sets were needed for the MultiSent setup than was available. For this project, an additional English dev set of approximately 1,500 samples was produced. French and Italian dev sets, however, were obtained by extracting 800 random samples (about 14%) from their respective training

XED data	train	dev	test
English	5773	1505	652
French	4975	800	652
Italian	4974	800	652

Table 3: Number of lines per file in the XED data.

sets. The total number of samples are outlined in 3.

As mentioned in section 4.2, the Finnish language data also considered in Kajava's work can be added to the MultiSent framework in future experiments once enough suitable data is available.

MultiSent: M-BERT test accuracy on XNLI

	ar	bg	de	el	en	es	fr	hi	ru	sw	th	ur	vi	zh
MLP	33.3	33.3	47.7	33.3	55.9	50.2	50.9	33.3	33.4	35.8	38.2	33.3	33.3	33.3
LR	33.8	33.6	47.3	33.5	54.4	49.7	48.5	33.5	32.8	36.5	36.3	33.0	33.3	34.3

Table 4: Test accuracy with MLP and Logistic Regression.

6 Results and discussion

This chapter describes the results obtained with the MultiSent framework, insofar as it was developed within the ramifications of this thesis project. The sentence embeddings produced by the pre-trained Multilingual BERT model were evaluated on two types of cross-lingual tasks: Natural Language Inference and Emotion Detection. These embeddings are essentially contextualised word vectors extracted from hidden states of the model output, which are then used as input to the classifiers trained on top. As such, the main practical objective of the thesis was achieved with some success, although there is limited material for proper analysis of the system's performance and functionality. Despite a significant time investment, the technical execution of the project remained on an experimental level and the scores obtained are questionable at best. The technical intricacies going on beneath the hood of MultiSent – ensuring that the desired embeddings were obtained and that they represented a suitable input to the downstream classifiers – was the biggest hurdle of the experiment, and therefore rigorous analysis of the meager results is difficult to perform.

A full table of scores obtained with M-BERT on all the downstream task and language combinations is displayed in tables 4 and 5.

For comparison with other results in previous work, table 6 shows the obtained test accuracy across different models for selected languages. State-

MultiSent: M-BERT test accuracy on XED

	English	French	Italian
MLP	34.5	27.3	25.3
Logistic Regression	38.3	27.9	25.1

Table 5: Test accuracy on each of the XED languages.

of-the-art results with BERT fine-tuned for NLI are not available for all languages; Chinese is included here for the sake of comparison with results reported in related literature. The first row contains the "Translate-Train" results from the XNLI paper, in which the sentence encoder has been trained on MultiNLI training data using a bidirectional LSTM with max-pooling (Conneau et al., 2018b). (At this time, their best results were still obtained by running the encoder and classification in English and simply translating the target language into English at test time.) The second and third rows are some of the recent results reported on the Multilingual BERT GitHub webpage.[19] Out of these BERT models, the first was trained on the machine-translated MultiNLI data, whereas the second was fine-tuned only on English MultiNLI data and then evaluated on the XNLI data set (zero-shot classification). Lastly is reported the scores from the feature-based Multilingual BERT implementation of MultiSent.

There are a multitude of ways in which classification parameters could have been tweaked and only some of them were explored. Both the logistic regression and MLP (multi-layer perceptron) types of classification were tested, although the differences appeared minor across the experiment runs. The Chinese accuracy of 34.3 was obtained with logistic regression, whereas the highest English and French accuracy scores (by a margin of less than two per-

[19]https://github.com/google-research/bert/blob/master/multilingual.md#models

XNLI: TRANSLATE TRAIN approach

Model	English	French	Chinese
BiLSTM-max	73.7	68.3	67.0
M-BERT (translate-train)	81.9	-	76.6
M-BERT (zero-shot)	81.4	-	63.8
Multi-Sent: M-BERT	55.9	50.9	34.3

Table 6: Classification accuracy on the XNLI data-set.

centage points) were obtained through MLP with 5 hidden units. As every premise sentence is associated with three hypothesis sentences in the data, one with each label, the distribution of classes is even in the case of XNLI.

For Chinese, the comparably low accuracy – close to the random chance of 33.3% – suggests poor transfer of inference capability when task-specific fine-tuning is omitted. It is, however, also possible that something along the stream of encoding the data, extracting the features, and performing the classification is not working as intended. This proved to be a rather challenging process to debug, especially with practical limitations such as queuing for computing resources.

As a MultiSent evaluation baseline, a simple continuous bag-of-words (CBOW) model was intended to be constructed using multilingual fastText word embeddings for the relevant languages. Regrettably, this step was not successfully completed within the time-frame of this project. Likewise, a version of Multilingual BERT fine-tuned on the NLI task was planned to be evaluated in MultiSent. The Hugging Face library enables fine-tuning on XNLI, but this option was not explored further due to time constraints.

As for XED (table 7), the accuracy scores are fairly low, which is not surpris-

XED: Classification accuracy

Classifier type	English	French	Italian
MLP (Kajava)	49.3	37.7	41.5
MultiSent: MLP	34.5	27.3	25.3
MultiSent: Logistic Regression	38.3	27.9	25.1

Table 7: Comparison of classification accuracy on the XED data-set.

ing given the number of classes alone. In line with the existing fine-grained sentiment classification task, only the micro-F score is reported; that is, the overall accuracy of the classifier (portion of samples correctly classified out of all classified samples). The first row is from the paper by Kajava (2018). She reports accuracy scores for four different classifiers; only MLP is considered here. It is contrasted with the MultiSent classification results obtained with both MLP and logistic regression.

It bears mentioning that there are substantial differences in the MLP classifier parameters used: in Kajava (2018), the MLP contains three hidden layers, each with 100 neurons, whereas the MultiSent classifier only contains one layer. The results here were obtained with 5 hidden units in the layer. Other differences include activation functions used (Rectified Linear Unit versus Sigmoid, respectively) and the value of k in k-fold cross-validation (10 versus 5). Certainly for further experiments, the parameters could stand to be much more carefully tuned.

With the total number of eight classes, an even distribution would yield a 12.5% chance of randomly estimating the class correctly. In practice, as shown in section 4.2, the split between the emotion classes varied by a few percentage points, the most frequent class being *joy* with 15% of the samples. Additionally, as discussed in section 4.2, there is the matter of emo-

tion multi-dimensionality being less straightforward than fine-grained one-dimensional classification. The significant drop in performance for French and Italian compared to English can likely at least partially be attributed to the clearly diminished size of the training and development sets: the truncated French and Italian training sets are 14% smaller than the English training set, whereas the French and Italian development sets are as much as 47% smaller. The random extraction of the dev sets also minorly affected the class distributions, which is exacerbated by the small number of samples.

7 Conclusion

The result of this project is MultiSent, a version of SentEval with partially working modifications to enable evaluation of sentence embeddings in a cross-lingual setting. While executed on a smaller scale than originally intended, mostly due to technical challenges encountered during the practical implementation, some tentative results were obtained. The accuracy scores obtained for XNLI were far from competitive (in the 33.3–55.9 range), whereas XED scores mostly suffered from data scarcity (25.2–34.4). More rigorous analysis of the results would require more extensive tuning of the considered models, careful extraction of the embeddings, as well as more coverage across possible classification parameters. Another use case mentioned by Conneau et al. (2018b) is the possibility of mixing languages in tasks with dual-sentence input (such as pairing a premise sentence in one language with a hypothesis sentence in a different one), which was not explored within the scope of this project. With the basic modifications in place, XED can be easily extended with additional languages for further experiments. It would also be recommendable to implement a naive cross-lingual CBOW encoder based on multilingual word embeddings as an evaluation baseline.

References

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W13-3520`.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. Massively multilingual word embeddings. *CoRR*, abs/1602.01925, 2016. URL `http://arxiv.org/abs/1602.01925`.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1042. URL `https://www.aclweb.org/anthology/P17-1042`.

Yoshua Bengio, Rejean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015.

Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://www.aclweb.org/anthology/D15-1075.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing *almost* from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1269.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL https://www.aclweb.org/anthology/D17-1070.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://www.aclweb.org/anthology/P18-1198.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://www.aclweb.org/anthology/D18-1269.

Mathias Creutz. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1218.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association

for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. *LERC*, 2018. URL http://arxiv.org/abs/1802.06893.

Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1162. URL https://www.aclweb.org/anthology/N16-1162.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080, 2020.

Kaisla Kajava. Cross-lingual sentiment preservation and transfer learning in binary and multi-class sentiment classification. Master's thesis, University of Helsinki, 2018.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information*

Processing Systems - Volume 2, NIPS'15, page 3294–3302, Cambridge, MA, USA, 2015. MIT Press.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1090.

Neha Nayak, Gabor Angeli, and Christopher D. Manning. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 19–23, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2504. URL https://www.aclweb.org/anthology/W16-2504.

Donald A Norman. *Memory, knowledge, and the answering of questions*. California Univ., La Jolla. Center for Human Information Processing., 1972. 57p.: Paper presented at the Loyola Symposium cnCognitive Psychology (Chicago, Illinois, 1972).

Emily Öhman, Timo Honkela, and Jörg Tiedemann. The challenges of multi-dimensional sentiment analysis across languages. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 138–142, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/W16-4315.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove:

Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4302. URL https://www.aclweb.org/anthology/W19-4302.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://www.aclweb.org/anthology/P19-1493.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language process-

ing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-5004. URL https://www.aclweb.org/anthology/N19-5004.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1036. URL https://www.aclweb.org/anthology/D15-1036.

Laura Smith, Salvatore Giorgi, Rishi Solanki, Johannes Eichstaedt, H. Andrew Schwartz, Muhammad Abdul-Mageed, Anneke Buffone, and Lyle Ungar. Does 'well-being' translate on twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2042–2047, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1217. URL https://www.aclweb.org/anthology/D16-1217.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Jög Tiedemann. Parallel data, tools and interfaces in opus. In Nico-

letta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

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