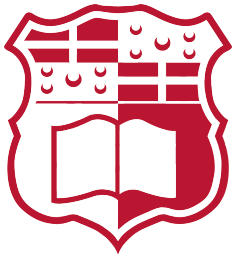

Towards Polyglot Machines: Cross-lingual Natural Language Inference

Jake J. Dalli

Supervised by Prof. Albert Gatt

Co-supervised by Dr Claudia Borg

Department of Artificial Intelligence

Faculty of ICT

University of Malta

November, 2020

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. Artificial Intelligence.

WWW.UM.EDU.MT

First edition, April 4, 2021

Acknowledgements

I would like to express my gratitude to my supervisors, Prof. Albert Gatt and Dr Claudia Borg, whose perspective and enthusiastic guidance were vital throughout this year of study. I would also like to thank my parents, John and Judith, and my brother Jeffrey, for their unconditional support throughout my studies.

Abstract

Inference is a central aspect of Natural Language Processing (NLP); the Natural Language Inference task (NLI), also called Recognizing Textual Entailment (RTE), is the task of determining whether a *hypothesis* text fragment corroborates (positvely entails), contradicts (negatively entails) or bears no relation to (no entailment) a *premise* text fragment. Prior work on this task has nearly exclusively focused on the monolingual English inference; in this study, we aim to address cross-lingual NLI. We study the area of cross-lingual natural language inference by addressing two different formulations of the task; cross-lingual transfer, where we explore how an inference model trained for English can be fine-tuned to perform inference in another language; and purely cross-lingual inference, where we train a model to detect inference for sentence pairs in different languages. Within our study, we experiment with two neural network architectures to address these tasks, a bidirectional LSTM and a decomposable attention model, employing aligned word embeddings to represent language. Results show that the bidirectional LSTM neural network performs best across all tasks. Moreover, we also show that employing machine translation to deal with cross-lingual NLI provides the best results. Although the use of word embeddings to encode sentences does not perform as well sentence embeddings, our proposed architecture using word embeddings requires significantly less computational resources due to the lower dimensionality of the embeddings. Our approach presents a results with less than a 10% loss of accuracy, and as little as a 5% loss in the best case, while using a fraction of the computational resources required by solutions employing sentence embeddings.

Contents

List of Figures

List of Tables

List of Abbreviations

NLP Natural Language Processing
NLI Natural Language Inference
RTE Recognizing Textual Entailment
RNN Recurrent Neural Network
CNN Convolutional Neural Network
LSTM Long Short-Term Memory Cell
GRU Gated Recurrent Unit
LDA Latent Dirichlet Allocation
LSA Latent Semantic Analysis
CBOW Continuous Bag of Words
CSLS Cross-Domain Similarity Local Scaling
SNLI Stanford Natural Language Inference corpus
MultiNLI Multi-Genre Natural Language Inference corpus
XNLI Multi-Lingual Natural Language Inference Corpus
TRAN-XNLI Translated XNLI Corpus
CROSS-XNLI Cross-Lingual XNLI Corpus
AWS Amazon Web Services
EC2 Amazon Elastic Compute Cloud

1

Introduction

Language is diverse. The five most spoken languages are used by around 25% of the world's population, whilst 75% share approximately an additional 6,500 languages. Furthermore, over half of the world's population is bilingual (Simons and Fennig, 2017). This diversity is not reflected in the field of natural language processing; a digital divide exists between languages (Bender, 2019). Only a few *high-resource* languages possess collections of digitized text and speech which are annotated sufficiently to be used for language processing tasks. The list includes English, Mandarin, Arabic and French; to a lesser extent, German, Portuguese, Spanish and Finnish also qualify. The remaining *low-resource* languages have far fewer resources, with several hundreds of languages having close to no available resources (Bender, 2019).

As a result, research in the field of natural language processing (NLP) is predominantly monolingual, and carried out in English by default. In fact, a large proportion of English-language work neglects to specify the language for which its approach caters, with work on other languages being considered niche, and, therefore, secondary to their monolingual English counterparts (Bender, 2019, 2011).

This status quo is undesirable for several reasons. Firstly, with NLP progression only available for a few languages, the majority of the world's population risks falling behind in terms of technological progress (Rehm, 2013) as work within the field of multilingual NLP is key to the fostering of technological inclusion. More importantly, the lack of consideration for how approaches to NLP may be applied to different languages is not amenable to a true natural language understanding system. Several approaches perform poorly when ported to different languages, calling their validity into question with respect to how they contribute to achieving computational natural language understanding (Ruder, 2020). Finally, it is expensive to develop a new monolingual NLP for each language or task (Ruder et al., 2019).

While potentially overcoming these disadvantages, computational NLP on a large, multilingual scale requires a different approach, one which is better equipped to reflect the ultimate goal of a maintainable and inclusive natural language understanding. For these reasons, our goal is to explore the area of cross-lingual natural language processing, particularly with respect to building a 'generalized' natural language understanding system.

In this chapter, we introduce the task of natural language inference as the cornerstone for language understanding in order to explore the limits of NLP within a cross-lingual context. We propose a set of aims to assess language understanding across languages, as well applying current approaches to address the task.

1.1 | Motivation

Searle (1990) illustrates computational linguistic understanding by introducing the "Chinese Room" experiment. Searle imagines himself in a room, where he is prompted by Chinese characters which are passed under his door. Searle is able to respond to the prompts in an intelligible way, by following a set of instructions and a database of Chinese symbols, even though he doesn't know any Chinese and is not understanding either the prompt or his own reply. The system described by Searle allows questions to be answered without any comprehension of the Chinese words involved, disputing that form alone is not sufficient to achieve true understanding. In the experiment, Searle satisfies the Turing Test, but the manner in which he satisfies it raises the question of whether the mastery of form alone is sufficient to constitute understanding. The task of inferring non-symbolic representation, semantic meaning, from symbolic form is described as the symbol grounding problem (Harnad, 1990).

A key property of symbol grounding is the ability to identify referents between words, in order to infer meaning (Fodor, 1975) - given a set of linguistic representations, one observes the semantic interpretation of the representations by inferring meaning through their relationship. Computational formulations of the inference task broadly attempt to determine whether two text representations are likely to be inferred from each other. Cooper et al. (1996) describe natural language inference as "not only a central manifestation of semantic competence but is in fact centrally constitutive of it."

The inference task is initially formalized in the Recognizing Textual Entailment Challenge (RTE) (Bentivogli et al., 2011; Dagan et al., 2005; Giampiccolo et al., 2007), where a model is tasked with determining whether two text fragments, a *premise* and a *hypothesis*, entail each other. In more recent literature, the same task has been referred to as

Natural Language Inference (NLI) (Conneau et al., 2018; Hu et al., 2020; Wang et al., 2018).

The task has predominantly taken the form of a three-way classification problem, with models classifying whether a *premise* (P) positively entails (substantiates), negatively entails (contradicts) or has no relation to a *hypothesis* (H):

$$RTE \in \{(P \vdash H), (P \vdash !H), (P \nvdash H)\} \tag{1.1}$$

For example, the premise *"The boy jumped over the wall in the garden."* would *positively entail* a hypothesis sentence *"There is a wall in the garden."* Conversely, the same premise would *negatively entail "The boy is inside the living room"* and bear *no relation* to (neutral) the hypothesis *"The boy is wearing a green shirt."*

In section 2.1 we provide a more nuanced selection of examples, explaining how the inference task relates to determinacy of entities, events and time.

The SNLI (Stanford Natural Language Inference) (Bowman et al., 2015) and MultiNLI (Multi-Genre Natural Language Inference) (Williams et al., 2018) benchmarks provide corpora for modeling inference. However, due to its inherently nuanced nature, developing corpora which allow the study of inference in all its complexity remains an open objective (Bender and Koller, 2020; Glockner et al., 2018). This fact alone is testament to the complexity of the NLI task, and its relationship to language understanding.

Recent work within the area of machine learning have seen several advances related to architectures used to approach natural language processing, particularly in the area of deep learning (Liu et al., 2017). The use of deep neural networks, particularly Recurrent Neural Networks, has replaced previous state of the art models built using hand-crafted features (Gers et al., 1999; Hochreiter and Schmidhuber, 1997; Rocktäschel et al., 2015). Other, more recent approaches employ attention mechanisms to capture contextual relationships that exist between different words or phrases (Bahdanau et al., 2015). The most recent advances suggest approaches which exclusively employ attention, replacing traditional neural network architectures altogether (Bender and Koller, 2020; Vaswani et al., 2017).

These developments have triggered advances in the way word representations are learnt from corpora, adopting predictive approaches as opposed to frequency-based representations (Baroni et al., 2014; Mikolov et al., 2013a). Predictive language models and word embeddings provide another area for research in cross-lingual NLP through alignment of word embedding representations to provide translation for different languages (Ruder et al., 2019).

Within the cross-lingual context, the inference task can be approached with different aims in mind. For instance, one can seek to improve NLI performance within *low-resource* languages by leveraging models learned on *high-resource* languages. Inspired by recent advancements in the field of transfer learning (Aytar and Zisserman, 2011; Pan and Yang, 2010; Thrun and Pratt, 1998; Torrey and Shavlik, 2009), researchers have proposed a number of benchmarks for transferring inference learned on English to other languages (Conneau et al., 2018). This line of work follows other transfer learning approaches dealing with reusing models for different tasks (Hu et al., 2020; Wang et al., 2018). Another approach is to re-frame the inference task within a cross-lingual context, where the *premise* and the *hypothesis* are in different languages (Mehdad et al., 2010). For example, the inference relation is classified over the text fragments *"The boy jumped over the wall in the garden"*, in English, and *"Hay una pared en el jardín"* in Spanish. The goal of such tasks is to build truly cross-lingual inference capabilities.

Beyond achieving the ultimate goal of massively multilingual natural language understanding, there exist more practical applications of addressing the inference task. Such applications include question answering (Harabagiu and Hickl, 2006), information retrieval (Clinchant et al., 2006), information extraction (Romano et al., 2006), document summarization (Lloret et al., 2008) and content synchronization (Negri et al., 2013).

Within our work, we contribute research towards achieving massively multilingual natural language understanding in a practical manner, in order to ensure technological inclusion within NLP advances. As the field of NLP progresses, we must ensure that applications improve across all languages.

1.2 | Aims and Objectives

As discussed above, the area of cross-lingual natural language processing is under-explored, and in its early stages. The overarching aim of our study is to apply current state-of-the-art approaches within a cross-lingual context, in order to examine different possibilities within the field. Our secondary goal is to address the larger challenge of cross-lingual understanding. We do this by selecting the language inference task as the target task for our study. Thus, the main aim of this work is to explore two key ideas:

- Investigate the use of transfer learning to improve the inference task across different languages. This can be done by taking a model trained for NLI within a *high-resource* language (English), we aim to improve inference within other languages.

- Investigate the application of deep learning approaches for the cross-lingual NLI formulation where the premise and the hypothesis are in different languages.

Our aims will be accomplished by developing a number of different components with the following objectives:

1. Build a number of word embedding models which cater for our aims, particularly a set of aligned word embeddings for cross-lingual natural language processing;
2. Design and adapt neural architectures for transfer learning within the inference task. We explore two options, the first consisting of a *Cross-lingual transfer* scenario, and the second considering a situation where translation is employed as an intermediate step, porting the inference task to a translation task.
3. Design and adapt neural architectures for approaching the cross-lingual inference formulation, where the premise is in one language (English) and the hypothesis is in another language.
4. Evaluate and investigate the performance of our methods with established benchmarks.
5. Suggest further work in the area of cross-lingual inference, and more broadly, cross-lingual natural language processing.

Within our research, we use the term MONOLINGUAL to refer to English language approaches.

1.3 | Approach

The aim of our research is twofold. Our first aim is to explore whether inference models learned on a high-resource language can be transferred to other languages; we refer to this as the *cross-lingual transfer learning* task. In this task, the premise and the hypothesis are both within the same language.

Secondly, we aim to address the alternative cross-lingual formulation of the inference task, where the *premise* and *hypothesis* are in different languages. We refer to this task as *purely cross-lingual* inference. Moreover, for the purposes of our research we only consider the case where the *premise* is in English and the *hypothesis* is in another language.

The XNLI data set (Conneau et al., 2018) provides a benchmark for evaluating multilingual inference models. The primary challenge posed by XNLI is the application of

fine-tuned inference models, initially trained for English NLI, to perform inference in different languages. Baselines presented within the paper are trained on the monolingual English MultiNLI corpus (Williams et al., 2018). We provide a brief history of these corpora and their predecessors in Section 2.1.2.

We adopt the XNLI challenge as the primary resource for our research since it provides adequate representation of languages. The corpus is composed of 7,500 premise and hypothesis pairs, collected through crowd sourcing, which are translated into fourteen different languages. Available languages include French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu. With the exception of English, French, German and Chinese, all of these languages can be considered relatively *low-resource*, with Swahili and Urdu representing the most poorly supported languages (Bender, 2019).

Our approach adopts a sentence encoder architecture. We use word embeddings to create vector representations which are uses to initialize the neural networks. To construct these embeddings, we train separate word embedding models for different languages. At a later stage, these embeddings are aligned into a single vector space, creating cross-lingual word representations.

We select the fastText framework (Joulin et al., 2016a) to generate our word embeddings, selecting word level alignment as our choice of alignment. In Section 2.3 of the next chapter, we describe the possible approaches for aligning embeddings.

The aligned embeddings are used to initialize different neural architectures which we design to address our goals. For the Cross-lingual transfer learning task, we design two target architectures in order to compare their performance. Our first architecture is a stacked bidirectional LSTM model (Bowman et al., 2015; Liu et al., 2016a); our second approach follows a decomposeable attention model (Parikh et al., 2016). We outline the techniques behind these architectures in Section 2.2.

Our models are tested in three of scenarios to address our objectives:

- **Cross-lingual transfer:** We train the model on the MultiNLI (Williams et al., 2018) corpus, and fine-tune the model for the XNLI task.

- **Translation Test Inference:** We investigate a scenario where cross-lingual inference is ported to a translation task. Pairs from the XNLI data set are translated to English, thus converting the cross-lingual task to a monolingual task and offloading the complexity of dealing with different languages to a machine translation task. Next, we train our models on the MultiNLI (Williams et al., 2018) corpus and evaluate them on the translated XNLI corpus. The translation software used for this experiment is Google Translate (Google, 2006).

- **Purely Cross-Lingual Inference:** We construct a variation of the XNLI corpus where English premises are paired with hypotheses in a different language. Next, we train the model on the MultiNLI corpus and fine-tune the model using the constructed data set.

Within our evaluation and discussion, we aim to assess the effectiveness of our models employed within each separate scenarios. Moreover, we compare and contrast the different models in terms of their utilization of attention mechanisms.

1.4 | Chapter Overview

The contents of this document is organized as follows:

Introduction Our Introduction provides an overview of the current state of natural language processing in terms of multilingual research, exploring the motivations for conducting such research. We explain why we chose to address the task of natural language inference, and why it is desirable to do so within a multilingual context. Finally, we provide a brief overview of our aims and objectives and the approach undertaken to achieve the said aims.

Background and Literature Review In this chapter, we explore the different ideas which relate to our research area. We first outline the inference task with examples, paying particular attention to the complexities of the task; we then introduce the task within a cross-lingual context, making a clear distinction between transfer-learning approaches and alternate formulations for cross-lingual tasks. In subsequent sections, we delve deeper into neural network approaches within NLP, with a particular emphasis on inference architectures and word embeddings.

Methodology Within the methodology chapter, we briefly summarize the findings within our background research and explain the approaches selected in terms of the research. We explain our target architecture, the design of our models, our development process and the methodology for designing experiments to reach our aims.

Evaluation and Results In this chapter we outline a series of experiments conducted to investigate the aims described in the Introduction Chapter. In particular, we assess the quality of the word embeddings built and the performance of our implemented neural architectures.

Discussion In this chapter we discuss the results achieved within our evaluation in the context of the current baselines and our stated aims. Furthermore, we seek to provide an overview of our contribution to the area of cross-lingual NLP.

Conclusion In our conclusion we reiterate the motivations for our work, summarizing our approach and the results achieved. Finally we propose future work which can be undertaken to further improve the current state of multilingual natural language processing.

2

Background and Literature Review

The task of natural language inference, also known as 'recognizing textual entailment' (RTE), is widely studied within the field of monolingual natural language processing. Various approaches have been employed in tackling the problem within a monolingual setting, including symbolic logic, knowledge bases, neural networks and distributed representations. In this section, we introduce the task of language inference within a monolingual context, explaining the core ideas behind neural networks. Subsequently, we explore recent advances within the areas of distributed representations and neural learning, to explore how these could be applied to language inference within a cross-lingual context. Finally, we explore different methods for evaluating such systems.

2.1 | Natural Language Inference

Natural language inference (NLI) is the task of establishing whether a pair of text fragments possess an inferential, contradictory or neutral relationship as judged by human reasoning. Cooper et al. (1996) initially identify the inference task to be "the best way of testing NLP system's semantic capacity", identifying the logical concept of inference in relation to linguistic phenomena.

The task has since been redefined to adopt a more probabilistic approach as opposed to previous formulations, which rely on theoretic semantics. Dagan et al. (2005) define inference as a directional relationship between two text fragments. Given a pair of text fragments, a premise P and a hypothesis H, P is said to entail H ($P \Rightarrow H$) if H can be inferred from P. This relation is referred to as *positive entailment*. Conversely, if P contradicts H ($p \nRightarrow h$), the relation is categorized as *negative entailment* (or, more commonly, *contradiction*). Text fragments which possess neither an inferential nor a contradictory relationship are said to be *not entailed* or *neutral*.

Premise	On 15 April 2019, France came close to losing its most famous cathedral when a fire broke out beneath the roof of Notre-Dame de Paris cathedral in Paris.
Hypothesis	The Notre-Dame cathedral is located in France.
Label	*Positive Entailment/Inference*
Premise	At 18:52 smoke was visible outside the cathedral, flames appeared in the next ten minutes, as firefighters arrived.
Hypothesis	Firefighters arrived on site within ten minutes after the fire alarm was raised.
Label	*Neutral/No Relation*
Premise	Paris prosecutors said that no sources have come forward with information and that there is no evidence of the cathedral fire being caused by deliberate act.
Hypothesis	A volunteer church assistant has confessed to starting a fire that severely damaged the cathedral.
Label	*Negative Entailment/Contradiction*

Table 2.1: Examples of entailing, neutral and non-entailing text fragments.

Table 2.1 shows a number of examples of entailing, neutral and contradictory text fragments related to a fire which broke out at the Cathedral of Notre-Dame. The first is positively entailing, given that a human would infer the location of the cathedral from the leading statement that the country came within reach of losing an iconic cathedral. The second statement is neutral; while it may be implied that the visual indicators of smoke and flames triggered the alarm, the statements contain no specific information about the relationship between the events. Lastly, the third statement is negatively entailed, given a direct contradiction between the facts that no sources spoke of the fire and that a confession was received by a church volunteer.

As shown in the above examples, the task of language inference is inherently nuanced, depending on several inferential factors including temporal relationships, prior knowledge, and semantic meaning:

- Consider the first example, omitting leading statement "*France came close to losing...*" would require a human to depend on prior knowledge that Paris is located in France, especially considering that there are several cities called Paris (Paris is also a city in Texas, introducing ambiguity), in order to conclude positive entailment.

- Within the second example, the visual indicators (smoke and flames) are not necessarily indicative of the fire alarm (a device) being raised. The order of events and semantic relationship between the indicators have no relation. Such tight coupling of temporal and semantic factors may even cause disagreement among humans.

- A human's conclusion that the final statement is contradictory is dependent on the knowledge that the semantics of the word *confession* implies that an individual has *come forward* with relevant information.

- Minor changes to the samples may change the resultant label entirely. In the second statement, omitting the word *"fire"* from "fire alarm" would change the context. In that case, it would be implied that the visual indicators of smoke and flames constitute an alarm, thus resulting in positive entailment.

A key element of language inference is identifying the determinacy of different entities and events in reference to each other, strictly restricted to information within the texts. One exception to this restriction is the incorporation of multi-modal representations, such as images, within the task (Bergsma and Van Durme, 2011; Calixto et al., 2017; Gella et al., 2017). Moreover, our examples only address a single context: that of a news article. Other contexts may present different challenges due to variations in the choice of language, presupposition of prior knowledge and context from which the text is sourced. The data used within our research, MutliNLI and XNLI, source data from different contexts (Conneau et al., 2018; Williams et al., 2018).

2.1.1 | Cross-Lingual Natural Language Inference

The study of natural language inference within the cross-lingual context is in its early stages, with most research revolving around monolingual English-language inference. Nonetheless, two lines of work exist, providing task formulations:

1. **Cross-lingual transfer:** Catering specifically to improving monolingual inference for different languages, leveraging models learnt in a source language (typically English) to improve accuracy in other high-resource (for example French) or low-resource (for example Swahili) languages (Conneau et al., 2018). A related line of work focuses on cross-lingual learning across different task, expanding the transfer learning idea across domains (languages) and other tasks (paraphrasing, part-of-speech tagging, question answering) alongside NLI (Hu et al., 2020; Wang et al., 2018).

2. **Alternative Formulations:** A second line of work exists in the form of *purely cross-lingual inference*, where the inference task is adapted to work premise and hypothesis fragments in different languages (Mehdad et al., 2010).

In this section, we briefly introduce the different approaches to dealing with cross-lingual inference, outlining the methods and benchmarks that they provide.

2.1.1.1 | Cross-lingual transfer Learning

Inspired by recent advances within the field of computer vision, one approach is to frame cross-lingual NLP as a cross-domain transfer learning problem. In such approaches, models are trained to transfer knowledge for a given task across different domains (Aytar and Zisserman, 2011; Pan and Yang, 2010; Thrun and Pratt, 1998; Torrey and Shavlik, 2009).

Within the context of language processing, domains are represented by different languages. A model is trained on a source language L_1 to perform a task in target language L_2, this task remaining unseen during the course of the training. A large body of research surrounds the performance of transfer learning across languages for a variety of tasks, particularly machine translation (Aharoni et al., 2019; Artetxe and Schwenk, 2019b; Conneau and Lample, 2019; Eriguchi et al., 2018; Schuster et al., 2019). In such cases, research aims to leverage knowledge learned from a high-resource source language to increase performance in low-resource target languages.

Such approaches typically follow three phases: a pre-training phase, an adaptation phase and a transfer phase. Within the pre-training phase, a cross-lingual representation is learned, including vocabulary from both languages, with the aim of building a generalized cross-lingual language model. The pre-trained model is then adapted to a specific task within the source language; such tasks include machine translation, or, in our case, natural language inference. During the adaptation phase, the model learns task-specific parameters on top of the cross-lingual representation. During a final transfer phase, the model is fine-tuned using data from the target language, known as *Cross-lingual transfer*, to apply inference in the target language. Cross-lingual transfer learning methods are frequently referred to as *Cross-lingual transfer* methods.

A cross-lingual transfer learning approach focuses on employing existing NLP architectures across different languages to improve accuracy. Within the context of NLI, such approaches deal with applying cross-lingual transfer to improve monolingual inference in different languages.

2.1.1.2 | Alternative Formulations

Mehdad et al. (2010) suggest re-framing the NLI task to address purely cross-lingual inference, where text fragments are in different languages. The task is adapted such that the task is to detect inference between a premise P in one language L_1 (for example English), and the hypothesis H in another language L_2 (for example German). As discussed in Section 1.1, such cross-lingual inference is classified over the text fragments where for example *"The boy jumped over the wall in the garden"*, in English, and *"Hay una*

pared en el jardín" in Spanish. Such cross-lingual inference can be achieved using two general approaches.

In the first approach, we port the cross-lingual challenge to a machine translation (MT) task, by translating H to L_1 and carrying out NLI within a monolingual context. This approach allows for a modular system which is tightly coupled to machine translation. Given strong NLI performance within L_1, one can leverage advances in *MT* in order to improve *NLI* performance in L_2. The main disadvantage of this approach is that *MT*, particularly within low resource languages, may not be sufficient to deliver the necessary improvements within the *NLI* task.

A second approach suggests the use of translation as a pre-processing step, prior to the inference task itself. Rather than translating either the premise or hypothesis fragments, we learn about statistical relations between text fragments within and between the language fragments. This approach is possible by extracting relations between phrases in L_1 and L_2 to fuel inference mechanisms. For example, such approaches would detect entailment relations between *landmark*, in English, and the phrases *das Wahrzeichen*, *der Orientierungspunkt* and *der Grenzpfahl* in German (all of which loosely translate to landmark); all three German phrases in this example express entailment to the English phrase to varying degrees. This approach provides for building more expressive entailment relations. However, they further tightly couple the NLI task to the MT task, resulting in a less modular (and therefore less portable) cross-lingual NLI system.

2.1.2 | Benchmarks and Corpora

The complexity of the inference task is well-researched in linguistics. In particular, the notions of presupposition (Grzymala-Busse, 1999) and indeterminacy (Halpern, 1990; Kenney and Smith, 1996) are addressed. Several researchers have presented the task in different forms in order to address the issues within a computational context; initially within the monolingual English domain, and eventually extended to the cross-lingual transfer context. We also review multi-task benchmarks, which provide alternative baselines for our research.

2.1.2.1 | Monolingual Natural language Inference

The *PASCAL Recognizing Textual Entailment Challenge* (RTE) presented the inference task for the first time in 2005, investigating different scenarios in subsequent annual challenges (Bar-Heim et al., 2006; Dagan et al., 2005). The task includes text fragments sourced from English Wikipedia and news articles and is initially framed as a binary

inference challenge, presenting a data set where text pairs have two potential target labels, *entailment* and *no entailment*. In subsequent challenges, the problem is re-framed as a three-way challenge, including a 'neutral' class, as presented in Section 2.1 (Giampiccolo et al., 2008).

Other challenges have presented longer sentences in order to mimic more realistic scenarios (Giampiccolo et al., 2007), scoping the problem as a search task or a novelty detection task (Bentivogli et al., 2009, 2010, 2011). More widely, and outside of the *PASCAL RTE Challenge*, there have also been attempts to address text fragments' appeal to world knowledge or to tie the inference task to question-answering (QA) applications (Khot et al., 2018; Levesque et al., 2012). Although these attempts have provided interesting alternative benchmarks, the three-way multi-class inference task described in the previous section remains the predominant target inference task within the field.

Following the initial framing of the problem as RTE, several other works have proposed alternate methods of collecting corpora, in order to increase the size of the corpus and to address different linguistic nuances. Marelli et al. (2014b) propose the *SICK* corpus for inference, composed of image and video captions which are modified to contain particular linguistic properties such as alternate syntax, negation and quantifiers. Both data sets provide a larger benchmark in comparison to the initial RTE task, but such automated corpus construction has produce data of questionable quality (Marelli et al., 2014a).

Bowman et al. (2015) present the *Stanford Natural Language Inference* (SNLI) challenge data set, generated by prompting humans to produce *entailment*, *neutral* or *contradictory* hypothesis sentences. Data is generated by prompting human annotators using images, resulting in a data set of text pairs addressing the task. The SNLI data set provided a larger and richer benchmark to its predecessors, triggering wide interest in the inference task. Yet the data set shared a common disadvantage due to reliance on image captions; such descriptions do not effectively represent linguistic phenomena such as tense.

The *RepEval 2017 Shared Task* poses the inference challenge inspired by the SNLI corpus, proposing an alternative benchmark to address its shortcomings. The task presents the Multi-Genre Natural Language Inference (MultiNLI) corpus which captures data covering a wider range of genres, representing both written and spoken styles of language (Nangia et al., 2017; Williams et al., 2018).

Several monolingual inference solutions to the SNLI and MultiNLI challenges have been proposed, particularly ones that leverage *deep learning* approaches which use convolutional neural networks (CNN) and recurrent neural networks (RNN). We provide an in-depth overview of these approaches in Section 2.2.2.1.

Chen et al. (2017b) present a CNN model with bidirectional LSTM cells, reporting

the highest accuracy for the MultiNLI task at 74.9%. The input layer is composed of two concatenated embeddings: a word-level vector representation of the sentence as represented by a pre-trained GloVe model (Pennington et al., 2014) (described later in Section 2.3.1), and a character-level vector representation of the sentence. Inputs are fed to three bidirectional LSTM layers, in a sentence encoding layer. Another hidden layer employs a specific gated-attention LSTM (Xue et al., 2020) with maximum pooling before feeding the final result multi-layer perception layer. The authors report that the application of the gated-attention LSTM increase accuracy from 73.9% to 74.9%.

In contrast, Nie and Bansal (2017) propose a three-layer bidirectional LSTM RNN, achieving 74.5% accuracy without incorporating cross-attention. The input sentences (premise and hypothesis) are separately encoded into fixed length vectors which are fed as inputs; a third input composed of both the premise and the hypothesis is input to the network. The outputs are fed to a multi-layer-perceptron classifier to calculate the output.

2.1.2.2 | Cross-Lingual Natural Language Inference

Conneau et al. (2018) present XNLI, a benchmark data set for the cross-lingual context, inspired by previous monolingual data sets (Bowman et al., 2015; Williams et al., 2018). The task treats the inference task for different languages when only English data is available at training time, also known as cross-lingual transfer from English. XNLI is composed of 7500 premise and hypothesis pairs from the English MNLI (Williams et al., 2018) corpus translated into fourteen languages, spanning across high resource and low resource languages. The languages include French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu.

The authors propose two distinct approaches to cross-lingual NLI, translation-based approaches and sentence encoding approaches, suggesting possible solutions using each approach and providing a benchmark. The approaches make use of cross-lingual word embeddings and bidirectional LSTMs (BiLSTM) which we review in Sections 2.3.2 and 2.2.2 respectively.

- **Translation-based Approaches:** In such systems, one ports the cross-lingual NLI to a translation task; either by first translating the source language data to the target data and then training the inference classifier on the target language (*TRANSLATE TRAIN*), or by training the classifier in the source language, and translating at target time (*TRANSLATE TEST*). In both cases, the solution relies on the quality of the machine translation for the target language. The authors benchmark

two BiLSTM models for the *TRANSLATE TRAIN* and *TRANSLATE TEST*, using a proprietary translation system.

- **Multilingual Sentence Encoders:** Such systems exclusively rely on word embeddings to build a multilingual classifier (cross-lingual word embeddings). In turn, the neural architecture performs a cross-lingual encoding process. The authors provide two baseline architectures: a continuous bag of words architecture (X-CBOW); and a BiLSTM architecture (Hochreiter and Schmidhuber, 1997) (*X-BILSTM*) trained on English MultiNLI data (Williams et al., 2018). Word embeddings for these approaches are trained using the MUSE framework (Lample et al., 2018), and aligned using a sentence-level mapping-based approach. This is further explained in Section 2.3.2.

Results show that the translation-based approaches outperform all other models in all languages, with accuracy ranging from 59.3% for Urdu to 73.7% for English. The X-CBOW model preforms worst of all the models, corroborating previous claims that the inference task requires sequence information as opposed to only word information (Bowman et al., 2015; Conneau et al., 2017a).

To date, the XNLI corpus is the only multilingual NLI benchmark available. The task also forms part of the XTREME multilingual multi-task challenge (Hu et al., 2020), which includes the XNLI baseline alongside baselines for paraphrasing, part-of-speech (POS) tagging, named entity recognition (NER), question answering (QA) and sentence extraction, all within a cross-lingual context.

A number of baseline models are provided for the XTREME task, particularly transformer based models (explained in section 2.2.2.2) with different training objectives:

- Cross-lingual transfer: Models are trained using English data, and tested using the target language data.
- Translate-train: Models are trained on English training data which is translated to the target language, and tested in the target language.
- Translate-test: Models are trained on the English data, and evaluated on the target language data translated to English.
- In-Language models: Models are trained on the target language data and tested on the target language data.

For the cross-lingual transfer case, mBERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) transformer models achieve 65.4% and 79.2% accuracy respectively.

An additional translation encoder achieves 67.4% accuracy. Meanwhile, mBERT obtains accuracy between 74% and 76.8% with the translate-train and translate-test objectives. Additionally, the benchmark provides metrics for human evaluation of each task; the human accuracy for cross-lingual inference stands at 92.8%.

2.2 | Approaches to Cross-Lingual NLI

In the previous section, we introduced the natural language inference task, also referred to as the 'recognizing textual entailment' task, as a multi-class classification problem, which classifies whether two text fragments positively entail, negatively entail or do not entail each other. We also briefly introduced the current state-of-the-art benchmarks and corpora for the task within an English monolingual and cross-lingual context. Natural Language Inference solutions are typically composed of two components: the first, a language modeling component, which aims to model inferential relationships in text; the second, an inference model, which aims to classify the inferential relation described in the task.

In this section we explore different approaches to inference. First, we briefly introduce traditional inference models which employ feature engineering to model the language and classify the relationship. Second, we explore more recent approaches employing neural networks.

2.2.1 | Feature Engineering in Inference Models

Negri et al. (2013) introduce the task of cross-lingual language inference within the context of content synchronisation. The task further extends our previous definition in section 2.1 by specifying the relation of entailment (bidirectional, forward, backward) in synchronisation, addressing one of the main applications for cross-lingual NLI. The task presents a data set of 1,500 textual entailment pairs for a combination of languages coupled with English, these include Spanish, Italian, French and German.

Solutions to the task can be broadly categorized across two dimensions; the reliance on initial translation, as initially proposed by Mehdad et al. (2010); and the target label for the inference task - partly relying on a binary YES or NO meta-classification (as in Levesque et al. (2012)), or a multi-class classification as described by the task definition.

One approach, **BUAP** (Vilarino et al., 2013), uses English as a *pivot language*, by translating the premise and hypothesis into each language in the pair. The authors propose a number of different approaches for extracting features from the pairs of same-language fragments. Sentences are represented as N-grams, spanning across words and part-

of-speech tags, and similarity measures, including Eucledian distance, Manhattan distance and Jaccard coefficient. Additionally, other features include a system of various binary classification models representing different judgements. All features are combined within a voting system that uses a majority criterion to detect entailment. The model's best performance achieves a 39% accuracy, with the application of similarity measures proving to be the most effective.

Softcard (Jiménez et al., 2013) also uses English as a pivoting language, computing the Edit Distance and Jaro-Winkler similarity measures. However, it uses *character-level* N-grams (Q-Grams). In addition, it applies Support Vector Machine (SVM) to predict the multi-class label, achieving an overall average accuracy that varies between 42.6% and 45.8%.

ALTN, an alternate approach proposed by Turchi and Negri (2013), trains an alignment model on sets of parallel texts in different languages. Features are subsequently extracted from the alignment process in order to train an SVM for multi-class categorisation. All features are language agnostic, including the proportion, length and counts of aligned word sequences. The approach yields a relatively consistent performance across language pairs ranging from 38.8% to 45.2% accuracy.

2.2.2 | Deep Learning Approaches

Neural Network models have been proposed as an alternative to traditional machine learning approaches with success for several language processing tasks. As discussed in the previous section, traditional NLI approaches require the manual construction of features in order to cater for different linguistic attributes; these include linguistic and syntactic features. However, natural language is notoriously complex, as it consists of a high-dimensional space. This results in highly-complex models, which are specific to the task and thus difficult to port to different domains. Meanwhile, neural networks allow for a generalized, self-correcting method of modeling non-linear and complex relationships. The application of neural networks within NLP has produced several advances across different tasks. Such methods are employed separately at different levels of NLI solutions. The concepts behind such neural architectures feature strongly in our research, since we consider applications at two different phases for cross-lingual natural language inference:

1. **Language Representations:** In the first phase, we attempt to learn a language model from a given corpus. In Section 2.3.1, we describe different neural network applications to building language models.

2. **Classification:** We also consider a neural architecture to specifically address the multi-class classification problem of natural language inference. The models extracted during the first phase are used to initialize the neural models described for the second phase.

We will now provide a brief overview of neural networks, in order to provide adequate background for a discussion of the language modeling and the target architecture of our inference solutions.

2.2.2.1 | Neural Networks

The most primitive form of neural network is the Feed-Forward network (FFNN), which is composed of an input layer that encodes the training data, a hidden layer, an intermediate layer of weighted values, and the output layer, which presents the result of the learning process. Neural networks with single hidden layers are described as *shallow*; conversely, networks with several hidden layers are described as *deep*. Thus, neural networking models with several hidden layers are frequently described as *deep learning* approaches (Duda et al., 2012). Figure 2.1 shows the *FFNN* alongside more complex neural architectures. Weights from hidden layers are combined by an activation function, the choice of which is critical in neural network design. Several options exist, including Binary Step Functions, Linear Activation Functions and Non-Linear Activation Functions (Duda et al., 2012).

Rumelhart et al. (1986) introduce the idea of back-propagating neural networks, in which weights for hidden layers are repeatedly adjusted based on the output vector. Within back-propagation, we evaluate candidate weights in terms of the desired output; this is achieved by calculating a loss function, quantifying a model's loss (or gain). Consequently, neural network weights are adjusted at every learning step based on the the loss function to achieve a global maximum. In Section 2.3.2 we discuss different loss functions for aligning word embeddings.

Several variants of the initial *FFNN* have been proposed to address different learning tasks. These variants typically involve *deeper* networks, broadly described as deep neural networks (DNN): these include Convolutional Neural Networks (CNN), networks designed to learn temporal data; and Recurrent Neural Networks (RNN), networks designed to learn sequential structures (Gers et al., 1999; Liu et al., 2017; Wen et al., 2016; Yin et al., 2017). Figure 2.1 shows the three neural architectures side-by-side.

Convolutional Neural Networks (CNN) employ the concept of a sliding window (or kernel) function applied to an input matrix in order to learn weights. Given an input

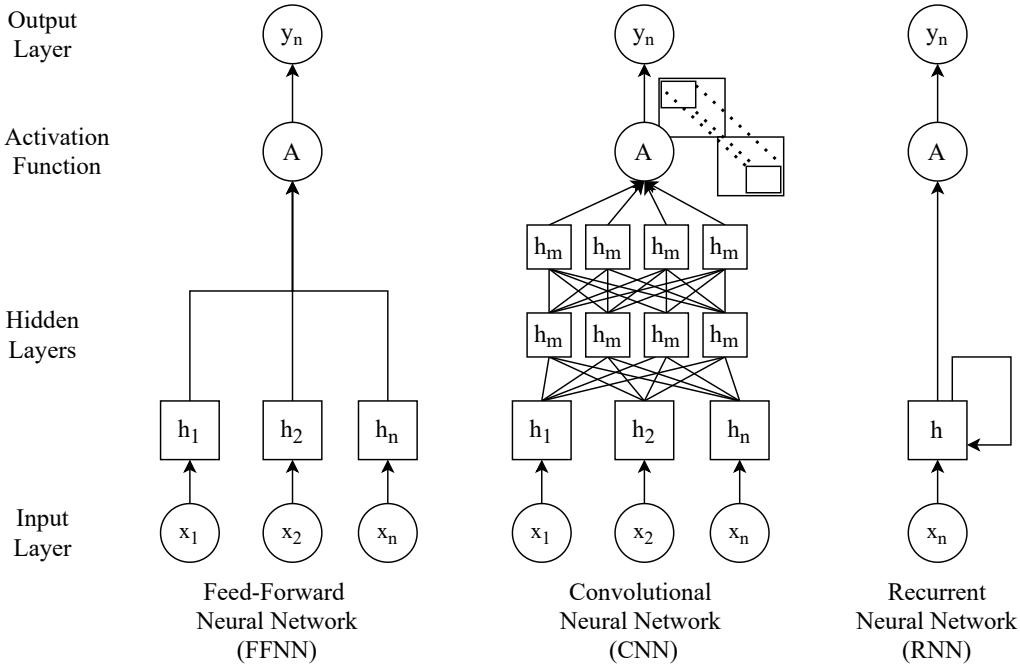

Figure 2.1: The Feed-Forward, Convolutional and Recurrent Neural Architectures, side by side, distinguished by the structure within hidden layers.

matrix, the kernel is applied on a subsection of the matrix to compute elements of a convolved feature matrix. Convolutions over the input layer compute the output. These convolutions are expressed as one or more hidden layers (Duda et al., 2012; Liu et al., 2017). After convolution, CNNs apply *pooling layers* to sub-sample the resultant output. We explain this process in terms of an NLP task below.

Given an input sentence of length n with a vocabulary size of d, we represent the sentence as an $n*d$ matrix input. Every row in the matrix input would correspond to one word. A kernel of width d and height m would select m words at every convolution. Convolution vectors are sub-sampled and fed to an activation function to produce the output.

The CNN model disregards the position of input entities; this intuition may not be useful within language processing, where the location of words within a sentence is of great importance. However, there are cases where such an approach is desirable. Vu et al. (2016) report a higher performance for CNN for relational classification tasks. Moreover, the CNN approach reportedly performs better with longer text fragments and local context dependencies (Adel and Schütze, 2017; Wen et al., 2016; Yin et al., 2017).

Recurrent Neural Networks (RNN) allow previous outputs from hidden layers to be uses as inputs for other hidden states. This has several benefits. Hidden layers are trained within the context of their previous inputs; this is beneficial in order to represent relationships between sequences of inputs. On the other hand, the size of the hidden layer is not limited as in the CNN case, where a kernel size must be pre-defined.

In figure 2.2, we show an RNN whose hidden layer feeds to itself; the RNN can also be represented as FFNN, where each neuron within the hidden layer is fed values from another hidden layer. The relationships between hidden layers can take different forms: one-to-many, where outputs are feed hidden layers; many-to-one and many-to-many , where hidden layers feed other hidden layers (Duda et al., 2012; Gers et al., 1999; Liu et al., 2017; Luong et al., 2016).

The ability to leverage RNNs to represent sequences is critical to neural natural language processing, particularly in order to learn dependencies along long sequences. However, the initial construct does not perform this function well when combined with backward propagation; this is described as the vanishing gradient problem (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997). As the network adjusts weights to search for the global minimum of the cost function, weights are multiplied across several recurring cells. Such multiplications cause gradient values to shrink such that the values become inconsequential; this causes the network to lose memory of previous sequences.

Hochreiter and Schmidhuber (1997) initially propose the idea of a Long Short-Term Memory (LSTM) variation of the RNN, which is the predominant form of recurrent networks in contemporary literature. The idea defines an LSTM *memory cell* as the basic building block of a hidden layer; a hidden layer is of composed of several LSTM cells.

The cell's architecture is composed of several gate units; an input gate, an output gate and a forget gate, representing read, write and reset operations for the memory. Such gates deduce which information is to be discarded, while processing new states to be fed to other hidden states (Gers et al., 1999). Other cells, such as Gated Recurrent Units (GRU), with a different gate configuration have also been proposed (Cho et al., 2014). The application of LSTM and GRU cells within RNN and also CNN architectures is largely responsible for the recent renaissance of *deep learning*. Comparison between the two architectures and hidden cell types for different tasks remains an active area of research (Yin et al., 2017).

Another concept which builds upon the idea of LSTMs is that of **neural attention**. Initially proposed by Bahdanau et al. (2015), attention is intended to calculate a set of attention weights which quantify how much a word from one sentence attends to another; particular outputs may naturally give higher importance to particular inputs.

This is particularly useful within machine translation, where each word in the output sentence specifically attends to particular words in the input sentence (Sutskever et al., 2014). The attention $\alpha^{<t,t'>}$ is defined as the amount of attention an output y^t should pay to an input $a^{t'}$; this is achieved by training a separate FFNN concurrently. Rocktäschel et al. (2015) apply the concept of attention to the natural language inference problem, suggesting word-by-word attention across the premise and hypothesis.

2.2.2.2 | Transformers

Recent optimizations for advancements in neural models has given birth to a new breed of neural architectures, which forego the recurrence and convolutions concepts discussed in the previous section. Vaswani et al. (2017) propose exclusively leveraging a neural network's *attention* to increase the reference window of a neural network, providing a theoretically-infinite memory mechanism. Chiefly, the framework presents a decoder which generates the output in steps using the encoder's representation and previous decoder outputs. Both the encoder and decoder a series of self-attention layers and FFNN utilizing the ReLu activation function. Input vectors are augmented with positional encoding, giving the encoder knowledge of the sequence. This is in contrast with a traditional recurrent approach, which learns the sequence at different time steps.

The key innovation in this method is the self-attention layer. In self-attention layers, inputs are each fed into distinct input layers to compute distinct query, key and value vectors. The query and keys are used to compute a score matrix, which determines how much a word attends to another. Scaled attention scores are fed to a Softmax activation function, and multiplied by the values to gain the output. The self-attention mechanism is repeated on several vectors throughout the process, in conjunction with position-wise feed-forward networks.

Transformer based models have achieved state-of-the-art results on several tasks (Devlin et al., 2019). However, the achieved results are poorly understood (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019), and some researchers have indicated that attention-based models rely on biases and shortcuts to provide predictions (Kovaleva et al., 2019).

2.2.3 | Neural Network Inference Architectures

Challenge datasets such as the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) provide a rich reference point for monolingual neural inference architectures. The latter of these presents training examples which are more representative of linguistic phenomena such as tense and belief (Nangia et al., 2017). Several approaches have been

proposed to address the inference task, employing a variety of techniques which include attention (Wang and Jiang, 2015), memory (Munkhdalai and Yu, 2017) and parse structure (Mou et al., 2016). In this section, we give a brief overview of two classes of models. The first class employs recurrence, typically using bidirectional LSTM networks, while the second class employs attention-based models as a form of neural alignment (Parikh et al., 2016).

Bowman et al. (2015) initially propose a baseline architecture that makes use of sentence embeddings in the input layer. The network accepts two concatenated 100d (embedding of 100 dimensions) inputs which are in turn fed to three consecutive *tanh* layers of 200d. In turn, the output is fed to a softmax classifier, trained jointly with the sentence embedding model. The approach is benchmarked against a generic bidirectional LSTM (BiLSTM) RNN. Both models are initialized using GloVe (Pennington et al., 2014) word embeddings (see Section 2.3), and trained using the AdaDelta (Zeiler, 2012) optimizer.

Building upon ideas within the baseline, another sentence encoding model presented by Liu et al. (2016a) also employs a biLSTM with average pooling. The authors also introduce an attention mechanism operating on the same sentence, quantifying which words attend to each other within the source and target sentence. This is in contrast to initial formulations of attention, which deal with words attending to each other between the source and target sentences (Sutskever et al., 2014). Inter-attention relates the intuition that humans can approximate meaning from a single sentence. The sentence vectors, together with their attention weights, are passed through a sentence matching module composed of three relations: the concatenation of the two representations, the element-wise product and the element-wise difference (Mou et al., 2016).

Parikh et al. (2016) also make use of an attention model, with three distinct phases. In the first two phases, elements of the input sentences are aligned and subsequently compared to produce a set of vectors for each sentence. These phases are constructed through a feed-forward network which normalizes the attention weights and compares aligned phrases, resulting in vectors which represent the aligned sub-phrases for each sentence. Finally, the resultant vectors are aggregated and used to predict the entailment relation. The comparison vectors are aggregated using a summation which is fed to a classifier. Other similar approaches suggest using BiLSTM cells instead of feed forward networks for constructing attention weights (Chen et al., 2017a).

A different class of neural architectures exists in Transformer models, which were introduced in the previous section. Such approaches exclusively employ attention outside of the traditional RNN and CNN constructs. However, as discussed in the previous section, results behind such models are poorly understood (Jain and Wallace, 2019; Kovaleva et al., 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019). Thus we con-

sider such models to be outside the scope of our research, as we have chosen to focus on deep neural networks exclusively. We outline the results achieved by these methods, as well as the techniques employed in table 2.2.

Approach			Test Accuracy	
Publication	**Model**	**Attention Mechanism**	**SNLI**	**MultiNLI**
Qian Chen et al. '17	Inter Attention with BiLSTM Encoders	Inter-Attention	85.5%	74.9%
Tao Shen et al. '17	Bidirectional Self-Attention	Self-Attention	85.6%	67.1%
Parikh et al. '16	Decomposable Attention	Self-Attention	86.5%	-
Nie Bansal et al. '17	Stacked Bidirectonal LSTM RNN	None	86.1%	73.5%
Bowman et al. '16	SNLI Baseline: Sum of Words	None	75.3%	-
Bowman et al. '16	SNLI Baseline: LSTM RNN	None	77.6%	-
Radford et al. '18	Transformer Model	Attention	89.9%	81.4%
Wang et al. '19	sturctBert Transformer Model	Attention	91.7%	-

Table 2.2: An overview of results for different inference architectures.

2.3 | Cross-Lingual Language Representations

In the previous section, we discussed language inference models in terms of two different components; a word embedding component, which aims to model inferential relationships between words; and an architecture component, which carries out the final classification. We now focus on advances in language modeling brought about through the application of neural networks, as opposed to traditional feature engineering approaches.

2.3.1 | Word Embeddings

Several models dependent on hand-crafted features incorporate *N-Gram* and *term frequency* for modeling language as discussed in Section 2.2.1. Typically researchers have employed word and N-Gram co-occurance to model sequences of words. However, as the length of modeled sequences increases, an ever-increasing number of features is required to accurately build generalizations, creating high-dimensional spaces (Bengio et al., 2003). This phenomenon, commonly referred to as the 'curse of dimensionality,' necessitates a different approach to language modeling, where one aims to learn an entire distributed representation of words using distribution probabilities. Such representations take the form of distributed word vectors or word embeddings, describing a word's position within a vector space learned using a probabilistic approach.

Several models exist for learning distributed word representations, including Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer et al., 1998), Latent Dirichlet Allocation (LDA) (Zhila et al., 2013), Feed Forward Neural Networks (FFNN) (Rumelhart et al., 1986) and Recurrent Neural Networks (Mikolov et al., 2010). Rumel-

hart et al. (1986) initially propose the application of neural network back-propagation to learn such word representations, which has been applied successfully to several NLP tasks (Collobert and Weston, 2008; Mikolov et al., 2012; Schwenk, 2007; Weston et al., 2011).

The initial formulation of the Feed Forward Network for constructing word embeddings involves encoding sentences as neural network inputs. Given a word as part of a sentence, *N* previous words are encoded and projected to a layer of *NxD* dimensionality, where *D* is the selected length of the vector. Given an increasing *N*, (size of the projection and hidden layers) complexity increases significantly with either *N* or the source vocabulary. While more computationally efficient, a Recurrent Neural Network possesses similar attributes. While precise, these techniques alone pose the challenge of increasing computational complexity as source data increases (Mikolov et al., 2013a).

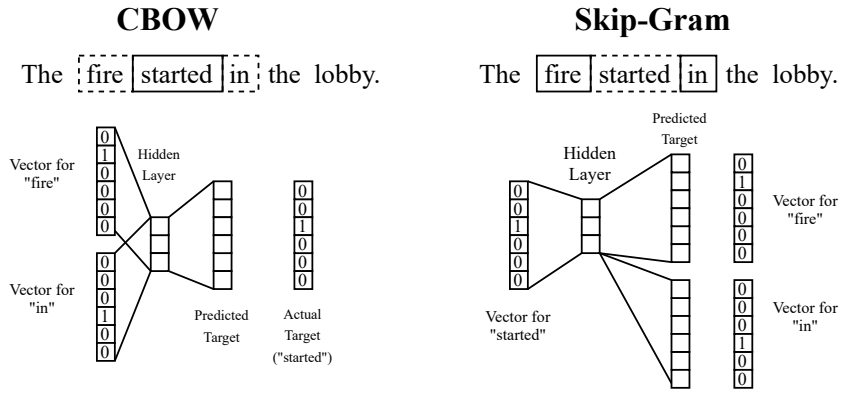

Figure 2.2: A CBOW model identifies a central word ("*STARTED*") from a context of words ("*FIRE*", "*IN*"), while a Skip-Gram model identifies context words ("*FIRE*","*IN*") from a central word ("*STARTED*"). In this example *N=3*.

Mikolov et al. (2013a) initially introduce two techniques for simplifying neural networks in order to reduce the classifier to log-linear complexity. Instead of encoding entire sequences of prior words, we aim to model the word's context. Both models use a two layer neural network. Consider that a target word is surrounded by context words, the two approaches are described as follows (Goldberg and Levy, 2014):

- **Continuous Bag of Words (CBOW)**: Given the context words, the model attempts to identify the central word. The model aims to maximize the probability of a central word using context word co-occurrences within a distance *N*. The network is composed of an input layer, a hidden layer and an output layer (omitting the

projection layer). Context words are encoded as inputs and the predicted output denotes the central word.

- **Continuous Skip-gram Negative Sampling Model (SGNS)**: Given the central word, the model attempts to identify potential surrounding words. The skip-gram model aims to maximize the probability of a word, based on co-occurrences within a *[-N,+N]* window. The input contains an encoded central word, with the target vectors denoting surrounding words.

The quality of a word embedding is typically determined by evaluating the model's lexical induction capability. An effective monolingual word embedding would map other related words close to each other. Within an effective monolingual word embedding, a word's nearest neighbours typically include synonymous words. From a quantitative perspective, the measured cosine similarity between two words (n and m) should be indicative of semantic relatedness:

$$\cos(n, m) = \frac{n \cdot m}{\|n\|\|m\|} = \frac{\sum_{i=1}^{n} n_i m_i}{\sqrt{\sum_{i=1}^{n} (n_i)^2}\sqrt{\sum_{i=1}^{n} (m_i)^2}} \tag{2.1}$$

An over-concentration of nearest neighbours, based on the respective relatedness metric, reduces the quality of word embeddings. Radovanovic et al. (2010) describe this phenomenon as hubness, where some observations are surrounded by many other observations, impacting the quality of embedding spaces.

2.3.1.1 | Frameworks for Learning Word Embeddings

Figure 2.2 shows the difference between the two models, with the Continuous Skip-Gram model representing the opposite of the CBOW model. Various frameworks have been proposed using such models. Mikolov et al. (2013b) introduce Word2vec, a framework utilizing the Skip-gram method with a Hierarchical SoftMax activation function and Negative Sampling. Data fed to the network is pre-processed by replacing named entities with unique tokens; this allows named entity phrases to be modeled without significantly increasing the vocabulary size. The resultant model achieves accuracy ranging from 66% to 72% on a word analogy task, reporting that the Skip-gram approach significantly outperforms complex neural models. Moreover, the model demonstrates additive compositionality, allowing reasoning to be extracted from arithmetic vector calculations.

Another application is the Global Vectors (GloVe) framework (Pennington et al., 2014), which leverages a hybrid method, applying a co-occurrence matrix alongside the

skip-gram model. The primary motivation for this approach is to learn sub-structures within the vector space, to increase performance in word analogy tasks. The model performs slightly better on the word analogy task, although requiring more computational resources.

Such word embedding frameworks have shown promising results in speech recognition and machine translation, but they do not convincingly outperform prior state-of-the-art techniques in all tasks (Baroni et al., 2014; Joulin et al., 2016b; Wang and Manning, 2012). One drawback of such frameworks is poor performance in relation to rare, or out-of-vocabulary, words. A key innovation which treats out-of-vocabulary words is the application of similar techniques to character level n-grams.

Joulin et al. (2016a) propose Fasttext, a framework for learning character-level word embeddings. Whereas previous systems such as Word2vec and GloVe learn vectors for n-grams of words, the proposed system learns characters within the words alongside the entire word. Given the word *"<grand>"*, the Fasttext framework would consider the tokens *[<g,gr,ra,an,nd,d>]*, where < and > denote the start and end of the word. This would allow the model to learn sub-word information. If the model encounters the out-of-vocabulary word *grandiose*, the rare word can be embedded near the word *grand* due to shared sequences of text. Within word-level word representations, rare words share fewer context words, resulting in poor results, but sub-parts of words neighbour a greater number of tokens. However, increasing the granularity of the n-gram model increases the computational requirements.

2.3.1.2 | Contextualized Word Embeddings and Attention Models

Another approach in the area of word representations is *contextualized* word embedding; in such methods, representations are processed as functions of an entire sentence, rather than singular words or sub-words. The main benefit of such a approach is dealing with *polysemy*, the various possible meanings of a word (Peters et al., 2018). This method shows significant improvements in the NLP tasks of named entity recognition, entity co-referencing and question answering. More recent innovations build upon the idea of contextualized word embeddings, adopting different approaches. For example, one such trend leverages a neural network's *attention*, specifically the weighted average of neural network inputs, in order to derive a composite embedding. We briefly outline attention and transformer models in Section 2.2.2.2. Attention-based models rely on deriving context within the same sentence; attention functions as a relationship score for inter-sentence relationships (Devlin et al., 2019; Radford et al., 2018b). Such models have achieved a new state of the art on several tasks, including inference. However, several

researchers indicate that the results behind these improvements are poorly understood (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019).

Kovaleva et al. (2019) quantify the influence of attention on the model's increased performance, indicating that attention is not directly responsible for improvements, particularly for the task of natural language inference. At the time of writing, the benefits of attention-based models remain contested, with indications that improvements could possibly be attributed to undiscovered phenomena. Moreover, the ability of such approaches to capture meaning has been called into question. Bender and Koller (2020) discuss attention models within the context of *form*, representing the concrete representation of text, and *meaning*, representing relationships which relate to concepts external to language. The authors argue that given that such approaches do not sufficiently address the meaning within different asks in spite of achieving new, state-of-the-art results.

2.3.2 | Cross-Lingual Word Embeddings

Recent advances in the area of word representations have proven highly effective in improving the latest deployments of natural language processing. One of the primary advantages of vector representations is their capability to perform arithmetic logic that can express the relationship between words (Goldberg and Levy, 2014; Mikolov et al., 2013b). While such research has predominantly focused on the English language, there is a renewed interest in cross-lingual natural language processing, fueling a sub-line of research related to creating cross-lingual word embeddings. This idea is attractive for two reasons. Firstly, a single word embedding may be used across different languages, reducing training cost; secondly, high-performing models within a high-resource language (such as English) may be leveraged to improve models in other low resource languages (such as Swahili).

Cross-lingual embeddings aim to create a common representational space containing text from both languages. This is achieved by aligning two monolingual models (Artetxe et al., 2018; Lample et al., 2018; Smith et al., 2017; Xing et al., 2015), or by simultaneously learning a common representational space (Gouws and Søgaard, 2015; Klementiev et al., 2012; Kočiskỳ et al., 2014; Xiao and Guo, 2014). Ruder et al. (2019) provide a comprehensive overview of the state of word embedding alignment for cross-lingual tasks. In this section, we summarize the findings while delving deeper into the predominant approaches.

Approaches differ along three different dimensions; the linguistic unit of alignment; the data leveraged to learn the embedding; and the method by which the representations are learned. We briefly outline the differences in terms of linguistics units and

source data in order to adequately scope our discussion. Vector representation learning can occur at different linguistic units such as *words*, *sentences* and *documents*. In each case, we learn vectors which represent different units and map them to a cross-lingual vector space. The data used to learn these representations can be *parallel* or *comparable*.

Most research within the area treats parallel word- and sentence-aligned multilingual representations. *Parallel* word-level data is readily available in multiple languages (including low-resource ones) in the form of dictionaries. One example of such parallel data is the Europarl Corpus (Koehn, 2005), composed of direct translations of text fragments from dialogue within European institutions, providing a relatively rich source of data for *parallel sentence alignment methods.*

Whereas *parallel* alignment treats text pairs across different languages, which are typically obtained through translation, *comparable* alignment methods use pairs of corpora treating the same subject or domain which are not direct translations. *Comparable* alignment methods at the word and sentence level may leverage a multi-modal embedding space, using images and linguistic features such as part-of-speech tags (Bergsma and Van Durme, 2011; Calixto et al., 2017; Gella et al., 2017; Gouws et al., 2015). This area of work is under explored, and the multi-modal nature of the work does not directly relate to textual entailment.

At the document level, *parallel* document data, where bilingual corpora consist of directly translated documents, is scarce, particularly when considering low-resource languages and different subject areas. *Comparable* document alignment typically involves learning representations of documents about a common topic in different languages. However, while it is theoretically possible, research has shown that it only provides minor auxiliary value (Ruder et al., 2019). Moreover, document-level alignment methods are contextualized in terms of sentence-level approaches. Thus, we limit our research to word-level and sentence-level parallel alignment, intentionally excluding *comparable* and *document level* approaches. In this section, we first explain approaches within the *parallel word-level* area, and then explain how they are extended to the sentence-level.

Lastly, we must distinguish between *bilingual* alignment and *multilingual* alignment, i.e. the alignment of word representations from two languages rather from than multiple languages. While some research suggests that there are several benefits to considering multilingual application (Duong et al., 2016; Levy et al., 2017), most research is focused on aligning bilingual pairs. The multilingual scenario can be scoped to several bilingual sub-alignments by selecting a single resource-rich language, such as English, as the *pivot* language (Duong et al., 2017). Thus, approaches to bilingual alignment can thus be easily leveraged within a multilingual setting.

2.3.2.1 | Word-Level Alignment

In Section 2.3.1 we introduced the concept of learned word representations as a vector space, discussing the *Continuous Bag of Words* and *Continuous Skip-Gram Negative Sampling* approaches. The approaches in this section employ either CBOW or CSGNS as the core learning mechanisms, depending on the learning objective.

Mapping-Based Alignment The most prominent method for aligning word embeddings, particularly at the word-level, is the mapping-based approach. The approach assumes that the geometric structure of words which are mapped in one monolingual space closely resembles the structure of their translated counterparts within another monolingual space. Thus, we aim to transform one vector space to another, using a seed vocabulary.

Consider source language s and target language t. The learning process aims to learn a transformation matrix $W^{s \to t}$. The seed lexicon is derived from n most frequent words in s, translated to t. The seed lexicons used can be acquired through a variety of methods; some approaches use off-the-shelf data sets (Lazaridou et al., 2015; Mikolov et al., 2013a), while others attempt to learn a lexicon from the corpus using supervised and unsupervised methods (Lample et al., 2018; Smith et al., 2017; Søgaard et al., 2018).

A common method to learn this mapping is a regression which minimizes the mean squared error (MSE) of the transformation matrix $W^{s \to t}$ between a source seed word $x_i^{s \to t}$ and its translation $x_i^{t \to s}$. The loss function can be expressed as follows:

$$\Omega_{MSE} = \sum_{i=1}^{n} ||W x_i^{s \to t} x_i^{t \to s}|| \tag{2.2}$$

Given that the monolingual representations are trained separately, and that the resultant cross-lingual representation is trained as expressed above, the objective can be revised to express the combination of optimizing all three learning processes. Thus the objective for skip-gram negative sampling can be expressed as follows, where X represents the embedding matrices for a seed vocabulary:

$$J = L_{SGNS}(X^s) + L_{SGNS}(X^t) + \Omega_{MSE}(X^s, X^t, W) \tag{2.3}$$

One challenge with this approach is in similarity measures within the monolingual word representation task and the alignment method. The CBOW and SGNS methods proposed use *cosine similarity* to estimate the relatedness of words, whereas the mapping process uses a *Euclidean MSE*. Resultant alignment causes high monolingual variance. Xing et al. (2015) propose limiting the regression, using an orthogonal normalization

constraint, in order to address this issue. This constraint allows us to create more consistent mappings (Smith et al., 2017; Zhang et al., 2016).

Haghighi et al. (2008) propose an alternative approach to the regression method. Rather than aligning two monolingual word embeddings, the source and target language are mapped to a single shared space. The proposed *Canonical Correlation Analysis* determines the *orthogonal linear* correlations for each corpus, learning two separate transformation matrices. In turn, vectors with the highest correlation are selected. However, improvements gained from this approach are not widespread across all language pairs, and the approach is considered equivalent to orthogonal constrained regression (Artetxe et al., 2017).

The above approaches frequently incorporate nearest-neighbour retrieval to refine or evaluate the resultant word embeddings. Translations for source words are sourced from the cross-lingual mapping by selecting the most similar words using a similarity measure (Artetxe et al., 2018; Lample et al., 2018). In turn, this retrieval can be used to refine the model by incorporating the translation into the seed lexicon. To enhance the quality of retrieval, a symmetry constraint is typically imposed; a pair is considered to be a translation if they possess mutual nearest neighbours in the vector space.

Significant work has revolved around improving the retrieval process in order to reduce the challenge of 'hubness,' (Radovanovic et al., 2010) or the over-concentration of nearest neighbours. Smith et al. (2017) propose a globally corrected retrieval process, normalizing the probability over source words. Lample et al. (2018) propose a new measure, cross-domain similarity local scaling (CSLS) as an alternative to cosine similarity. The CSLS metric builds a bipartite neighborhood graph, with the aim of maximizing the selection of word pairs within the selected neighbourhood.

Pseudo-bilingual Corpora and Joint Methods While less popular, there are alternative approaches to learning word-level mappings. Xiao and Guo (2014) propose to construct pseudo-bilingual corpora. Instead of learning the mapping between the source and target language, a pseudo-bilingual corpus is constructed by replacing words in a source language with their translations. Similarly, Gouws and Søgaard (2015) construct a pseudo-bilingual corpus by concatenating the corpora for both languages, replacing words whose translations are available with their translated equivalent. Such approaches have shown increased benefits with dealing with polysemy in the cross-lingual context, and also show promising results when applied to low-resource languages. A disadvantage of this approach is that such word embeddings can only be used for a single alignment task, whereas mapping-based approaches can reuse their monolingual component across several alignments. Mapping-based approaches are more computa-

tionally efficient on several levels: both in terms of training on a concatenated corpus, and in terms of the re-usability of models.

The idea of constructing pseudo-corpora has also been used in conjunction with mapping-based approaches. Each of the methods discussed above seeks to optimize the same task; the mapping-based approach seeks to optimize monolingual losses alongside the cross-lingual regularization term, while pseudo-bilingual approaches aim to optimize the single cross-lingual loss through data manipulation. A particular line of research is centred around applying both approaches in tandem (Klementiev et al., 2012)(Kočiskỳ et al., 2014). While alternatives to mapping-based approaches are interesting to explore in the case of low-resource languages, the approaches described above are theoretically equivalent.

Ruder et al. (2019) show that such approaches can be reduced to a *Constrained Bilingual Skip-Gram* approach, which makes use of negative sampling. The word embedding models initially suggested by Mikolov et al. (2013a), leveraging skip-gram negative sampling, are equivalent to pseudo-bilingual sampling.

2.3.2.2 | Sentence-Level Alignment

As discussed in Section 2.3.2, parallel sentence-level alignment requires a corpus of sentences which are directly translated between the source language s and the target language t. Such data is expensive to collect, particularly for low-resource languages; however, one rich data set comes in the form of the Europarl corpus. We discuss different approaches to parallel sentence level alignment, some of which are conceptually similar to word-level mapping-based approaches.

Word-Alignment Matrices: Considering two sentences in the source and target language, S_s with length n and S_t with length m, words from S_s are mapped to words from S_t (Dyer et al., 2013). Thus we capture an alignment matrix of words, $M^{s \Rightarrow t}$, which denotes the occurrence of alignment.

Such approaches minimize the difference between the source embeddings with the corresponding alignment matrix, and the target embeddings (Zou et al., 2013). The method draws a direct parallel to the mapping-based approach of parallel word-level data; with the slightly modified minimum squared error being expressed as follows, where M is the alignment matrix for given seed vocabularies X:

$$\Omega_{s \Rightarrow t} = ||X^t - M^{s \Rightarrow t} X^s|| \tag{2.4}$$

$$J = L(X^t) + \Omega_{s \Rightarrow t} = ||X^t - M^{s \Rightarrow t} X^s|| \tag{2.5}$$

Drawing further comparison to mapping-based word-level alignment, several researchers have sought to better fulfil the optimization goal by applying constraints to the alignment. The adjacency matrix can be factorized to be constrained accounting for a separate monolingual objective, contextual relationship or sparseness (Huang et al., 2015; Shi et al., 2015; Vyas and Carpuat, 2016).

Sentences as Sums of Word Embeddings: Other research proposes considering the sentence embedding as a sum of word embeddings. Hermann and Blunsom (2013) propose conceptualizing the representations of sentences in s and t as the sum of embeddings, optimizing the minimum distance between the sentences. Similarly Lauly et al. (2014), encode the source and target sentence within a single auto encoder, optimizing the cross-lingual task jointly.

Skip-Gram Models: Another approach is to extend the monolingual skip-gram model (Mikolov et al., 2013a) to learn cross-lingual embeddings. Analogous to the orthogonal regression mapping-based word-level alignment, these models optimize for losses (L) for each language together and an additional cross-lingual regularization term (J):

$$J = L^s + L^t + \Omega$$

Research in the area of parallel sentence-level alignment is lacking in comparison to its word-level counterpart. Moreover, there is no definite benefit to adopting an exclusively sentence-level approach; however there are indications that augmenting sentence-level alignment with word-level alignment can reduce the impacts of hubness (Radovanovic et al., 2010)(Ruder et al., 2019).

2.4 | Summary

In this chapter, we discussed literature relevant to the task of cross-lingual natural language inference. We observe that over the past decade, natural-language processing approaches in general have been overhauled significantly. The introduction of neural-network models and word embeddings have ushered in a renaissance of new advances and state-of-the-art approaches. This is particularly evident with the recent proliferation of transformer models over the past year, which forego the recurrent and convolutional neural network approaches considered modern and state-of-the-art until recently. However, as discussed, such models are poorly understood.

Cross-lingual natural language processing is a relatively under-researched area, with most task-specific research hitherto focusing only on monolingual cases; neural network architectures and word embeddings are still active and under-explored research areas in the cross-lingual context. The task of cross-lingual natural language inference is also under-explored; the corpus and benchmark provided through XNLI Conneau et al. (2018) is the first of its kind. To our knowledge, there are no alternative RNN or CNN architectures aside from the XNLI benchmarks. The only other architectures which exist are generalized multi-task transformer models, proposed as baselines in the XTREME Hu et al. (2020) task. The two tasks will serve as a strong baseline for our research.

Both XNLI and XTREME were proposed within the past two years, while the XTREME task is currently open at the time of writing. Thus, we conclude that the research area is nascent and rapidly evolving, with strong potential for novel contributions.

Within this section, we dedicated a considerable amount of attention to monolingual neural network approaches, particularly word embeddings and neural architectures. This serves as an important background to source ideas and concepts which can be adapted within the cross-lingual context. Moreover, we also dedicated a considerable amount of attention to word alignment processes; these methods underpin the quality of word embeddings proposed in the XNLI baselines.

In the next section we will outline our system's design based on our research through which we conclude that:

- The current state of the art (SOTA) in cross-lingual NLI is the transformer-based XLM (Conneau and Lample, 2019) model, with an accuracy of 79.2%. Human cross-lingual NLI has an accuracy of 92.8%.
- While showing strong results, transformer models are not as well understood as recurrent and convolution neural networks. The SOTA for LSTM-based English monolingual NLI is 71%.

- Translation based approaches, which are considered undesirable for cross-lingual NLI approaches, currently outperform all cross-lingual transfer approaches.
- Word-level word embedding alignment is the most popular word embedding alignment method. Research indicates that there is potential in using word-level and sentence-level alignment methods simultaneously (Ruder et al., 2019).

3

Methodology

Our study poses two research questions. First, we aim to design a system where inference knowledge learned in a *high-resource* language is transferred to other languages. Second, we wish to examine whether current neural architectures can be applied in an inference scenario where the premise and hypothesis are in different languages.

In this chapter, we detail our approach for achieving these goals, discussing our choices in terms of the research cited within the previous chapter, in order to achieve our objectives (restated below):

1. Build a number of word embedding models catered to our aims, particularly a set of aligned word embeddings for cross-lingual natural language processing.

2. Design and adapt neural architectures for transfer learning within the inference task. We explore two options, the first consisting of a *Cross-lingual transfer* scenario, and the second considering a situation where the inference task is ported to a translation task.

3. Design and adapt neural architectures for approaching the cross-lingual inference formulation, where the premise is in one language (English) and the hypothesis is in another language.

We begin by reviewing the overall solution architecture, describing how each component relates to the other and the application of our architecture for different tasks. Next, we describe each component in detail, reviewing the process for learning word embeddings, and two neural architectures proposed in our research.

3.1 | Solution Overview

Our system consists of a central inference model which is trained for monolingual entailment classification on one language L_1; the model is then fine-tuned for an inference task in a second language L_2. In our examples, L_1 is English and $L2$ is Spanish. However as discussed in Chapter 4, we evaluate our approach using a number of different languages.

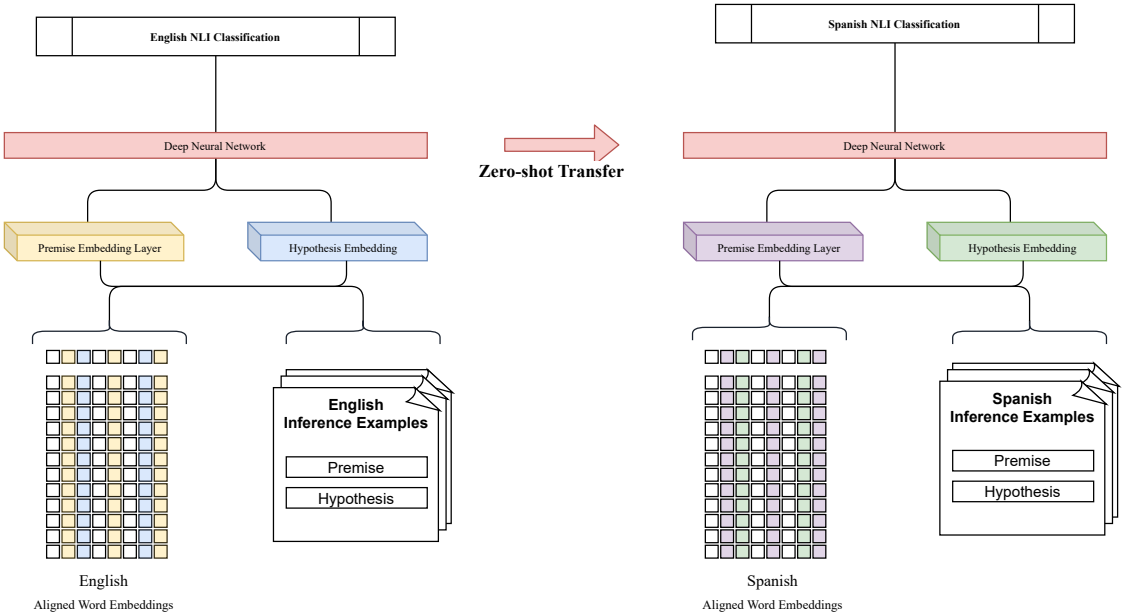

Figure 3.1: The target architecture, primarily consisting of a neural architecture which is trained on English and adapted or fine-tuned to another language. The network employs sentence representations built from learned word embeddings. Within our solution, we first train a neural network using English word embeddings, and an English NLI data-set. Next, we re-use and fine-tune the model using word embeddings and inference examples in the source language (for example, Spanish). The figure shows the fine-tuning process for the transfer learning case, where a model learned in English is fine-tuned to address inference in Spanish. Within subsequent sections, we explain how we modify the data sets used within this example to address other cross-lingual inference scenarios.

Each architecture is composed of three distinct components: an embedding model, which creates a language model based on an input corpus; a sentence encoding component, which leverages the embedding model to represent sentences; and an neural network architecture which learns the inference relation or is fine-tuned for the target language. The latter of these components is adapted and fine-tuned based on the language.

In Section 3.2 we outline our choice of word embeddings and the training and alignment process used, specifying how our approach differs from previous approaches. In Section 3.3, we compare and contrast two different neural network architectures employed within our solution.

3.1.1 | Data sets

The *MultiNLI* (Williams et al., 2018) data set is used to train the initial inference model in English. Fine-tuning is performed in English and other languages using data sourced from the *XNLI* data set (Conneau et al., 2018), which provides examples in fourteen different languages alongside English. Alongside the *XNLI* data set, fine tuning is performed using two additional data sets, constructed to tackle a total of three distinct scenarios. We outline each scenario below.

In our first scenario, we address the *Cross-lingual transfer* learning task using the original *XNLI* data set. An inference model learned on one language is transferred and fine-tuned to perform inference on a second language, as shown in Figure 3.1. This application is bench-marked against a second application in our second scenario, where inference examples are translated to English. We do this by constructing a second data set, which we call *TRAN-XNLI*, where examples in the original *XNLI* data set (Conneau et al., 2018) in languages other than English are translated to English. As in Figure 3.1, a model is adapted and fine-tuned to perform inference on the target language, whose examples are translated. The translation is carried out using the Google Translate API (Google, 2006) and the original English-language examples are omitted from the data set. In this second scenario, we are effectively porting the inference task to a machine-translation task. This approach has shown to be the strongest baseline in previous research (Conneau et al., 2017a), and will therefore serve as an important benchmark for evaluating the results of our own research.

For our final scenario, we construct an additional data set where the premise is in English and the hypothesis is in another language, also using examples from the original *XNLI* corpus. We refer to this corpus as *CROSS-XNLI*. The data set is constructed by iterating over all non-English inference pairs, and replacing the premise by their English counterparts. Similarly to the *TRAN-XNLI* data set, original English-language inference pairs are also omitted from this corpus.

Our solution is unique in terms of previous solutions to the inference task. Conneau et al. (2018) propose baseline architectures which employ sentence embeddings and aligned encoders to address the *Cross-lingual transfer* and *Translation* scenarios described above. Such sentence embeddings are high dimensionality, consisting of 1024

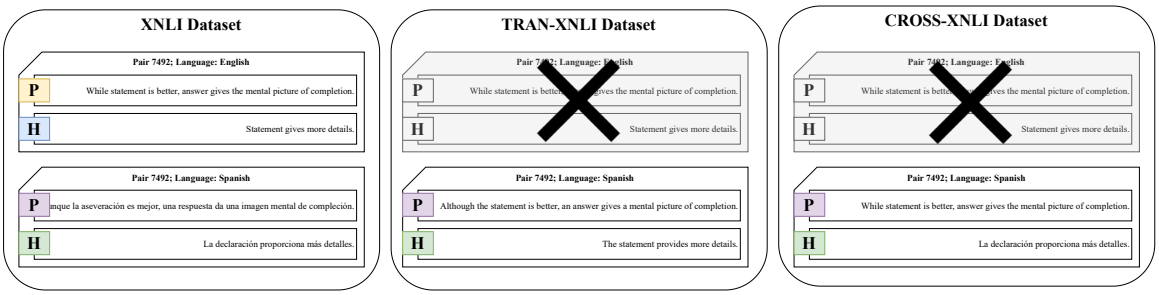

Figure 3.2: Alongside the original *XNLI* data set used for fine-tuning, we construct two additional benchmarks; *TRAN-XNLI*, which ports the inference task to a translation task; and *CROSS-XNLI* which addresses pure cross-lingual inference.

dimensions. Conversely, our solution employs word embeddings which results in a training process which is faster and consumes less resources. Moreover, our solution employs alignment at the word embedding layer as opposed to the encoding layer. We discuss the implementation details for these components in the next section.

3.2 | Cross-Lingual Word Embeddings

The first task in applying our solution architecture is to represent language in a way which can be employed by a neural architecture. In Section 2.2 we introduce the concept of word embeddings, which represent words as vectors within a high-dimensional space. Within our research, we employ the fastText (Joulin et al., 2016a) framework to learn word representations from input corpora in different languages from the *XNLI* corpus. Subsequently, we align these embeddings to a common vector space, in order to provide a single cross-lingual representation of words. The aligned word embeddings which result allow us to encode sentences in different languages to be employed within our target architecture. In this section we briefly outline our decisions in terms of learning such word embeddings and their aligned vector space.

3.2.1 | Monolingual Word Embeddings

Our system employs the fastText (Joulin et al., 2016a) framework for constructing monolingual word embeddings. The framework learns word representations for character *n*-gram vectors using the continuous skip-gram model. We consider this choice of framework along two dimensions; on one hand, the selection of *word representations* as opposed to *sentence representations*; and on the other hand, the particular method for learn-

ing the word representations, which makes use of character *n*-grams.

Conneau et al. (2018) propose a sentence-encoder architecture for cross-lingual inference which uses multilingual sentence embeddings (Schwenk and Douze, 2017). The sentence embeddings are constructed through a separate neural architecture to that which performs inference. Within this architecture, input corpora are tokenized using the *fastBPE* (Sennrich et al., 2016) and fed to a BiLSTM sentence encoder. The resultant vectors are of 1024 dimensions. In contrast, word embedding frameworks typically achieve promising results using smaller vectors of 300 dimensions (Joulin et al., 2016a; Mikolov et al., 2013b; Pennington et al., 2014). Larger representations, as in the case of sentence representations, are intuitively better suited for complex classification tasks, although modeling increased complexity requires increased resources. One of the primary motivations for our study is to perform cross-lingual learning with as little resources as possible. To this end, research has shown that, in some cases, word embeddings may render comparable results using less computational resources (Li et al., 2018). Thus, we select word embeddings to be employed within our solution architecture as an alternative to sentence embeddings. Our approach can thus be contrasted with previous approaches employing sentence embeddings.

Another consideration in our selection process is the method though which the word embeddings are learned. Several frameworks for learning word representations exist (Joulin et al., 2016a; Mikolov et al., 2013b; Pennington et al., 2014; Peters et al., 2018); a distinguishing factor between different examples of such frameworks is their representation of word morphology. Words are typically composed of smaller meaningful units, base words, in conjunction with optional prefixes and suffixes. Learning representations for different languages require us to consider such morphological differences between languages, particularly in cases where less data is available. Within smaller corpora for particular languages, certain words may appear to be rare if morphology is not taken into account (Joulin et al., 2016a). The fastText word embedding framework represents words as character *n*-grams, applying the continuous skip-gram model (described in Section 2.2) on a character level as opposed to a word level. Within our approach, we construct fastText word embeddings for each language learned from a corpus composed of a concatenation of the all Wikipedia (Wikipedia contributors, 2004) entries for the language. Embeddings are constructed for all languages available within the *XNLI* corpus: English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu.

3.2.2 | Aligned Word Embeddings

Given a set of monolingual word embeddings learned on our input corpora, the second phase of word embedding construction involves aligning the embeddings into a single vector space. We apply the approach proposed by Joulin et al. (2018), which aligns neighbourhoods of words for source and target languages L_s and L_t using cross-domain similarity local scaling. The approach considers a bipartite neighborhood graph as shown in Figure 3.3; given a word w, its word vector x and its K connected nearest neighbours in a different language, a graph is constructed such that vertices can be divided into two separate sets U and V and such that every vertex in U is connected to V.

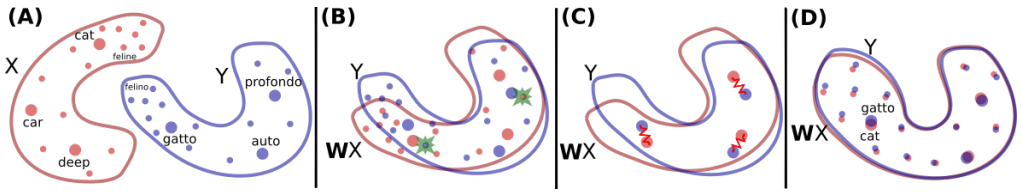

Figure 3.3: The word embedding alignment process using CSLS as described by Conneau et al. (2017b)

Given $N_y(x)$, the set of K nearest neighbours of a word vector x, we calculate the similarity to near target word vectors y using cosine similarity (Joulin et al., 2018):

$$CSLS(x,y) = -2cos(x,y) + \frac{1}{k}\sum_{y' \in N_y(x)} cos(x,y') + \frac{1}{k}\sum_{x' \in N_x(y)} \tag{3.1}$$

As a result, the approach presents a metric for matching words between the two languages; a neighbourhood for a word in source language corresponds to a similar neighbourhood within the target language. The alignment process is assisted by a number of bilingual dictionaries, sourced from the *MUSE* framework (Conneau et al., 2017b). In order to address the issue of *hubness*, where particular words are surrounded by a significant number of neighbours, convex relaxations are applied, constraining the mapping along orthogonal matrices as described in Section 2.3.2.1. Following the alignment process, we obtain a set of word embeddings (with words represented by 300-dimension vectors) which are aligned to a single vector space.

3.3 | Natural Language Inference Architecture

In this section, we briefly outline the neural network architectures which are implemented within our solution. Within our solution, we employ word embeddings as inputs for our neural architectures. In turn, the networks encode the sentences using these embeddings in order to perform the inference task. We implement two neural architectures to assess different approaches to using word embeddings, which are benchmarked for the three selected cross-lingual tasks. Our first model is an attention-based model, which decomposes the task into sub-problems to be processed in parallel. We implement an adapted version of the decomposable attention model initially proposed by Parikh et al. (2016). In Section 3.3.1, we discuss the implementation of this network within our architecture.

We select the BiLSTM architecture as our second solution since, as discussed previously, the *XNLI* benchmark provides reliable benchmarks for our solution, and serve as a guideline for potential target architectures. These benchmarks propose a stacked bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber, 1997) as the baseline architecture for the task (Conneau et al., 2018). Other approaches have employed similar architectures for the inference task (Chen et al., 2017b; Ghaeini et al., 2018) achieving similar state of the art results. However, our work is different from the approach proposed within the *XNLI* benchmark. Whereas the benchmark uses sentence embeddings, our networks employ *word* embeddings as inputs. In this, our approach is influenced by previous monolingual inference architectures employing word embeddings. In fact, our methodology can be considered as adapting prior methods which employ word embeddings to a cross-lingual context.

Novel architectures exclusively employing attention mechanisms have been proposed, achieving promising results for transfer learning (Hu et al., 2020; Vaswani et al., 2017). We briefly outline the rationale behind such models, Transformer Models, within our literature review. However, we choose to exclusively concentrate on employing recurrent architectures to the inference task. Within our evaluation, we refer to the implemented decomposable attention architecture and BiLSTM as *DATTEN* and *BiLSTM* respectively. Below, we outline the implementation of both algorithms.

3.3.1 | Decomposable Attention

The decomposable attention model is inspired by attention mechanisms in sequence-to-sequence (seq2seq) learning (Luong et al., 2016; Sutskever et al., 2014); within such systems, given two encoded sentences, the input states for words are *softly* aligned through

an attention mechanism. Attention serves the purpose of aligning states within a neural context.

Consider two sentences: "*the boy is asleep*" and "*the boy is not awake*." The neural network is tasked with learning dependencies between phrases such that the words "*asleep*" and the phrase "*not awake*" attend to each other. Similarly, the phrase "*the boy*" attends to each other within both sentences, referring to the same subject.

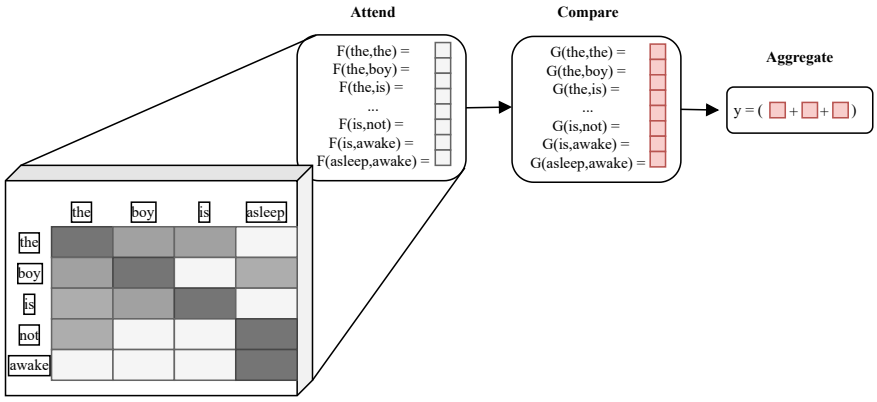

Figure 3.4: The decomposable attention model consists of three distinct phrases; the attend, compare and aggregate phases. (Parikh et al., 2016)

Parikh et al. (2016) propose an approach for attention modeling within inference tasks. The approach follows a three-phase process. In the first phrase, we align sub-phrases of the two sentences using a feed-forward network. The *attend* phase produces a series of aligned vectors which are subsequently compared within a *comparison* phase. These phases can be expressed as two functions: F for the alignment and G for the comparison.

Within the attend phase, F consists of computing the dot product e_{ij}, for which we take the weighted sum to compute attention weights α_j and βj respectively. The attention weights α_j and βj calculate the weighted attention for a sub-phrase in the premise attending to a sub-phrase the hypothesis and vice versa. Subsequently the attention weights are compared to create a $v_{n,i}$ and $v_{n,j}$ as outputs for each sentence:

$$e_{ij} = F(a_i)^T F(B_j) \tag{3.2}$$

$$\begin{aligned} \alpha_j &= \sum_{j=1}^{n} \frac{exp(e_{ij})}{\sum_{k=1}^{n} exp(e_{kj})} \overline{\beta_j} \\ \beta_j &= \sum_{j=1}^{n} \frac{exp(e_{ij})}{\sum_{k=1}^{n} exp(e_{kj})} \overline{\alpha_j} \end{aligned} \tag{3.3}$$

$$v_{1,j} = G(\alpha_i, \beta_i)$$
$$v_{1,j} = G(\beta_j, \alpha_j) \tag{3.4}$$

The two vectors are summed and inputted to another feed-forward network in order for the inference classification to be performed.

3.3.2 | Bidirectional LSTM

We propose an alternative architecture in the form of a bidirectional LSTM (BiLSTM) model (Hochreiter and Schmidhuber, 1997) to be considered in contrast to the decomposable attention model. The chosen architecture is inspired by the initial baseline provided by the XNLI data set (Conneau et al., 2018).

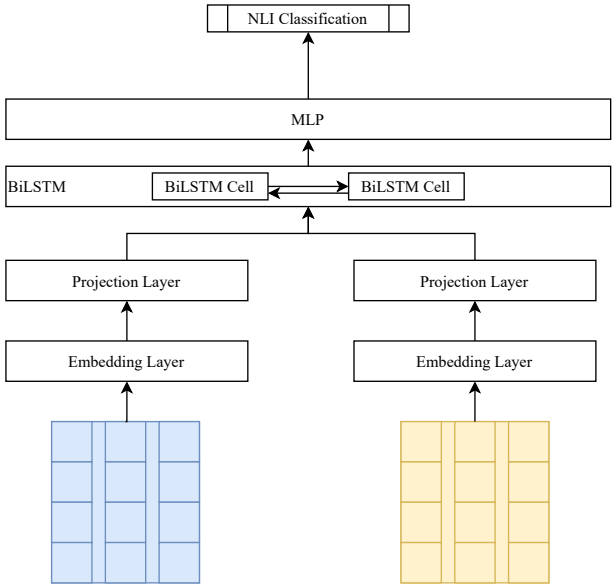

Figure 3.5: The proposed BiLSTM architecture, composed of embedding, projection, bidirectional LSTM and multi-layered perceptron layers.

Our model follows the recurrent neural-network architecture, leveraging LSTM cells. Embeddings for the premise and hypothesis sentences are fed into the input layer of the network, which are then passed to a projection layer. The baseline approach involves the alignment of the encoding layer (shown above as projection and embedding layers), but alignment is carried out within the source embeddings in our case.

3.4 | Implementation Details

In this section we briefly specify the tools used within our implementation. As previously discussed, our development process consists of two phases; learning of word representations and neural architectures for classifying inference. All implementation described in this section is carried out using Python Version 3.6.9.

We learn our embeddings using Wikipedia (Wikipedia contributors, 2004) corpora downloaded from publicly-available Wikipedia database backups for different languages (Wikipedia, 2004). The backup is subsequently cleaned and restructured using the Python Genism package (Řehůřek and Sojka, 2010), providing a clean, plain-text version of the corpus. Word embedding learning and alignment employ the Fasttext framework (Joulin et al., 2016a) and MUSE frameworks (Conneau et al., 2017b), deploying parallel resources for the different languages. Minor modifications to the process to refine alignment are performed in Python.

All processing for constructing the word embeddings is carried out on a CPU optimized AWS EC2 (Amazon, 2006) instance, with 36 virtual cores and 72GB of RAM. Throughout our implementation, we prioritize CPU optimization for our server, since the fastText framework solely relies on CPU cores for processing.

During the second phase of our implementation, we implement neural networks using the PyTorch Python package (version 1.3.0), making use of data sets provided by the TorchText Python package (version 0.7.0). We develop custom data loaders for all our data sets, including MultiNLI (Williams et al., 2018), XNLI (Conneau et al., 2018), TRAN-XNLI and CROSS-XNLI; of these, MultiNLI and XNLI are constructed programmatically, employing the Google Translate API (Google, 2006) where applicable. Our networks are run on a GPU-optimized AWS EC2 instance (Amazon, 2006) with 16 VPU cores, 1 GPU core, 122GB of RAM and 8GB of GPU memory. All training employed the GPU for processing, with a few exceptions to optimize for memory usage.

3.5 | Summary

In this chapter we gave a brief overview of the methodology to be used to explore the cross-lingual inference task towards the stated aims. Three separate components of our methodology can be identified:

1. **Data:** We source and construct the relevant data sets for evaluating our system. We source the *MultiNLI* (Williams et al., 2018) data set for training our neural networks in English, and the *XNLI* (Conneau et al., 2018) data set for fine-tuning the networks to perform the inference task on different languages. Additionally we build two additional datasets: *TRAN-XNLI*, a version of the *XNLI* data set where all example pairs other than English are translated to English; and *CROSS-XNLI*, a version where hypothesis examples in languages other than English are replaced with their English counterpart.

2. **Word Embeddings:** We learn a number of monolingual word embeddings using the Wikipedia corpus for different languages, and align them using cross-domain similarity local scaling.

3. **Neural Architectures:** We propose two different neural architectures: a decomposable attention neural network (referred to as DATTEN here under) and a Bidirectional LSTM (BiLSTM) network to perform the inference task.

Within the next section, we perform a number of experiments to assess the performance of our system. For the *Cross-lingual transfer* task, we train our networks on the *MultiNLI* corpus, and fine-tune the network using the *XNLI* corpus. Subsequently, our other two tasks employ the *TRAN-XNLI* and *CROSS-XNLI* data sets for fine-tuning.

4

Results and Evaluation

In this chapter, we shall outline the evaluation process carried out in our research on cross-lingual natural language inference. The aims of our study is to assess two scenarios for cross-lingual inference: the *Cross-Lingual transfer* scenario, which additionally includes the *Translate Train* experiment; and the *Purely Cross-Lingual Inference* scenario. For each experiment, we present two different neural architectures, a bidirectional LSTM and a decomposable attention network. We also briefly describe the hyper-parameters selected for each model, highlighting the differences in the training processes. Our evaluation is structured as follows:

1. **Aligned Word Embeddings:** We evaluate the word embedding models constructed to encode sentences within our neural networks. In particular, we aim to evaluate the quality of the aligned word embeddings.

2. **Cross-Lingual Transfer:** We evaluate our neural architectures on the cross-lingual transfer by comparing the performance of the two networks across the two scenarios. In the first scenario, we fine-tune a network trained on English to perform inference in another language, covering the *Cross-Lingual Transfer* case; in the second scenario, we translate the inference pairs to English in order to perform monolingual inference, offloading the cross-lingual component of the task to a machine-translation task. We also take a deeper look into our results by evaluating the best model's performance across different inference labels (contradiction, entailment, and neutral).

3. **Purely Cross-Lingual Inference:** We evaluate our models on the cross-lingual inference formulation where the *premise* is in English and the *hypothesis* is in the target language. Similarly to the previous task, we also evaluate the best model's performance across different inference labels.

For the purposes of our evaluation, we consider a subset of languages provided by the XNLI data set—English, French, Spanish, German, Greek, Bulgarian and Russian, which provides a mix of Romance, Germanic and Indo-European languages.

4.1 | Aligned Word Embeddings

As described in Chapter 3, we train word embeddings for each language using the fastText framework (Joulin et al., 2016a). Our word embeddings are generated using an unsupervised Skip Gram Negative Sampling approach in 300 dimensions. The model is fed data from the Wikipedia (Wikipedia, 2004) and Europarl (Koehn, 2005) corpora for each language, where available. Subsequently, we align our word embeddings using the MUSE framework (Conneau et al., 2017b), and a modified version employing orthogonal constraints as described by Joulin et al. (2018). Within all our experiments, we select the English embedding as the source language for the alignment, mapping other target languages to align words, finally producing a set of aligned embeddings for both the source and the target language. We perform an intrinsic evaluation for our aligned embeddings, assessing the quantifiable accuracy of our alignment (Mikolov et al., 2013a; Søgaard et al., 2018; Vulić and Moens, 2013).

Our evaluation metric follows a standard benchmark for machine translation. We source English dictionary translations for each language from the MUSE framework (Conneau et al., 2017b). For each translation pair, we retrieve the top ten nearest neighbours from the aligned embedding using Euclidean distance. Accuracy is scored by evaluating whether the selected nearest neighbours contain the direct translation as provided by the dictionary; calculated as the fraction of entries found in the dictionary whose translation is found in the top k retrieved words (where k=10). Thus, we report our results as the percentage of words whose translation was correctly retrieved by the nearest-neighbour retrieval:

$$Accuracy = \frac{count\_selected\_words}{count\_translations\_top\_k\_selection} \tag{4.1}$$

Our results show that the orthogonal constraints applied during the training process increased the accuracy for translation retrieval. These results corroborate previous research on the topic Joulin et al. (2018), achieving comparable results across different language pairs.

Although our chosen approach aligns two separate word embeddings, resulting in two separate but aligned vector spaces, we also experiment with different approaches. For example, we merge the aligned embeddings into a single vector space, creating

	Europarl Sentences	Wikipedia Articles
en	2,218,201	6,181,238
fr	2,190,579	2,260,765
es	2,123,835	1,636,226
de	2,176,537	2,494,151
el	1,517,141	183,044
bg	411,636	266,604
ru	-	1,671,334

Table 4.1: Sizes of the Europarl (Koehn, 2005) and Wikipedia (Wikipedia contributors, 2004) corpora which were used for constructing Word Embeddings

Language	MUSE	Constrained
fr	78.3%	84.2%
es	76.2%	86.5%
de	68.9%	75.8%
el	53.1%	64.2%
bg	53.3%	64.8%
ru	58.2%	70.3%

Table 4.2: Accuracy for the nearest neighbour retrieval task for each alignment method, where MUSE refers to embeddings aligned using the MUSE framework and constrained refers to the same approach with orthogonal constraints applied as described in Section 3.2.2. Our results show that orthogonal constraints have a positive effect on word embeddings.

a single-word embeddings space composed of aligned vectors. Given two aligned vectors for a given language pair, such as English and French, we simply append the vectors into a single file prior to evaluation. This results in a single vector space which would supposedly be used to encode sentences in different languages. Similarly, we also experiment with learning word embeddings over a mixed language corpus to create a single multilingual vector space without the need of alignment. However, both these approaches yield poor results when applied to our neural networks. For this reason, we will opt to employ our word embeddings separately within our neural networks.

4.2 | Cross-Lingual Natural Language Inference

We present a number of different experiments in our exploration of cross-lingual inference. As discussed in Chapter 3, we train a neural network on entailment pairs for *English* and fine-tune the network to perform entailment in a different languages. All experiments employ the MultiNLI (Williams et al., 2018) data set for initial training in English. Meanwhile, we experiment with different cross-lingual entailment scenarios

during our fine-tuning phase: *Cross-Lingual Transfer*, *Cross-Lingual Transfer relying on Translation* (Translate-Train) and *Purely Cross-Lingual Inference*.

For each of our experiments, we test two neural architectures; a stacked Bidirectional LSTM (bilstm) network and a Decomposable attention (datten) network as discussed in Chapter 3. During the tuning process, we adopt a learning rate schedule wherein if the validation loss does not decrease within three epochs (the patience parameter), the learning rate is reduced. We perform a simple grid search over the selected optimizer, dropout ratio, epochs, learning rate and patience to select the ideal hyper-parameters for each approach. While we do not attempt to exhaustively evaluate this process, we provide an overview of the selected hyper-parameters in Table 4.3. Both approaches were tested using 1000 epochs and a batch size of 128. We observed that the algorithms typically reached a maxima within this training period.

	Batch Size	Hidden Layer Size	Dropout Ratio	Epochs	Learning Rate	Patience	Optimizer
bilstm	128	200	0.5	1000	0.01	3	adam
datten	128	200	0.3	1000	0.1	3	adagrad

Table 4.3: The hyper-parameters selected for each architecture.

We find that the Bidirectional LSTM model performs best using an Adam optimizer and a learning rate of 0.01. Meanwhile, the decomposable attention model performs best using an Adagrad optimizer and a larger learning rate of 0.1. Both algorithms employ three ReLU dropout layers, which achieves best performance with 30% and 50% dropout respectively.

Prior to testing our methods on different cross-lingual inference tasks, we carried out our initial experiments on the *monolingual* inference task presented by the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). The monolingual inference models trained during this phase of our evaluation are later fine-tuned to address the cross-lingual inference task. Table 4.4 shows a summary of our results on the monolingual inference task, compared to other approaches. The monolingual task is outside of the scope of this research, but the task serves as an important reference point to validate our approaches. The current state of the art within the monolingual task is achieved by Transformer Models, a different architecture to that which we employ. Thus, we seek to compare our methods to similar approaches that employ RNN and CNN models. When compared, the results for our approaches achieve similar performance, thus validating our implementation.

Dataset	Approach	Accuracy
MultiNLI	bilstm	**71.7%**
	datten	70.5%
	Transformer Model (SOTA) (Radford et al., 2018a)	**81.4%**
	CNN + Max-Pooling (Chen et al., 2017b)	74.9%
SNLI	bilstm	**84.7%**
	datten	69.2%
	Transformer Model (SOTA) (Pilault et al., 2020)	**92.1%**
	600D BiLSTM Encoder (Liu et al., 2016b)	83.2%

Table 4.4: Results for our approaches to monolingual inference tasks in English versus previous approaches. Although the current state of the art methods employ transformer models, the results for our approach are comparable to previous approaches using similar architectures, thus validating our approach.

4.2.1 | Cross-Lingual Transfer

In our first two experiments, we explore cross-lingual transfer learning for monolingual inference. Given a neural network trained for inference in English, we fine-tune the network to learn inference for text fragments in a different language.

We study the *Cross-Lingual transfer* scenario of cross-lingual inference, where we assess whether a proposed neural architecture trained in one language can be fine-tuned to perform inference classification in another language. We compare this approach against the alternative of relying on machine translation, where the neural network is fine-tuned for inference pairs which are translated from the target language to English. In the *Cross-Lingual transfer* case, for example, we train a neural network on entailment pairs for *English* and fine-tune the network using entailment pairs in *Spanish*. Within the *Translate Train* case, we train a neural network on entailment pairs for *English* and transfer the network to classify *Spanish* entailment pairs translated to *English*. This effectively reduces the cross-lingual transfer task to a monolingual task, where translation acts as an intermediary. Previous work has shown that this method yields the best results, providing a strong baseline upon which it is difficult to improve (Conneau et al., 2018; Hu et al., 2020). The experiments are carried out using the XNLI (Conneau et al., 2018) and an additionally constructed *TRAN-XNLI* data set as described in Chapter 3. We compare our results with benchmarks from approaches presented alongside the *XNLI* (Conneau et al., 2018) corpus; the authors propose a bidirectional LSTM architecture which, in contrast to our approach, employs sentence embeddings as opposed to word embeddings. The sentence embeddings employed in the baseline approach are of significantly higher dimensionality: whereas our approach employs 300 dimension word embeddings, sentence embeddings are of 1000 dimensions.

Table 4.5 shows the results for our experiments in cross lingual transfer. There is a sizable difference in performance for the two approaches - the Bidirectional LSTM (*bil-*

stm) architecture consistently outperforms our Decomposable Attention (*datten*). Another observation is that our method, which employs word embeddings, yields poorer results than the baselines, which employ sentence embeddings. The loss of accuracy is significant, averaging a 10% loss in the cross-lingual transfer case, and a 7.5% loss in the *Translate Train* case. In some cases, the loss is as little as 5%.

While the loss described above is significant, it should also be assessed by considering the lower dimensionality of the word embeddings, and the impact this has on computational resources, of which our training process requires significantly less. In quantifiable terms, models employing sentence embeddings take at least 30 hours of training time (Schwenk and Schwenk, 2020), processing 2,000 sentences per second, whereas the approaches employed within our solution are trained in under 3 hours. Thus, we present a trade-off which results in models being cheaper to train, accepting a degree of loss of accuracy in favour of cheaper training. This provides an interesting direction for future work, where more optimal solutions employing word embeddings may approach the results achieved by sentence embeddings.

Cross-Lingual Transfer Approaches			**Languages**						
Task	**Dataset**	**Model**	**en**	**fr**	**es**	**de**	**el**	**bg**	**ru**
Cross-Lingual Transfer	xnli	bilstm	**64.3%**	**58.9%**	**59.9%**	**58.7%**	**59.6%**	**52.3%**	**53.4%**
		datten	52.6%	50.5%	51.0%	52.9%	53.6%	48.3%	49.8%
XNLI Baseline (Conneau et al., 2018)			73.7%	67.7%	68.7%	67.7%	68.9%	67.9%	60.6%
Translate Train	tranxnli	bilstm	-	**62.0%**	**62.5%**	**61.3%**	**63.2%**	**61.7%**	**60.6%**
		datten	-	48.8%	48.7%	50.5%	50.6%	48.3%	48.0%
XNLI Baseline (Conneau et al., 2018)			-	70.4%	70.7%	68.7%	68.1%	70.4%	67.8%

Table 4.5: Results for our experiments in cross-lingual transfer for the inference task. We present the results for two architectures, a bidirectional LSTM (bilstm) and decomposable attention (datten), which we compare to baseline Bidirectional LSTM approaches presented alongside the XNLI corpus. In contrast to our approach, the baseline architectures employ sentence embeddings as opposed to word embeddings.

When considering the cross-linguistic performance of our models, both models perform consistently with a 5% to 8% variation, although a number of key observations can be made. Firstly, results for languages which come from the same language family are similar. *French* and *Spanish*, languages from the *Romance* branch of languages, consistently achieve high results in our experiments; *German* and *Greek*, solely representing *Germanic* and *Hellenic* languages respectively, achieve similar results. Meanwhile, the pair of languages achieving the poorest performance, *Russian* and *Bulgarian*, achieve similar results, perhaps due to the fact that they both use the Cyrillic alphabet. We also observe that, within the case of cross-lingual transfer, the variation is higher, with the model performing worse on languages which are more distant to English. Slavic languages such as *Bulgarian* and *Russian*, where we achieve the worst results, differ signif-

icantly more from *English* when compared to *French* and *Spanish*, where we achieve the best results. In fact, this variation disappears within the *Translate Train* scenario where the text is translated to English; there is no observable variation across languages in this case.

The results also show that the *Translate Train* approach consistently outperforms the *Cross-Lingual Transfer* approach, corroborating previous research (Conneau et al., 2018). Complexity introduced due to the cross-lingual nature of the fine-tuning process is offset by the high accuracy of current machine-translation models. This posits the idea that massively-multilingual inference could be achieved by well-researched machine-translation approaches; translating a sentence from a target language (for example *Spanish*) to *English* yields better accuracy than performing inference in the target language. Furthermore, the difference in performance between the *Cross-Lingual Transfer* and *Translate Train* approaches, within our implementation as well as in the baselines which use sentence embeddings, is similar. Thus, we conclude that offloading the complexity of cross-lingual inference to machine translation remains the most promising approach to the task, even when employing word embeddings.

Next, we take a deep dive into the results by analyzing the performance of the models across different labels. A brief analysis of results for our Decomposable Attention (*datten*) model shows that the model fails to classify inference adequately for all tasks. Results for the *contradiction* and *entailment* classes are particularly low, ranging from 0% to 30% and 10% to 30% respectively. Conversely, the approach achieves a very high accuracy for the *neutral* class, ranging between 85% and 90%. For this reason, we exclusively scope our analysis to the bidirectional LSTM (bilstm) approach. Table 4.6 shows the performance of the Bidirectional LSTM model for different labels for the cross-lingual transfer task.

Task	Language	Contradiction	Entailment	Neutral
XNLI	en	65.3%	72.2%	55.7%
	fr	54.3%	73.9%	46%
	es	59.2%	72.9%	47.4%
	de	58%	66.2%	51.1%
	el	64%	62.4%	50.7%
	bg	55.7%	26.6%	72.2%
	ru	57.2%	23.8%	73.1%
TRANXNLI	fr	70.2%	74.7%	47%
	es	67.4%	68.4%	52.2%
	de	65.2%	67.8%	50.8%
	el	68%	70.4%	51.3%
	bg	68.9%	72.8%	46.5%
	ru	66.2%	66.8%	48.7%

Table 4.6: Table of results for the bidirectional LSTM (bilstm), by language and target label.

Our results show that predicting the *Neutral* label is most challenging - the results for the *Neutral* label vary the most, with the models obtaining roughly 50% accuracy. We conclude that detecting entailment, whether positive (*Entailment*) or negative (*Contradiction*), is the most challenging part of the task, both for the monolingual translated and the cross-lingual scenarios.

However, this does not always hold true - our models yield anomalous results in the cases of *Bulgarian* and *Russian* within the cross-lingual transfer case, where we achieve high accuracy for the *Neutral* label, but low accuracy for the *Entailment* label. We put forward the idea that these results may be related to the nature of the languages - all languages within our set, with the exception of *Bulgarian* and *Russian* exhibit some strict form of word order. Briefly, word order relates to how strict rules are in terms of word placement. Such order dictates whether a sentence is a statement or a question. Given a sentence "*Nigel started a fire.*", alternative ordering of words such as "*A fire Nigel started*" may convey different intentions. Certain languages, such as *Bulgarian* and *Russian*, exhibit *free word order*, such that the order of words is not as semantically deterministic. This is in contrast to other languages with strict word order, such as *English*, *French*, *Italian*, *Greek*, and to a lesser extent, *German*. We suggest that the free word order characteristic for the Slavic languages of our language set may yield higher results for the *Neutral* label. In fact, within the *Translate Train* task, this phenomenon disappears due to inference being carried out in English. Perhaps unsurprisingly, results for the *translate train* scenario are very similar to those presented for *English* within the *cross-lingual transfer* scenario.

4.2.2 | Purely Cross-Lingual Inference

In our third experiment, *Purely Cross-Lingual Inference*, we train our neural network on *English* and fine-tune the network to classify inference pairs where one sentence is in *English* and the other sentence is in another language such as *Spanish*. The data set used in this experiment is the *CROSS-XNLI* data set, which is based on the original XNLI (Conneau et al., 2018) data set as discussed in Chapter 3. Currently there are no baselines addressing this scenario, thus our work in this area is novel.

Table 4.7 shows the results of our approach when applied to the purely cross-lingual inference task. Similar to the previous task, the bidirectional LSTM outperforms the decomposable attention model, albeit with a lower margin. Results are consistent across all languages with no major variations. Moreover, performance across different languages varies less than in the *cross-lingual transfer task* described in the previous section. Although the purely cross-lingual task is inherently more complex due to detecting in-

Cross-Lingual Transfer Approaches			Languages						
Task	**Dataset**	**Model**	**en**	**fr**	**es**	**de**	**el**	**bg**	**ru**
*Purely Cross-Lingual Inference	crossxnli	bilstm	**-**	**55.3%**	**54.0%**	**52.4%**	**50.6%**	**50.2%**	**53.1%**
		datten	-	48.2%	47.8%	52.3%	46.1%	46.1%	50.6%
Baseline (Artetxe and Schwenk, 2019a)			-	72.2%	72.2%	72.8%	71.6%	72.0%	71.4%

Table 4.7: Table showing the results achieved on the purely cross-lingual inference task, where the premise is in English and the Hypothesis is in another language. We compare our results to a similar architecture which employs sentence embeddings as opposed to word embeddings (Artetxe and Schwenk, 2019a).

ference across two languages, one of the sentences is in English. We put forward the idea that such consistency is due to the fact that the cross-lingual task includes English sentences, and thus the model requires less fine-tuning. When compared to the baseline, our approach using word embeddings as opposed to sentence embeddings performs significantly more poorly than our previous experiment in cross-lingual transfer; thus the two approaches are not comparable. We posit that the loss of accuracy in this case may be due to the fact that sentence embeddings provide a better representation when comparing sentences across languages.

Task	**Language**	**Contradiction**	**Entailment**	**Neutral**
CROSSXNLI	en	64.1%	69.2%	51.3%
	fr	59.9%	63.7%	41.9%
	es	53%	48.6%	59.7%
	de	60.3%	52.2%	44.8%
	el	52%	63.2%	33%
	bg	56.9%	58.9%	30.1%
	ru	57.5%	63.2%	38.3%

Table 4.8: Table of results for the bidirectional LSTM for the purely cross-lingual inference task.

Similar to our analysis in the previous section, we analyze the results for our best performing architecture in more detail, by breaking down the accuracy by target label. Table 4.8 shows that the Neutral label is the most challenging label to predict - with few cases surpassing 50% accuracy. This corroborates the idea presented in the previous section, that detecting entailment is more challenging than detecting the nature of the entailment. We observe that poorer prediction for the *Neutral* class is responsible the loss of accuracy in our two poorest models - *Greek* and *Bulgarian*. However, interestingly our model for *Russian*, which hails from the same language family as *Bulgarian*, performs well amongst the other languages in the set. Thus we find no conclusive evidence that linguistic attributes may influence performance in this task.

4.3 | Summary

In this section, we presented the results for our approach for aligning word embeddings and cross-lingual inference, noting the following observations:

- Incorporating translation as an initial pre-processing to deal with cross-lingual inference remains a strong approach which is difficult to surpass.
- Word embeddings do not perform as well as sentence embeddings for the inference task. However the difference in performance is not that great, with a 5% to 10% percentage loss of performance. This presents a trade-off which can be made in favour of a more computationally efficient training process.
- Perhaps unsurprisingly, the *purely cross-lingual inference* task which is unique to our research is significantly more complex than the *transfer* task.
- Our implementation of *decomposable attention* did not sufficiently detect entailment in text.
- Languages hailing from the same language families tend to render the same results within our approach.

5

Conclusions

Our research proposes the exploration of cross-lingual natural language inference, primarily motivated by the lack of exploration in the area, and with the broader goal of achieving more maintainable and inclusive natural language processing. We select the inference task specifically, due to its nature as a manifestation of semantic competence (Cooper et al., 1996). To facilitate our exploration, we investigate two lines of work which provide different formulations of cross-lingual inference with the aim to investigate:

- The use of transfer learning to improve the inference task across different languages; this is done by taking a model trained for NLI within a high-resource language (English), aiming to improve inference within other languages. Within our experiments, we describe this as the *Cross-Lingual Transfer* task, where a model is trained for the English inference task, and fine-tuned to carry out inference on another (target) language. Additionally, we explore an approach which relies on machine translation, the *Translate Test* scenario, where examples within the target language are translated to English. The latter scenario offloads the complexity of the cross-lingual task to machine translation.

- The application of deep learning approaches for the cross-lingual NLI formulation where the premise and the hypothesis are in different languages. We call this scenario the *Purely Cross-Lingual Inference* task. Our choice to address this task provides novelty due to the lack of research in the area.

5.1 | Achieved Aims and Objectives

The above goals are reflected within the approach described in Chapter 3 which achieve two aims; the construction of a set of aligned word embeddings for different languages, to be employed within our neural network architectures; and the design and adaptation of neural architectures for the scenarios described previously. Both these aims are achieved within our implementation. To adequately represent cross-lingual inference, we selected a subset of languages which provide a strong mix of Latin, Germanic and Indo-European languages - English, French, Spanish, German, Greek, Bulgarian and Russian. Our phased approach, considering input word embeddings and neural architectures separately, provides for an evaluation at different levels of the architecture.

In Chapter 4, we present the results for our approach; the solution implemented and subsequent results obtained address the objectives set out in Chapter 1:

1. To build a number of word embedding models to cater for cross-lingual inference; particularly a set of aligned word embeddings for cross-lingual NLP. Our aligned word embeddings achieve similar accuracy to the current state of the art for the nearest neighbour translation tasks, validating our choice of approach. Subsequently, we validate the application of the embeddings as part of our overall solution.

2. To design and adapt neural architectures for transfer learning within the inference task. We experiment with two different architectures adapted to the task, addressing *Cross-Lingual Transfer* and translation-based approaches for the same task. Within the *Cross-Lingual Transfer* case, our results corroborate previous findings by Conneau et al. (2018) regarding the effectiveness of employing machine translation as a part of the method. However in contrast to previous approaches, our work is different insofar that it exclusively employs word embeddings as opposed to sentence embeddings - this provides for a more computationally efficient training process due to the lower dimensionality of word embeddings. While not negligible, the lower accuracy achieved by our approach should be considered against the gains made in computational efficiency. With increased training data during training time, our approach may prove to provide a beneficial trade-off.

3. To design and adapt neural architectures for transfer learning within the purely cross-lingual inference task, where the *premise* is in English and the *hypothesis* is in another language, as initially proposed by Mehdad et al. (2010). The results achieved in this task are notably lower than performance for the *Cross-Lingual*

Transfer tasks. We attribute this to the increased complexity of detecting inference across different languages.

We perform a thorough evaluation of our solution in Chapter 4, providing suggestions for future work in this Chapter. Thus we conclude that our work has achieved the goals set out within our introduction, with our main contributions being a more efficient approach to cross-lingual transfer providing comparable results; and a benchmark for purely cross-lingual inference. More broadly, our work provides a reference point for future work in cross-lingual inference, which remains a nascent research area.

Our findings clearly indicate that employing translation within cross-lingual inference remains a strong baseline which is difficult to surpass; offloading the complexity of a cross-lingual task to be handled by established and well-researched machine translation algorithms achieves higher accuracy than attempting fine-tuning. This gives rise to the question of whether it is indeed worth pursuing cross-lingual inference as a task. However, there are several arguments in favour of pursuing cross-lingual NLI nevertheless.

Chiefly, one must consider the methods through which translation systems are evaluated. When evaluating machine translation systems extrinsically, we are required to deduce the relationship between two text fragments in different languages, in order to evaluate whether the translation is correct. This deduction involves classifying the inferential relationship between the sentences; in essence, the cross-lingual entailment classification can be employed to evaluate translation systems. Work within the fields of machine translation and cross-lingual inference are symbiotically beneficial. In turn, work in the area of cross-lingual inference will assist in improving translation systems; and subsequent improvements in translation systems can be used to further improve cross-lingual inference task. Moreover, one must also consider the broader aims of AI and natural language understanding. Considering AI to be the successful replication of human learning, one must consider how humans acquire language. Bilingual and multilingual speakers do not acquire languages in isolation; on the contrary, humans acquire languages by drawing experiences from other learned languages. Thus from an abstract standpoint, pursuing cross-lingual inference remains a worthy goal in the hope of replicating human language acquisition within a computational context. We therefore conclude that not only does cross-lingual inference remain a worthy task to pursue, but it is also necessary for further improving machine translation tasks.

5.2 | Limitations and Future Work

In Chapter 4, we presented the results for our approach. Although our research provides a number of unique contributions to the area of cross-lingual inference, our methods present a number of areas for improvement. Firstly, we observed that one of our architectures, the Decomposable Attention architecture, did not learn inference sufficiently - we suggest that additional hyper-parameter tuning is required to exhaustively to rule out this approach. Moreover, our approach exclusively employs word embeddings as opposed to sentence embeddings - we propose that a combination of sentence embeddings and word embeddings, or alternatively a more complex encoding architecture, may provide a solution which achieves higher accuracy while striving to keep training time low.

As discussed within our introduction, research within the area of cross-lingual NLP, and more so within cross-lingual inference, is sparse. Task formulations within the area are not well defined - with the phrase 'cross-lingual' holding different meanings depending on the research considered. Recent research has nearly exclusively concentrated on cross-lingual transfer (Conneau et al., 2018). Thus our research casts a wide net in its exploration of cross-lingual inference with the aim of providing a reference point for future work. The three scenarios addressed, two within the 'cross-lingual transfer', and one which we term 'purely cross-lingual inference', cover a broad range of computational tasks. More work is required within the area of cross-lingual semantics to establish the road-map towards massively multilingual NLP, particularly in the form of better defined and understood task definitions. Moreover, each different formulation would benefit from dedicated research.

Our work is also motivated by improving NLP performance for low-resource languages. Within our experiments we select a few languages to sufficiently test our approach. Resources for the selected languages are available to a lesser extent than English, however other languages with fewer digitised resources exist, such as Swahili and Urdu. Applying our approaches to such languages presents an avenue for future work, possibly adopting variations to our approach, particularly in the area of training word embeddings.

Within our study, we choose to exclusively focus on employing traditional neural architectures. This is primarily due to the fact that in spite of recent advances, the application of neural networks within cross-lingual natural language processing remains an under-explored area. However as discussed previously, recent work within the area of natural language processing has suggested novel architectures which exclusively employ attention, Transformer models (Devlin et al., 2019; Hu et al., 2020). Such models

have achieved the a new state of the art on a variety of tasks, including inference, however the ability of such models to truly capture inference has been called into question (Kovaleva et al., 2019). At the time of writing, this area of research is rapidly evolving and thus provides an avenue for future work on the topic.

5.3 | Final Remarks

Cross-lingual NLP is a nascent area within natural language processing with the vast majority of natural language processing research has revolving around English, There are several reasons as to why achieving multilingual NLP is attractive, chief among which is fostering technological inclusion within communities which do not speak English.

The nature of cross-lingual intelligence adds an additional layer of complexity to NLP tasks, challenging previous approaches to be applied to address different linguistic phenomena. Within our research, we opt to address inference, which is arguably one of the more complex tasks within the area due to its inherently nuanced nature. Testament to this is the research around constructing inference data sets which represent a broad range of inferential reasoning. Moreover, in this chapter we also explain how the inference task itself can be used to improve machine translation, which in turn can also play an important role in improving cross-lingual inference. Further research in the area will result in more robust solutions, benefiting both the advancement of NLP in languages other than English, and NLP in English itself. Amongst these tasks, inference is particularly complex due to the nature of linguistic phenomena.

Although our methods do not provide improvements in accuracy, our research makes two unique contributions to further research - to our knowledge no approach has been presented using word embeddings, with previous baselines exclusively using sentence embeddings; moreover, we address the purely cross-lingual inference task using neural networks, a task for which no previous work is available. Thus we believe that our exploration offers a strong contribution in pushing for continuous development of cross-lingual inference approaches.

In conclusion, our work makes a number of contributions to the field of Artificial Intelligence; it provides an investigation into cross-lingual transfer learning for the inference task; and the application of deep learning approaches within such tasks. These research contributions help foster technological inclusion, ensuring that the field continues to progress NLP for different language; while also aiming to inspire systems which closely mimic human natural language acquisition.

References

Heike Adel and Hinrich Schütze. Exploring different dimensions of attention for uncertainty detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 22–34, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1003.

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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 3874–3884. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1388. URL https://doi.org/10.18653/v1/n19-1388.

Amazon. Amazon web services, 2006. URL https://aws.amazon.com/. [Online; accessed 22-July-2020].

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, March 2019a. doi: 10.1162/tacl_a_00288.

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, 2019b.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL https://www.aclweb.org/anthology/P18-1073.

Yusuf Aytar and Andrew Zisserman. Tabula rasa: Model transfer for object category detection. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool, editors, *IEEE International Conference*

on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, pages 2252–2259. IEEE Computer Society, 2011. doi: 10.1109/ICCV.2011.6126504. URL https://doi.org/10.1109/ICCV.2011.6126504.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

Roy Bar-Heim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognizing textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy*, page 49, 2006.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics.

Emily Bender. The #benderrule: On naming the languages we study and why it matters. *The Gradient*, 2019.

Emily M Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26, 2011.

Emily M. Bender and Alexander Koller. Climbing towards NLU: on meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5185–5198. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.463. URL https://doi.org/10.18653/v1/2020.acl-main.463.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. NIST, 2009. URL https://tac.nist.gov/publications/2009/additional.papers/RTE5_overview.proceedings.pdf.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The sixth PASCAL recognizing textual entailment challenge. In *Proceedings of the Third Text Analysis Conference, TAC 2010, Gaithersburg, Maryland, USA, November 15-16, 2010*. NIST, 2010. URL https://tac.nist.gov/publications/2010/additional.papers/RTE6_overview.proceedings.pdf.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*. NIST, 2011. URL https://tac.nist.gov/publications/2011/additional.papers/RTE7_overview.proceedings.pdf.

Shane Bergsma and Benjamin Van Durme. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1764, Barcelona, Catalonia, Spain, 2011. Citeseer, AAAI Press.

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 (EMNLP)*. Association for Computational Linguistics, 2015.

Iacer Calixto, Qun Liu, and Nick Campbell. Multilingual multi-modal embeddings for natural language processing. *arXiv preprint arXiv:1702.01101*, 2017.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1657–1668. Association for Computational Linguistics, 2017a. doi: 10.18653/v1/P17-1152. URL https://doi.org/10.18653/v1/P17-1152.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 36–40, Copenhagen, Denmark, September 2017b. Association for Computational Linguistics. doi: 10.18653/v1/W17-5307. URL https://www.aclweb.org/anthology/W17-5307.

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.

Stéphane Clinchant, Cyril Goutte, and Eric Gaussier. Lexical entailment for information retrieval. In *European Conference on Information Retrieval*, pages 217–228, Imperial College, London, England, 2006. Springer, Springer-Verlag Berlin Heidelberg.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, Helsinki, Finland, 2008. Association for Computing Machinery.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

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 2017a. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL https://www.aclweb.org/anthology/D17-1070.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017b.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1269. URL `https://doi.org/10.18653/v1/d18-1269`.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium, 1996.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc, editors, *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer, 2005. doi: 10.1007/11736790_9. URL `https://doi.org/10.1007/11736790_9`.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

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`.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1285–1295. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1136. URL `https://doi.org/10.18653/v1/d16-1136`.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Multilingual training of crosslingual word embeddings. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 894–904. Association for Computational Linguistics, 2017. doi: 10.18653/v1/e17-1084. URL `https://doi.org/10.18653/v1/e17-1084`.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N13-1073`.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*, 2018.

Jerry A Fodor. *The language of thought*, volume 5. Harvard university press, 1975.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1303. URL https://www.aclweb.org/anthology/D17-1303.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

Reza Ghaeini, Sadid A. Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Fern, and Oladimeji Farri. DR-BiLSTM: Dependent reading bidirectional LSTM for natural language inference. 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 1460–1469, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1132. URL https://www.aclweb.org/anthology/N18-1132.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In Satoshi Sekine, Kentaro Inui, Ido Dagan, Bill Dolan, Danilo Giampiccolo, and Bernardo Magnini, editors, *Proceedings of the ACL-PASCAL@ACL 2007 Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republic, June 28-29, 2007*, pages 1–9. Association for Computational Linguistics, 2007. URL https://www.aclweb.org/anthology/W07-1401/.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. The fourth PASCAL recognizing textual entailment challenge. In *Proceedings of the First Text Analysis Conference, TAC 2008, Gaithersburg, Maryland, USA, November 17-19, 2008*. NIST, 2008. URL https://tac.nist.gov/publications/2008/additional.papers/RTE-4_overview.proceedings.pdf.

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2103. URL https://www.aclweb.org/anthology/P18-2103.

Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722, 2014.

Google. Google translate. http://translate.google.com, 2006. [Online; accessed 22-September-2020].

Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1157. URL https://www.aclweb.org/anthology/N15-1157.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. 2015.

Jerzy W Grzymala-Busse. Fahiem bacchus. representing and reasoning with probabilistic knowledge. a logical approach to probabilities. the mit press, cambridge, mass., and london, 1991 (© 1990), xvii+ 233 pp. *The Journal of Symbolic Logic*, 64(4):1837–1837, 1999.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P08-1088`.

Joseph Y Halpern. An analysis of first-order logics of probability. *Artificial intelligence*, 46(3):311–350, 1990.

Sanda Harabagiu and Andrew Hickl. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia, 2006. Association for Computational Linguistics.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*, 2013.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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.

Kejun Huang, Matt Gardner, Evangelos Papalexakis, Christos Faloutsos, Nikos Sidiropoulos, Tom Mitchell, Partha P. Talukdar, and Xiao Fu. Translation invariant word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1084–1088, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1127. URL `https://www.aclweb.org/anthology/D15-1127`.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL `https://www.aclweb.org/anthology/N19-1357`.

Sergio Jiménez, Claudia Jeanneth Becerra, and Alexander Gelbukh. Softcardinality: Learning to identify directional cross-lingual entailment from cardinalities and smt. In *SemEval@NAACL-HLT*, Atlanta, Georgia, USA, 2013.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016a.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759, 2016b.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1330. URL `https://www.aclweb.org/anthology/D18-1330`.

R. Kenney and P. Smith. *Vagueness: A Reader*. A Bradford book. MIT Press, 1996. ISBN 9780262112253. URL `https://books.google.com.mt/books?id=-iWiQgAACAAJ`.

Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *AAAI*, volume 17, pages 41–42, 2018.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL `https://www.aclweb.org/anthology/C12-1089`.

Tomáš Kočiskỳ, Karl Moritz Hermann, and Phil Blunsom. Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947*, 2014.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, Phuket, Thailand, 2005. Citeseer, Association for Machine Translation.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4364–4373. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1445. URL `https://doi.org/10.18653/v1/D19-1445`.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=H196sainb`.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. Learning multilingual word representations using a bag-of-words autoencoder. *CoRR*, abs/1401.1803, 2014.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1027. URL `https://www.aclweb.org/anthology/P15-1027`.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, Rome, Italy, 2012. Citeseer, AAAI Press.

Omer Levy, Anders Søgaard, and Yoav Goldberg. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 765–774. Association for Computational Linguistics, 2017. doi: 10.18653/v1/e17-1072. URL `https://doi.org/10.18653/v1/e17-1072`.

Hongmin Li, Xukun Li, Doina Caragea, and Cornelia Caragea. Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks. *Proc ISCRAM Asian Pacific*, pages 1–13, 2018.

Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090, 2016a.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090, 2016b.

Elena Lloret, Oscar Ferrández, Rafael Munoz, and Manuel Palomar. A text summarization approach under the influence of textual entailment. In *NLPCS*, pages 22–31, 2008.

Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *International Conference on Learning Representations*, 2016.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland, August 2014a. Association for Computational Linguistics. doi: 10.3115/v1/S14-2001. URL `https://www.aclweb.org/anthology/S14-2001`.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223, 2014b.

Yashar Mehdad, Matteo Negri, and Marcello Federico. Towards cross-lingual textual entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324, Los Angeles, California, 2010.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. volume 2, pages 1045–1048, Chiba, Japan, 01 2010.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013a. URL `http://arxiv.org/abs/1301.3781`.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, Lake Tahoe, USA, 2013b.

Tomáš Mikolov et al. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80:26, 2012.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. How transferable are neural networks in NLP applications? In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 479–489. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1046. URL `https://doi.org/10.18653/v1/d16-1046`.

Tsendsuren Munkhdalai and Hong Yu. Neural semantic encoders. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 397, Vancouver, Canada, 2017. NIH Public Access, Association for Computational Linguistics.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. The RepEval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5301. URL `https://www.aclweb.org/anthology/W17-5301`.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 25–33, Atlanta, Georgia, USA, 2013.

Yixin Nie and Mohit Bansal. Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5308. URL `https://www.aclweb.org/anthology/W17-5308`.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2249–2255. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1244. URL `https://doi.org/10.18653/v1/d16-1244`.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014. doi: 10.3115/v1/d14-1162. URL `https://doi.org/10.3115/v1/d14-1162`.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1202. URL `https://doi.org/10.18653/v1/n18-1202`.

Jonathan Pilault, Amine Elhattami, and Christopher J. Pal. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters & less data. *CoRR*, abs/2009.09139, 2020.

A Radford, K Narasimhan, T Salimans, and I Sutskever. Improving language understanding by generative pre-training. openai, 2018a.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018b.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010.

Georg Rehm. Europe's languages in the digital age: Multilingual technologies for overcoming language barriers and preventing digital language extinction. In *Workshop "State of the Art of Machine Translation–Current Challenges and Future Opportunities",(STOA)*, 2013.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. Reasoning about Entailment with Neural Attention. *arXiv e-prints*, art. arXiv:1509.06664, September 2015.

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. Investigating a generic paraphrase-based approach for relation extraction. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006. Association for Computational Linguistics.

Sebastian Ruder. Why You Should Do NLP Beyond English. `http://ruder.io/nlp-beyond-english`, 2020.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65:569–631, 2019. doi: 10.1613/jair.1.11640.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*, 2019.

Holger Schwenk. Continuous space language models. *Computer Speech & Language*, 21(3):492–518, 2007.

Holger Schwenk and Matthijs Douze. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2619. URL https://www.aclweb.org/anthology/W17-2619.

Holger Schwenk and Holger Schwenk. Laser natural language processing toolkit, Mar 2020. URL https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/.

John R. Searle. Minds, brains, and programs. In Margaret A. Boden, editor, *The Philosophy of Artificial Intelligence*, Oxford readings in philosophy, pages 67–88. Oxford University Press, 1990.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2931–2951. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1282. URL https://doi.org/10.18653/v1/p19-1282.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 567–572, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2093. URL https://www.aclweb.org/anthology/P15-2093.

Gary F Simons and Charles D Fennig. *Ethnologue: Languages of Asia*. sil International, 2017.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=r1Aab85gg.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL https://www.aclweb.org/anthology/P18-1072.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, Monteral, Canada, 2014.

Sebastian Thrun and Lorien Pratt, editors. *Learning to Learn*. Kluwer Academic Publishers, Norwell, MA, USA, 1998. ISBN 0-7923-8047-9.

Lisa Torrey and Jude Shavlik. Transfer learning. In *Transfer Learning*, 2009.

Marco Turchi and Matteo Negri. Altn: Word alignment features for cross-lingual textual entailment. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 128–132, Atlanta, Georgia, USA, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, California, USA, 2017.

Darnes Vilarino, David Pinto, Saúl León, Yuridiana Alemán, and Helena Gomez. Buap: N-gram based feature evaluation for the cross-lingual textual entailment task. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 124–127, Atlanta, Georgia, USA, 2013.

Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. Combining recurrent and convolutional neural networks for relation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 534–539, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1065. URL `https://www.aclweb.org/anthology/N16-1065`.

Ivan Vulić and Marie-Francine Moens. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N13-1011`.

Yogarshi Vyas and Marine Carpuat. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1187–1197, 2016.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Shuohang Wang and Jing Jiang. Learning natural language inference with LSTM. *CoRR*, abs/1512.08849, 2015.

Sida I Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Ying Wen, Weinan Zhang, Rui Luo, and Jun Wang. Learning text representation using recurrent convolutional neural network with highway layers. 06 2016.

Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, page 2764–2770. AAAI Press, 2011. ISBN 9781577355151.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019,*

Hong Kong, China, November 3-7, 2019, pages 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002. URL https://doi.org/10.18653/v1/D19-1002.

Wikipedia. Wikipedia data dump, 2004. URL https://en.wikipedia.org/wiki/Wikipedia:Database_download. [Online; accessed 22-July-2020].

Wikipedia contributors. Wikipedia, the free encyclopedia, 2004. URL https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350. [Online; accessed 22-July-2020].

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://www.aclweb.org/anthology/N18-1101.

Min Xiao and Yuhong Guo. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1613. URL https://www.aclweb.org/anthology/W14-1613.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, editors, *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1006–1011. The Association for Computational Linguistics, 2015. doi: 10.3115/v1/n15-1104. URL https://doi.org/10.3115/v1/n15-1104.

Lanqing Xue, Xiaopeng Li, and Nevin L Zhang. Not all attention is needed: Gated attention network for sequence data. In *AAAI*, pages 6550–6557, 2020.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of cnn and rnn for natural language processing. 02 2017.

Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi S. Jaakkola. Ten pairs to tag - multilingual POS tagging via coarse mapping between embeddings. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1307–1317. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-1156. URL https://doi.org/10.18653/v1/n16-1156.

Alisa Zhila, Wen-tau Yih, Christopher Meek, Geoffrey Zweig, and Tomáš Mikolov. Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1000–1009, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural*

Language Processing, pages 1393–1398, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1141.