ZERO SHOT DIALOGUE ACT CLASSIFICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

İLIM UĞUR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2019

Approval of the thesis:

ZERO SHOT DIALOGUE ACT CLASSIFICATION

submitted by **İLIM UĞUR** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Göktürk Üçoluk
Supervisor, **Computer Engineering Department, METU** _____

Assoc. Prof. Dr. Sinan Kalkan
Co-supervisor, **Computer Engineering Department, METU** _____

Examining Committee Members:

Assist. Prof. Dr. Emre Akbaş
Computer Engineering Department, METU _____

Prof. Dr. Göktürk Üçoluk
Computer Engineering Department, METU _____

Assist. Prof. Dr. Burcu Can
Computer Engineering Department, Hacettepe University _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: İlim Uğur

Signature :

ABSTRACT

ZERO SHOT DIALOGUE ACT CLASSIFICATION

Uğur, İlim
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Göktürk Üçoluk
Co-Supervisor: Assoc. Prof. Dr. Sinan Kalkan

September 2019, 81 pages

Solutions to many natural language processing problems need language-specific labeled data to be learned. However, both the endeavor of compiling a new dataset in a new language and the practice of translating an existing dataset to another language require human expert effort which can not be automated. To learn a solution in a new target language in an automated manner without any extra data, we focus on the known problem of dialogue act classification and propose two solutions that combine existing dialogue act classification methods with machine translation techniques. We implement the proposed solutions Localized Dialogue Act Classifier (LDAC) and Universal Dialogue Act Classifier (UDAC) using two different dialogue act classification methods, and a state-of-the-art machine translation method. We test both solutions on two datasets that are frequently used in testing a dialogue act classification method, namely Switchboard Dialogue Act (SwDA) and Meeting Recorder Dialogue Act (MRDA) datasets, and use German, Spanish and Turkish as the target languages. The results show that the models trained on translated datasets perform worse than their monolingual counterparts, trained on a dataset in its original lan-

guage. Nonetheless, the results also indicate that acceptably accurate dialogue act classification is achieved on new target languages by LDAC, without having a dataset in that language. These results show that the automated dataset translation idea we propose deserves further exploration.

Keywords: dialogue act classification, zero shot learning, word embeddings, machine translation

ÖZ

SIFIR ATIŞ DİYALOG SINIFLANDIRMASI

Uğur, İlim
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Göktürk Üçoluk
Ortak Tez Yöneticisi: Doç. Dr. Sinan Kalkan

Eylül 2019, 81 sayfa

Birçok doğal dil işleme sorununa yönelik çözümler, öğrenilmesi için dile özgü etiketlenmiş verilere ihtiyaç duymaktadır. Bununla birlikte, hem yeni bir dilde yeni bir veri seti derlemek hem de mevcut bir veri setini başka bir dile tercüme etmek çabası, otomatikleştirilemeyecek bir uzman insan katkısını gerektirmektedir. Bu tez kapsamında, yeni bir hedef dilde bir veri kümesi olmaksızın o dildeki bir doğal dil işleme probleminin çözümünü öğrenebilmek kabiliyetini elde etmeyi araştırmakta ve bilinen bir problem olan diyalog sınıflandırma problemine odaklanmaktayız. Bu kapsamda, mevcut diyalog sınıflandırma yöntemlerini makine çevirisi teknikleri ile birleştiren, Yerelleştirilmiş Diyalog Sınıflandırıcısı (YEDİS) ve Evrensel Diyalog Sınıflandırıcısı (EDİS) adında iki çözüm önermekteyiz. Önerdiğimiz çözümler iki farklı diyalog eylem sınıflandırma yöntemi ve son teknoloji ürünü bir makine çevirisi yöntemi kullanarak uygulanıyor. Çözümleri, diyalog sınıflandırma yönteminin test edilmesinde sıklıkla kullanılan iki veri seti (SwDA ve MRDA) üzerinde, Almanca, İspanyolca ve Türkçe hedef dilleriyle test ediyoruz. Sonuçlar, çevrilen veri setleri üzerinde eğitilen modellerin, tercüme edilmemiş bir veri kümesi üzerinde eğitilen tek dilli eşlerine kı-

yasla daha kötü performans gösterdiğini belirtiyor. Yine de, sonuçlar aynı zamanda LDAC tarafından yeni hedef dillerde, bu dilde bir veri setine sahip olmadan, kabul edilebilir bir doğruluk oranıyla diyalog sınıflaması yapılabildiğini göstermektedir. Bu sonuçlar, önerdiğimiz otomatik veri kümesi çeviri yaklaşımının daha fazla araştırmaya değer olduğunu gösteriyor.

Anahtar Kelimeler: diyalog sınıflandırma, sıfır-atış öğrenmesi, sözcük vektörleri, makine çevirmesi

To my family,

ACKNOWLEDGMENTS

I want to thank my supervisor Professor Göktürk Üçoluk and my co-supervisor Associate Professor Sinan Kalkan, for their constant support, guidance and, at times, patience. They were the mentors who pitched the first crumbs of what would become the main contribution presented in this text. I appreciate both of their invaluable help, without which it would be much less probable for me to complete this research.

I want to thank my parents Müyesser Yıldız and Naci Uğur for being my biggest supporters at all times, and whom I have neglected as I worked on this thesis.

Thanks a lot to Bahar Şevket, who read this document, suggested formatting improvements, and provided her support even when I was not aware that I needed it.

Thanks to Fatih Semiz for his help throughout my process of writing this thesis. His guidance as to the thesis preparation and presentation has been much helpful.

I appreciate the emotional and professional support I received from each past and present member of Vedat Minör, the band of which I am lucky to be a part. Particularly its members Damla Koçkar, Deniz Beyaz, Mine Hasırcı Doğan, Ozan Atak, and Serhat Bayılı constituted the entirety of my social life during this long and challenging period.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DA	Dialogue Act
DNN	Deep Neural Network
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
LDAC	Localized Dialogue Act Classifier
LSTM	Long Short Term Memory
MRDA	Meeting Recorder Dialogue Act
MT	Machine Translation
NLP	Natural Language Processing
NN	Neural Network
RNN	Recurrent Neural Network
SVM	Support Vector Machine
SwDA	Switchboard Dialogue Act
UDAC	Universal Dialogue Act Classifier

CHAPTER 1

INTRODUCTION

The work reported in this thesis focuses on combining the advances made in two separate, yet related problems in Natural Language Processing (NLP) in order to remedy a problem. This chapter starts with introducing those NLP tasks. The section following the overview of those tasks outlines the problem addressed with this research and the incentives in tackling it. We then list the main contributions of this work and finally outline the structure of the thesis.

1.1 Dialogue Act Classification

Dialogue Act (DA) Classification is an important and well-studied task in NLP. Broadly defined, DA Classification task is a fundamental classification problem in NLP, the goal of which is selecting a semantically accurate label for each utterance in a given dialogue, from a predetermined set of tags, with minimum error. Within the context of the problem, the term dialogue refers to a series of chronologically-sorted utterances, spoken by two or more parties. An utterance refers to the textual representation of verbal expression, uttered by one of the parties involved in the dialogue. An utterance may include words, numerals, and punctuations, based on the data format used. The predetermined set of tags, mentioned in the problem definition above, represents a taxonomy that was formed by experts, where each tag included within the set represents a different semantic connotation, intended to correspond to a class of utterances. The error rate, or prediction accuracy, is determined by the ratio of the instances correctly classified over the entire test cases available.

DA Classification problem has a wide area of application. In recent years, many sys-

tems have been developed to enhance human-computer interaction. The ability of any computational system to have a written or vocal conversation with a human naturally relies on the capability to interpret the utterances expressed by humans. After all, classifying utterances and differentiating between statements, different types of questions and responses are vital to any such computerized dialogue system. Hence, this capability is essential to have and is relied on heavily by dialogue systems used in various industries, including medical care and commercial marketing.

1.2 Machine Translation

Similar to DA Classification task, Machine Translation is another fundamental natural language processing task. The goal of MT task is to process a text of arbitrary length in a source language and to produce a semantic equivalent of that text in the desired target language. How it differs from traditional translation, where one or more human experts in relevant languages do the work, is the effort to minimize, if not remove, the human contribution. An MT method typically tries to utilize minimal or no expert human input in order to be able to complete its task. In practice, the goal of building an equivalent text in a target language has become a task of replacing the words and phrases given as input in the source language into their counterparts in the target language, in a manner that preserves as much of the semantics of the original text as possible.

Evaluating the performance of an MT method is not trivial. Due to the human effort being expensive and not reusable, many researchers such as Papineni et al. [5] studied methods of automating the evaluation process. A typical approach used in some of the MT research is providing a text that is a single word or phrase. In that case, the output is more likely to be a single word or phrase, as well. This testing methodology can not test the proficiency of the method in question with longer texts, such as sentences or paragraphs. However, it can be automated and reused, by comprising a list of words and phrases in the source language in advance, along with their counterparts in the target language.

Through recent technological advents, most noteworthy of which is the Internet, the

amount of machine-readable data available for processing increased exponentially. Another feature of such texts is that they originate from all around the world, and they are in various languages. Any effort to handle such textual data requires a capability to translate some portion of it into other languages. As a result, various parties such as data mining initiatives, academic research endeavors, and individual enthusiasts require an automated, reusable MT solution. It is essential to note that there is no known system in existence, which can output a perfect contextually and semantically equivalent translation of any text given as input. Still, many institutions and individuals, as mentioned above, lean on existing proprietary or open technologies for machine translation. State-of-the-art systems that are capable of providing an adequate output can be utilized to, at the very least, gain an insight as to the context of the text. Weighed against the time and financial cost of using a human expert translator to do the task, an imperfect output provided by the modern automated systems is deemed sufficient by many parties.

1.3 Motivation and Problem Definition

There is a common theme among the solutions to various NLP problems. Many of those solutions either work for a specific language, or they need to be learned from language-specific data. This deficiency requires constructing new solutions or compiling new datasets for each language in which a solution to an NLP problem is to be devised. Both courses of action require manual human effort. Even though the latter can be automated for unlabeled data, unfortunately, much of the data required by NLP tasks needs to be labeled. This lack of labeled data presents a challenge in training an existing NLP solution in a new target language.

One possible approach to address the lack of labeled data in a target language with minimal manual human effort is to translate a labeled dataset, already available in a particular source language to another target language, using automated Machine Translation (MT) methods. If the machine translation is sufficiently good at translating a dataset, existing solutions to various NLP problems can be learned to work in many new languages, without requiring any new datasets to be compiled.

In order to test the efficacy of this method, the research should focus on the approach, rather than devising a new NLP solution, or a translation method. Instead, we elect to work on a specific NLP problem named Dialogue Act (DA) classification. As part of that effort, in this thesis, we combine existing solutions to DA classification task with a state-of-the-art MT method, by leveraging the shared use of word embeddings in all the selected methods. We offer two different DA classification solutions named Localized Dialogue Act Classifier (LDAC) and Universal Dialogue Act Classifier (UDAC) which can remedy the lack of data when learning to classify dialogues in any target language. As eliminating expert human effort was one of the main motivations behind our approach, we also devise an automated methodology to test the automated solution we propose.

Using the testing methodology we propose, we test LDAC and UDAC with three different target languages and with two different DA classification solutions, on two frequently used DA classification datasets, named Switchboard Dialogue Act (SwDA) corpus and Meeting Recorder Dialogue Act (MRDA) corpus. We additionally explore how the word order in an utterance affects the performance of LDAC and UDAC.

Examining the results of the experiments conducted, we observe that the accuracies of LDAC and UDAC are not as high as the monolingual DA classification accuracy of the DA classifier they employ.

On the other hand, it is also striking that, at least for MRDA corpus, when LDAC uses a state-of-the-art DA classifier, the models trained with translated data perform better than a relatively new DA classification solution performs on the original dataset, without any translation. This fact, combined with the other results we present, indicate that the translation-based solutions proposed in this thesis deserve further exploration, and can eventually be adopted by many NLP solutions.

1.4 Contributions and Novelties

The contributions of this thesis are as follows:

- We offer two solutions to remedy the lack of labeled datasets in many languages

for various NLP problems, which is a substantial setback in implementing solutions to those problems.

- We focus on DA classification problem to investigate the efficacy of the solutions we propose. By running experiments spanning multiple different target languages and multiple datasets, we quantify the viability of the solutions we offer.
- The solutions we propose combine prominent previous works in DA classification and MT. Although their architectures initially seem simplistic, they utilize existing methods in a novel way.

1.5 Outline of the Thesis

The structure of the thesis is as follows. Chapter 2 provides the definitions of the problems this thesis is focused on, namely, Dialogue Act Classification and Machine Translation. It also details the previous research efforts made in solving each of those problems. Chapter 3 presents the proposed solution and describes the DA Classification methods and MT method to be used in detail. Chapter 4 describes the setup for the experiments used to test the proposed solutions and analyzes the results obtained from the experiments. Finally, Chapter 5 presents the conclusion and proposes future work to be done.

CHAPTER 2

RELATED WORK

This chapter covers each relevant task separately, mentioning the significant studies made on each problem. It is important to note that, despite the diversity of the earlier studies featured below, the more recent studies covered below are more focused on the research that is more relevant to the scope of this thesis, since the work conducted on both problems is rather extensive.

2.1 Dialogue Act Classification

DA Classification problem has been known and studied for more than two decades. Naturally, there have been various approaches with which researchers tackled this task.

Early concepts to solve this problem were based heavily on the ideas from NLP domain, such as usage of language models. Another popular strategy was using decision or classification trees. For instance, Mast et al. [6] tested two separate algorithms, where one used Semantic Classification Trees, and the other used Polygrams. Within their research, they introduced classification trees that were dependent on the state of the dialogues, as well as competing language models within their algorithm that utilized Polygrams. Similarly, around the same time, Warnke et al. [7] used semantic classification trees as well and trained their algorithm by searching for an optimal tree on Word Hypotheses Graphs, using A* algorithm, which was initially devised by Hart et al. [8]

After the initial heavy influence of the NLP methods, a variety of different approaches

emerged. Reithinger et al. [9] observed that, by 2000, DA classification research had produced three significant methodologies that can be used to develop a model.

The first method used by researchers was the utilization of statistical classifiers that used language models. Using that approach, Reithinger et al. [10] observed the problem of cumulative error in traditionally built systems that relied on syntactic and semantic features. Instead of a lexicographical system, they studied a statistical method that yields 65% and 74% accuracy for German and English test data, respectively. Similarly, a few years later, Choi et al. [11] improved the statistical approach with a model that uses maximum entropy to acquire probabilistic information from a tagged corpus. Their model achieved 83% accuracy. Though it is important to note that their results can not be compared directly with prior research, as the corpus they used was not the one used by Reithinger et al. [10].

The second research route taken was the usage of transformation-based learning. Observing how Brill [12] introduced transformation-based learning into part-of-speech tagging problem, and yielded best results on that problem known to date, Samuel et al. [13] applied the same approach to the somewhat similar problem of DA Classification.

The third path was using neural networks. Kipp [14] proposed an approach based on splitting the data and getting it processed by different portions of the network. This research is particularly significant, as it embodies many ideas included in the state-of-the-art learning-based models described below, such as vectorizing the input and using the hierarchy of the dialogue to split the input into smaller, more meaningful components.

In one of the exploratory research efforts in that period, Shriberg et al. [15] evaluated almost all of the techniques used at the time, including Hidden Markov Models, N-gram language models, maximum entropy estimation, decision tree classifiers, and neural networks. Their study was one of the first to adopt a dataset called Switchboard corpus (SwDA) by Godfrey et al. [16], which later became one of the standard datasets used to test any DA classification solution, including the one presented in this thesis. Even as early as 1998, Shriberg et al. achieved a maximum accuracy of 72% in classifying dialogue acts. Considering that inter-personal agreement in utterance

tagging among the human expert assessors of SwDA dataset regarding the dialogue act tags is 83%, the result they obtained is quite significant.

Ang et al. [17] studied the dialogue act classification on a new dataset presented by Shriberg et al. [18]. The study uses a maximum entropy classifier, similar to some of the other studies highlighted above. Still, this research is significant to highlight as it is one of the first studies that first utilized another corpus that, after SwDA, is widely used in testing DA classification methods, called Meeting Recorder Dialogue Act (MRDA), dialogues in which have more than two parties. It also initiates specific heuristics as to the usage of MRDA in DA Classification problem that are used to this day, such as reducing the more complex labeling system used in MRDA down to five major labels.

Lafferty et al. [19] introduced Conditional Random Fields (CRF), a framework with which to build probabilistic models that can classify sequential data. Using a CRF for DA classification task has certain advantages. Using generative models like Hidden Markov models or stochastic grammars requires making certain assumptions regarding the independence of the utterances to achieve tractability. Alternatively, a conditional model can better handle the contextual dependencies of the current utterance and its label to the future or past utterances, and their respective labels. Additionally, using a CRF prevents a bias that occurs in maximum entropy Markov models and discriminative Markov models that have a directed graphical model within their foundation. Such models tend to be biassed towards states with fewer successor states, whereas a CRF does not. As covered below, CRF is used in the most successful DA classification solutions to this date, making the advent of CRF one of the cornerstones that led to the current state of the art in solving the DA classification problem. Although not a solution to the DA classification problem, one such example of how CRF is included into existing models would be how Nakagawa et al. [20] incorporated CRF into a dependency tree-based method for sentiment analysis, and their proposed method outperformed baseline methods.

As the computation hardware became affordable, machine learning and deep learning solutions started to extend outside the boundaries of the artificial intelligence field, penetrating through several other research areas within computer science domain,

including NLP. In addition to the statistical and probabilistic solutions mentioned above, different machine learning solutions were pursued, in both DA classification problem and in the problems that are related to it. Silva et al. [21] fused a rule-based approach with a learning-based approach which used a Support Vector Machine (SVM), and obtained state of the art results of the time.

A significant milestone in NLP research was paved by finding that the words can be efficiently represented as vectors. Mikolov et al. [22] showed that words could be represented as vectors on a multidimensional space of word embeddings. Though the idea of having a distributed representation of words was not new, and was studied years ago by Rumelhart et al. [23], this research shed new light into how the embeddings of words with similar meanings are clustered close to one another, and how the distance vector between them indicates their contextual relation. Up to that point, most researchers working on the DA classification problem treated words as atomic units but chose to represent them by enumerating the words and giving them an integer index. Having a multidimensional representation for each atomic unit of utterances in a dialogue helped fuel the methods that learned to extract implicit features in an automated way, due to the contextual relation between representations of each word. For instance, Kim [24] used the word vectors in a convolutional neural network, which was then trained for sentence classification. The results of that research highlight how remarkable a role pre-training word vectors play in learning-based approaches to NLP problems.

Most of the contemporary research on DA classification now uses variations of machine learning methods. Lee and Dernoncourt et al. [3] studied short-text classification by modeling recurrent neural networks (RNN) and convolutional neural networks (CNN). Their models consider this task in two parts. The first part generates a vector representation of entire utterances, based on the vector representation of words covered above, using either the RNN or CNN architecture. Then, the current utterance is classified, considering the vector representation of this utterance, as well as a few previous utterances. Their method achieved state-of-the-art results with widely used datasets SwDA and MRDA, which were also mentioned above.

Three studies should be mentioned in order to cover the current state of the art in

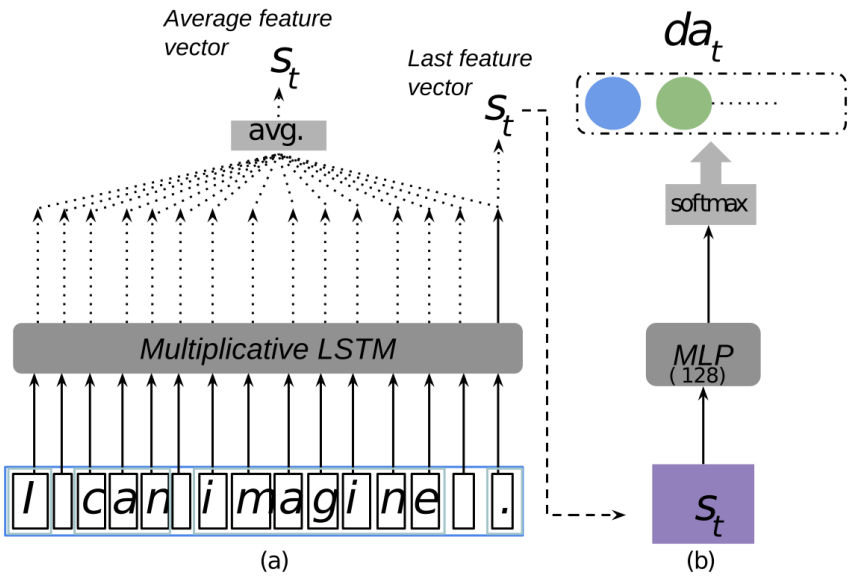

Figure 2.1: (a) Multiplicative LSTM (mLSTM) character-level language model to produce the sentence representation s_t. The character-level language model is pre-trained and produces the feature (hidden unit states of mLSTM at the last character) or average (average of all hidden unit states of every character) vector representation of the given utterance. (b) Utterance-level classification using a simple multi-layer perceptron layer with a softmax function. (**Figure source:** Bothe et al. [1])

DA classification using machine learning. Kumar et al. [4] developed a model based on bidirectional Long Short Term Memory (LSTM) units, introduced by Graves and Schmidhuber et al. [25]. Their model has access to the entire conversation and uses two layers of bidirectional LSTM units. The first layer forms a representation of the utterance, while the second layer considers all the representations of utterances in a conversation. Then, a CRF layer on top does the actual classification. This model achieved a near-human level of DA classification with SwDA dataset, considering that its results were only 5% less than the percentage of inter-annotator agreement for the dataset, which is 84%. [26] Alternatively, Bothe et al. [1] proposed a character-level RNN model for DA Classification problem. The model is a multiplicative LSTM network, which was studied by Krause et al. [27]. Importantly, to make the proposed model applicable to practical human-computer interaction scenarios, when the model is considering an utterance in a dialogue, it only has access to preceding and the current utterances in that dialogue, but not the future ones. Figure 2.1 visualizes

the baseline model of the authors, the top layer of which is improved by an RNN setup. Finally, Chen et al. [26] examined another CRF-based deep learning model. Instead of LSTM units, they used bidirectional Gated Recurrent Units (GRU) that were studied by Cho et al. [28] and included hierarchical semantic inference with memory. The results they obtained constitute the current state-of-the-art according to tests run on both SwDA and MRDA datasets, which yielded the accuracies 81.3% and 91.7%, respectively.

2.2 Machine Translation

Similar to DA Classification problem, MT is a task various versions of which have been studied for decades. Initial efforts mainly involved statistical and probabilistic approaches, but as covered below, the recent studies are more focused on solutions that are based on learning methods.

In a 1993 article, Brown et al. [29] studied statistical methods in machine translation. They used a readily available bilingual corpus, in which each dialogue was available both in English and in French. For each pair, their method attempts to find out the words in each sentence that correspond to one another, by using a statistical strategy to assign different probabilities to different word alignments in those sentences. They claimed that word by word alignment approach in machine translation is inherent to any sufficiently large bilingual corpus, due to the limited use of linguistic information in their statistical model.

Vogel et al. [30] used a similar probabilistic word alignment approach for sentence translation, utilizing a first-order Hidden Markov Model, a model which is used frequently in speech recognition solutions at the time. However, as opposed to assigning alignment probabilities using the absolute positions of the alignment of the words, they elected to use the relative positions of the words. Obtaining results that are on par with the research efforts at the time, they argued that their model was more robust in handling languages with many compound words, which may correspond to a permutation of multiple other words in another language.

Och and Ney [31] offered another statistical MT solution which uses a maximum

entropy model, and considers the sentences in source and target languages as features, along with any hidden variables. They demonstrated that the performance of a baseline statistical system for MT is considerably improved with this approach. It is noteworthy that this approach was extendible, as the model allowed adding new features, despite the authors recognized the computational complexity of handling complex features.

Och [32] stated that there was a potential mismatch between the methods used to quantify the success of a translation approach, such as BLEU mentioned above, and the expected performance of an MT solution in realistic scenarios. The methods to quantify the performance were using statistical analysis based on available test data. Hence, getting an adequate score would not conclusively indicate that a solution would work as successfully, especially with any text previously not seen. As an improvement, Och presented new criteria with which to count the errors in a machine translation solution. The results indicate that, trying to optimize the error rate during the training of the statistical models, which was the critical part of the new criteria proposed, yields solutions that work more successfully, even with previously unseen text.

As well as methods that focus on conducting the translation with techniques that use words as the atomic unit of information, researchers also studied various approaches that are based on treat phrases as the smallest unit. Chiang [33] presented a synchronous context-free grammar which learns to consider the phrases hierarchically, including any subphrases nested within a phrase. Considering the fundamental syntactical notion that a language is a hierarchical structure, and applying it to the translation task, their model surpassed the accuracy of another phrase-based translation system by Och and Ney [34], which was considered state of the art at the time.

One particular challenge of working on MT task is the fact that any NLP technique to be used in research should be applied to both languages and requires bilingual data to be learned and tested. Apart from improving the statistical methods to conduct MT, some studies also covered attempts to advance the linguistic foundation behind those approaches. Up to that point, for instance, many researchers used the available monolingual or bilingual data to learn parse trees, and rule extraction was automat-

ically done from the few best parse trees obtained. As those few trees covered only a fraction of the cases encountered, the performance of the translation systems that depended on such a rule extraction technique suffered. Mi and Huang [35] improved this automated rule extraction method by offering a representation of what they call a packed forest of parse trees in the form of a context-free grammar, which helps to store exponentially more parse trees compactly.

Similar to DA Classification task, as machine learning methods became affordable with cheaper hardware, the notion of an alternative representation of words became an attractive idea. Klementiev et al.[36] studied a distributed representation of words derived from unlabeled, parallel, bilingual texts. Their approach treats each word as a separate problem and similar to some of the other approaches mentioned below, tries to align the representations of words in two languages jointly, using statistical correlation and co-occurrence of each problem. (i.e., word) As covered in the previous subsection, another phenomenal research effort in this tract belongs to Mikolov et al. [22] who, as covered above, demonstrated the vector representation of words in multidimensional spaces, in a manner that establishes a proven correlation between the spatial distance of a pair of words represented in that space, and their contextual relation. Their work significantly facilitated machine learning solutions to be adopted by NLP tasks in recent years.

Another research by Mikolov et al. [2] showed that, with vector representation of words and the relationship between their meanings and placements in the multidimensional space, word translation task can be considered as a learning problem where one can construct a linear transformation between two monolingual multidimensional spaces, with relatively little, bilingual parallel data. Figure 2.2 represents a demonstration of the similarities between the relative placements of the words representing universal concepts within their respective monolingual vector spaces. This similar placement forms the foundation of the method proposed by the authors. Although their results did not come close to the accuracy of translation conducted by human experts, it was a remarkable demonstration of how practical this approach was, as it merely learned linear transformation by stochastic gradient descent. Further research efforts in this area build on this cornerstone approach and improve their results significantly. One example of an issue with the linear transformation between monolingual

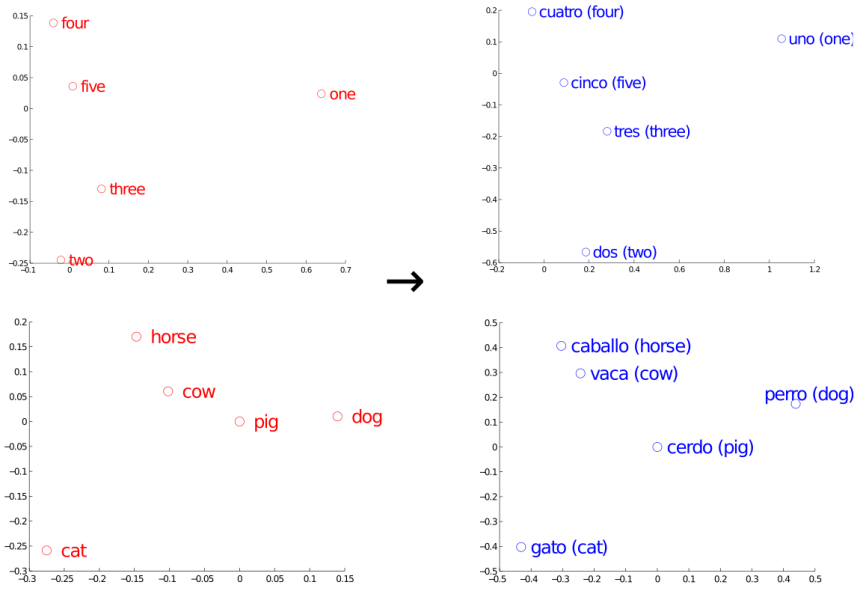

Figure 2.2: Distributed word vector representations of numbers and animals in English (left) and Spanish (right). The five vectors in each language were projected down to two dimensions using Principal Component Analysis, and then manually rotated to accentuate their similarity. It can be seen that these concepts have similar geometric arrangements in both spaces, suggesting that it is possible to learn an accurate linear mapping from one space to another. (**Figure source:** Mikolov et al. [2])

vector spaces is that there are hub words in any monolingual multidimensional space which can cause incorrect translations, as the hub words are the nearest neighbor to many other words due to their semantic correlation. Consequently, when vector representation of a word in a source language is translated to its representation in a target language, a hub word in the target language may be closer to the transformed vector, than a better semantic counterpart is. Dinu and Baroni [37] addressed this problem by using similarity vectors instead of distance alone, and by proportionally penalizing the similarity vector of each word, based on how big of a hub it is.

As the learning methods for MT further developed, one focus of research became the effort to minimize and eliminate the human expert supervision involved in creating parallel corpora on which the models can be trained. Artetxe et al. [38] demonstrated that almost unsupervised learning could be achieved by using a bilingual dictionary that has as few as 25 words. Their results showed that the performance of a machine

learning method need not suffer from the use of fewer words in bilingual parallel data.

Smith et al. [39] showed that it is possible to create the entire bilingual dictionary of words by considering the words written as the same character sequence in both languages to have the same meaning. Considering that many languages use different alphabets, this automated composition of the bilingual dictionary does not extend to all languages. Authors also present their evidence showing that learning model they propose performs better when it is trained with a bilingual dictionary that was formed by human experts. Despite the shortcomings of this automated bilingual data generation attempt, their usage of inverted softmax significantly increases the translation accuracy, primarily when the expert dictionary is used.

Using character-level information, such as the idea of considering the words with the same syntax to have the same semantics has its limitations. However, the most recent studies in MT task that attempt to translate without supervision show remarkable improvements. Conneau et al. [40] proposed one such combination of techniques which help align vector spaces of monolingual word embeddings. They utilize adversarial training to construct an unsupervised bilingual dictionary, where one of the trained models tries to distinguish the original language of a word, and another one tries to transform one of the vector spaces as similar as it can to the other one, thereby making the origin of the words included in the two bilingual spaces harder to distinguish. This approach, merged with their novel means of solving the hubness problem, gives the new method an edge, which, for some language pairs, helps it surpass even their supervised counterparts.

Criticizing previous research efforts in unsupervised MT for not being sufficiently robust in realistic conditions, Artetxe et al. [41] recently offered an alternative, where an automated bilingual dictionary is initialized using the structural similarities of each word embedding space, and both the dictionary and the performance of the model are incrementally improved iteratively, with no supervision. To form the bilingual dictionary, they first acknowledge that the respective axes of the monolingual word embedding spaces must be aligned in a tractable manner, so that the j_{th} axis of both spaces would have a similar semantic connotation. They conjecture that, once the axes are aligned, the similar vectors in different monolingual vector spaces can be

inferred and such word pairs with similar vector representations can be used to form a dictionary in an automated fashion. Still, the task of aligning the axes of word embedding spaces, as well as the exhaustive deduction of the most similar word pairs in each space are intractable. Taking this into account, they propose certain simplifications, such as using a word similarity matrix of each language, as opposed to the typical word embedding matrix, and sorting the values in each row of both of the monolingual word similarity matrices. According to the data they present alongside their research, the resulting model they propose is an improvement on all previous works in word translation by word mappings, supervised or otherwise.

2.3 Cross-lingual Text Classification

Compared to research on MT and DA classification, the idea of incorporating MT techniques into NLP solutions to remedy the lack of labeled data is relatively new. This section covers the advancements and studies by various researchers who studied different aspects of this idea and worked on implementing it.

Duh et al. [42] posed the question of whether MT methods matured enough to aid in learning cross-lingual solutions, by focusing on sentiment classification problem. The authors view the capability as a domain adaptation problem and run experiments to find out how translating the labeled data affects the accuracies achieved by a sentiment classifier. Based on the results obtained, they argue that due to the domain mismatch caused by the differences in the word distributions in the source and target domains (i.e., languages), a decrease in the accuracy of the classifier is bound to occur, even if a semantically perfect MT method was used. Nonetheless, they acknowledge the positive results achieved by previous work in the area and suggest adopting specialized adaptation methods that better address the differences between monolingual adaptation techniques and the cross-lingual domain adaptation problem.

In order to achieve cross-lingual sentiment analysis, Mohammad et al. [43] proposed two different constructs, which are strikingly similar to the LDAC and UDAC architectures proposed in this thesis. The first approach is based on translating text in a target language to English, for which there are multiple powerful sentiment analy-

sis tools. The second approach aims to translated labeled data in English to a target language and use it as an additional source of information in training a sentiment analysis system in the target language. These two approaches resemble our UDAC and LDAC solutions, respectively. Their experiments show that the first approach they offer yields results that are comparable to existing solutions in Arabic sentiment analysis of the time. Their second approach, which is based on using data translated to a target language as supplementary input, is also shown to increase the accuracy of sentiment classification in the target language. However, their experiments show that the classification accuracy decreases when translated data is used as the sole input in training.

While comparing different methodologies to achieve aspect-based cross-lingual sentiment classification, Barnes et al. [44] mention similar concerns regarding the usage of MT methods to transfer the information from the domain of a source language to the domain of a target language. Citing previous research, they reflect on how the noise introduced by the poor quality of a translation may deter a cross-lingual classifier from achieving a substantial accuracy. [43, 45] Noting that languages for which there is no sufficient data are also the ones that are likely to be poorly translated, they stress the importance of reducing the factors that may lead to a low-quality translation, when an MT method is used, including the issue of mismatching domains, which was mentioned by Duh et al.

Lee and Lee [46] applied a sentence-based MT methodology similar to the one we present on the task of Question Answering. They explore both the option of translating data to a target language and the option of translating the target language data to the source language, for which there are various existing QA solutions. The authors also compare this MT idea with an alternative solution that uses a Generative Adversarial Network (GAN). Their findings indicate that GAN-based approach can compete with MT-based solutions they investigate, using fewer linguistic input. Using both of these solutions together, they manage to set the new state-of-the-art in QA task in Chinese.

Perhaps the study which is most relevant to this thesis is the research conducted by Martínek et al. [47] that focuses on multilingual and cross-lingual DA classification.

They propose two main architectures. Their multilingual architecture requires a set *S* of different languages for which there is available, labeled DA data. Uniting these different sources of data as a single, multilingual dataset, they are able to train a multilingual classifier which can label DAs in any language that is included in set *S*. Although the model can be trained once and used for all languages in *S*, it needs to be retrained once a new language is added to the set. Their cross-lingual model, on the other hand, resembles the architecture of UDAC. It works by training a single classifier in a source language, for which there is available data, and projects the dialogues in a target language to the source language for classification. Similar to UDAC, once the classifier is trained in the target language, it does not need to be retrained for a new target language. However, a notable difference between UDAC, which utilizes an MT method based on an orthogonal transformation of vector spaces, in their cross-lingual model, the authors elect to use a linear transformation technique named Canonical Correlation Analysis. [48] As classifiers, the authors experiment with two different CNN configurations as well as a BiLSTM configuration. They test their proposed approaches on German and English dialogues in Vermobil dataset [49] using word2vec word embeddings. [22] Their results indicate that, despite being less flexible due to the requirement of retraining for new languages, their multilingual approach outperforms the cross-lingual one. They also found that BiLSTM classifier outperforms the alternative CNN configurations in most of the cases.

Although there are a few similarities between the solutions proposed in this thesis and the ones offered by Martínek et al., there are a few key differences between the two research efforts. Firstly, their focus is on producing a viable cross-lingual model, at least one of their solutions (i.e., multilingual model) requires datasets to be available in all the languages in which it can classify DAs. Even though their results show that the multilingual model is more accurate than its alternative, this thesis focuses on models that attempt to mitigate the lack of data in target languages. Secondly, Martínek et al. focus mainly on the overall accuracies of the methods they propose, while our research does a more thorough and in-depth analysis, covering confusion matrices and dialogue excerpts, which is a practice seen in numerous monolingual DA studies. [4, 26] Last critical difference is the datasets being used. Verbmobil dataset used by the authors may be a more effective tool for our research in evaluating the

solutions we propose. However, to download the dataset, one either has to be a paid member of one of the distributing organizations or to pay a fee. Thus, in order for any interested party to be able to verify our findings and improve them, we opt for the use of publicly available datasets SwDA and MRDA, even though neither of the datasets provides dialogues in two different languages.

As the research efforts provided above show, applying a cross-lingual approach to NLP problems to mitigate the lack of labeled data is an increasingly active field of research which is worth exploring.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter details the solution approach adopted in this thesis. We start by rationalizing the critical ideas behind the technique we propose to solve the problem. In the following sections, we introduce two solutions named Localized DA Classifier (LDAC) and Universal DA Classifier (UDAC). In order to solve the problem at hand, each of the solutions is designed to combine an existing DA classification solution with a machine translation technique. The details of their design are explored individually in the respective sections named after the solutions. Following the sections for the two proposed solutions, we present the translation method utilized within our proposed solutions. Lastly, we describe the DA classification models picked to be used, outlining their architectures, and providing the rationale behind selecting these particular DA classifiers.

3.2 Solution Approach

The most straightforward solution that can be conceived to remedy the lack of DA classification data in a language is to create a labeled dataset in that target language. However, one obvious shortcoming of this approach is the need to duplicate the effort of data collection and expert labeling for any new language in which DA classification problem is to be studied.

An idea to fix this problem without having to duplicate manual human labor for each new target language is to utilize an MT solution on an existent labeled dataset. If a

DA classification dataset in a source language can be adequately translated to a target language using an MT method, then without needing to label any new data explicitly, a dataset in the target language can be obtained. If the resulting dataset is not too low in quality, this method can also be extended to obtain DA classification datasets in numerous languages for which there is no available labeled corpus.

This thesis adopts this notion and tests whether the newly generated corpus in a target language provides a high-quality source of information for researchers to utilize. We offer two different mechanisms to combine this MT approach with a DA classification method, and the success of each solution is examined in Chapter 4.

3.3 Localized DA Classifier (LDAC)

At the highest level of abstraction, LDAC is a method that stems from the idea of training a DA classifier in a target language without having labeled data in that target language. A similar approach is utilized in the field of image captioning by Samet et al. [50] An intuitive construction that can be conceived to achieve this goal relies on keeping the architecture of the DA classifier intact and modifying the data on which it is being trained. LDAC makes use of this idea and employs an MT method to translates a dataset into the target language before the training of the classifier begins. Once the dataset is translated to the target language, the translated data is sent to the DA classification solution as input. As a result, the DA classification network trained on this translated dataset is expected to be able to classify the dialogues in the target language, without ever examining an actual labeled conversation in that language.

Figure 3.1 visualizes the general structure of the training process for LDAC. The architecture can be adapted to any language, and it does not specify a particular MT or DA classification. The fundamental approach is to translate an available, labeled DA dataset, and to train a DA classifier on that labeled, translated data. The compatibility of the specific translation and classification solutions to be used is left to be handled as a practical concern at the implementation level, as opposed to being handled in the architecture level. This way, the abstract idea LDAC embodies can be flexibly implemented using many MT and DA classification solutions available.

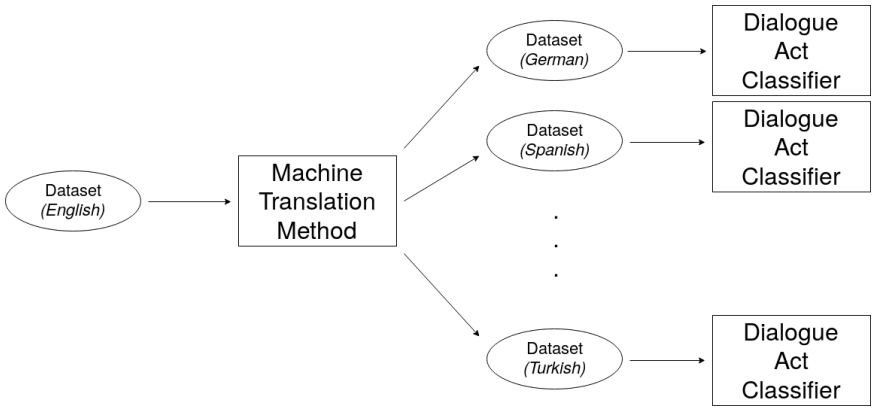

Figure 3.1: Depiction of the training and testing processes of LDAC

One clear challenge in proposing an approach that is based on translating a dataset is the inability of testing it. The goal of this thesis is to train a DA classifier in a target language without having labeled data in that language. Consequently, as no data in the target language is available, it is tricky to test the actual performance of LDAC. To remedy this, we propose a particular testing scheme for LDAC. We conduct the training with a translated dataset, which is translated using a selected MT method. However, we use a different, widely used MT method to translate the testing data. In other words, the testing process is identical to the training flow shown in Figure 3.1, with the distinction of using another MT method for testing. We consider that using different translation solutions for training and testing data helps achieve a more impartial and unbiased evaluation of the performance of the proposed solution. This testing approach is elaborated in greater detail in Chapter 4.

3.4 Universal DA Classifier (UDAC)

UDAC takes the idea behind LDAC and attempts to eliminate the need for training a separate DA classifier for each target language. UDAC construction first trains a classifier by using any labeled DA dataset. After the training, a DA classifier in that language is obtained. Then, any data in any target language can be translated to the language in which the classifier is trained, and therefore can be labeled by the classifier. Through this approach, UDAC achieves classification in any target language by training a single classifier, as opposed to having to train a separate classifier in each

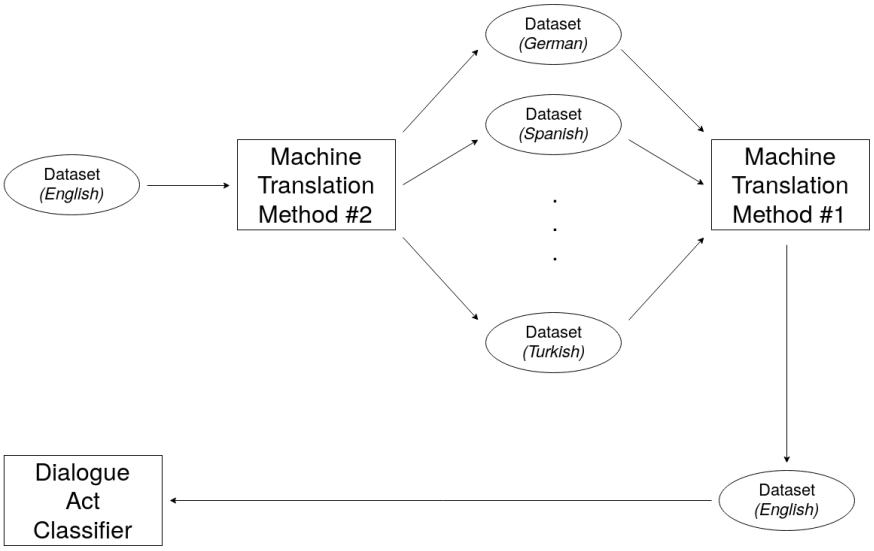

Figure 3.2: Depiction of the testing process of UDAC

target language. Similar to LDAC, UDAC is a design that can achieve DA classification in a language without being trained on a dataset compiled in that language.

Both the DA classification and the MT methods selected are independent of the proposed methodology. UDAC only requires the capability to train a DA classification method in a source language, and the facilities to translate any dialogue data in a target language to the source language in a manner that makes it possible to feed the data to the classifier as input.

The idea to have a classifier in a source language and to get any dialogue in a target language classified through translation seems painfully trivial. Nonetheless, it is still an idea worth investigating, considering its possible implications in DA classification, as well as many other NLP tasks. However, much like LDAC, it is challenging to come up with a way to evaluate the accuracy of UDAC, as we assume that there is no labeled dialogue data available in the target language.

To test UDAC, the only dialogue data we have at hand is the dataset in its original language, with which we train the DA classifier. In order to run tests with dialogues in target languages without having any labeled data in those languages, the data has to be translated, so that labeled testing data in the target language is obtained. This method of translating the original dataset into a target language needs to be done

through another automated, reusable MT method than the one being used in UDAC. A second MT solution is employed to prevent any bias of a single particular MT method affecting the evaluation process of the accuracy of UDAC. Assuming that testing data translated into the target language constitutes a sufficiently good method to evaluate, we then translate it back to the language in which the DA classifier was trained, using the MT method included in UDAC, and feed the result into the classifier for testing. Figure Figure 3.2 and Figure 3.3 features a straightforward demonstration of the testing and evaluation processes of UDAC, respectively. The testing process is further detailed in Chapter 4, along with a discussion of the effectiveness of this evaluation method by analyzing the results obtained.

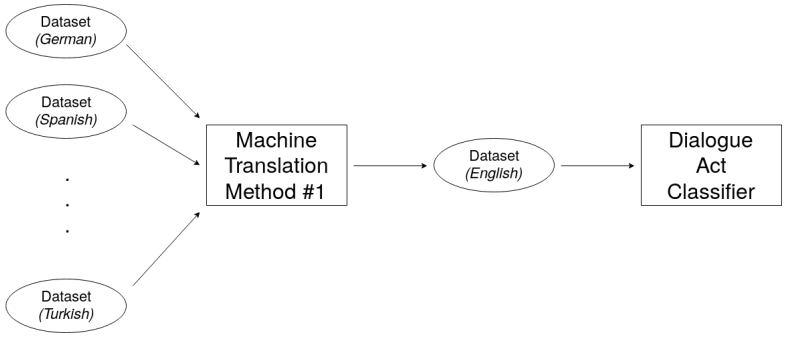

Figure 3.3: Depiction of the evaluation process of UDAC with dialogues in any target language

3.5 Translation Solution

Mikolov et al. [2] note that a central idea behind the capability of translating a vector in a word embedding space to another is how words are placed in the multidimensional, monolingual space, based on their semantics. Specifically, as they demonstrate, a tuple of words that represent concepts that are shared by all languages (e.g., numbers, animals) have similar relative placements in a monolingual word embedding space, regardless of whichever language is represented by that space. This notion of similar relative placements is rooted in the fact that such universal concepts, as well as the words that are used to represent them, carry contextual information on the concept itself. Although conducting a word-based MT on utterances may cause losing sentence-level grammatical and structural information, a word-based translation ap-

proach can help preserve the contextual information a word embedding represents. Additionally, the translation method to be used in this endeavor should be compatible with most of the recent studies in DA classification. Many of the contemporary studies in that field consider words as the atomic unit of information and use word embeddings to represent words in their models. Hence, picking a word-based MT method that uses word embeddings to represent words emerges as a rational choice.

The translation method used in the experiments given below was published by Smith et al. [39], and it uses a word-level translation method. Their method trains a linear orthogonal translation matrix between multidimensional word embedding spaces of a source and a target language, by using inverted softmax to train the matrix. As covered above, their study also introduces a method to form the bilingual dictionary without introducing any human expert signal. To comprise the bilingual dictionary data in such a manner, they consider the word pairs with the same syntax in each monolingual space to have the same semantics. In other words, they consider such pairs as translational pairs that belong to the bilingual dictionary. However, their results highlighted that this method of comprising the bilingual dictionary performs worse than having it formed by an expert. As a result, this thesis adopts their approach but trains the translation matrix by using bilingual data that is compiled using an expert signal. The authors of the method form the expert dictionary by using the most popular words in the English language, and their translations in the target languages, which were obtained using Google Translate. Further details as to the training of the translation matrix are covered in Chapter 4.

As discussed above, the translation method used in the experiments is the one proposed by Smith et al. [39]. Their study features an approach making use of inverted softmax to find the orthogonal, linear transformation between two monolingual vector spaces. They use 5K most common words in English and their counterparts in Italian as the training dictionary of the translation matrix. Similarly, our experiments form the bilingual dictionaries using the most 5K words common words in the source language (i.e., English), and their counterparts in the target languages. The most common words in English are obtained by picking the first 5K words from the fastText pre-trained monolingual word vectors for English, as the vectors are ordered by their frequency. The rest of the training process adheres to the model presented in the orig-

inal paper, although we did not apply dimensionality reduction to any of the trained alignment translation matrices. Adhering to the preference in the code sample provided by the authors for alignment matrix training, the target languages are aligned to the vector space of English and identity matrix is used to align English to its own vector space. [51] Finally, testing of the MT accuracy of the resulting alignment matrices was not conducted. The performance evaluation of the studied word-based MT method is already presented by the authors of the study as %38.0, %58.5 and %63.6 for Italian to English, for precisions 1, 5, and 10, respectively.

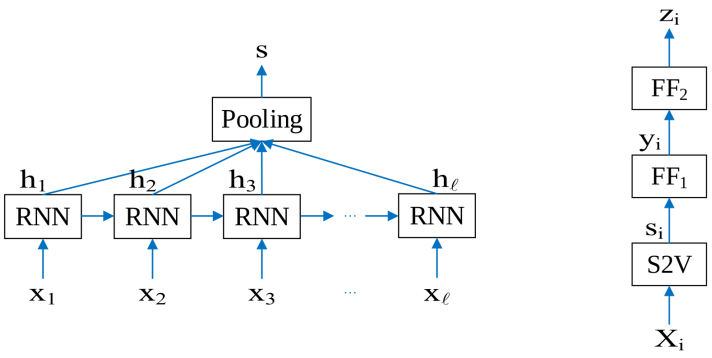

Figure 3.4: RNN architecture proposed by Lee and Dernoncourt. On the left, the first level of the network that generates the vector representation (i.e. the first level) of a short text $x_{1:l}$. On the right, the second level of the network which consists of a two-layer feedforward ANN used for predicting the probability distribution over the classes z_i for the i^{th} short-text X_i. S2V stands for short text to vector, which is the RNN architecture that generates s_i from X_i. (i.e. first level of the architecture) (**Figure source:** Lee and Dernoncourt [3])

3.6 DA Solutions

LDAC and UDAC are designed to work with a variety of DA classification and MT solution pairs. However, to achieve better compatibility of the methods to be used, the details of each DA and MT technique should be considered. As mentioned in the previous section, a word-based translation method was selected. Utterances translated

using a word-based method may lose some, if not all, of the sentence-level structural information. (e.g., conjugations, tenses) This notion implies that the resulting translated dataset may be considered noisy, and DA classification solutions to use this dataset should be effective in handling noisy data and inferring the cumulative contextual semantics of the entire utterance from the embeddings of the words included in them. This aspect of the translation process makes machine learning approaches a natural choice to be paired with the chosen MT method. Consequently, to achieve the best possible accuracy with the proposed LDAC and UDAC solutions, the DA classification techniques used should be learning-based.

In order to observe how LDAC and UDAC perform with different DA classification models, two different models were picked. Among the four state-of-the-art learning-based DA classification studies reviewed and featured in Chapter 2 above, two of the studies used models that employ CRF in their architecture. In order to see how the idea presented in this thesis performs with different models, among the studies employing CRF, only the one by Kumar et al. [4] was picked.

The second DA classification study is selected from the remaining two recent studies. Although the study by Bothe et al. [1] was considered, their method is based on processing the utterance as a single string and deducing character-level information. Considering how word-based translation and possible loss of structural information of the sentence in the new dataset may affect the efficiency of their character-based method, the second study with which to test our method was selected to be the model proposed by Lee and Dernoncourt [3].

3.6.1 Lee-Dernoncourt Model

The first DA classification architecture to be tested is the RNN model studied by Lee and Dernoncourt. Lee and Dernoncourt propose a neural network architecture with two levels. In the first level, using word embeddings, they encode the word-level information into producing an utterance-level vector representation using a layer of LSTM nodes, followed by a pooling layer. In the second level, the utterance-level representations are given as input to a 2-layered feedforward artificial neural network structure (ANN) output of which is the label prediction for the utterance. Figure 3.4

shows the proposed architecture. The authors experimented with incorporating representations of the previous utterances while classifying the current utterance label as well, by changing how the ANN nodes are connected. However, to keep the network structure more simple, and to offer an alternative architecture to the second solution, which covers an approach with better utterance-level memory, no utterance-level memory is adopted, and each utterance is fed to its separate node in the ANN, as demonstrated on the right of Figure 3.4.

3.6.2 BiLSTM-CRF Model

The second DA classification solution we utilize is the method proposed by Kumar et al. [4], unofficially named BiLSTM-CRF. As its name suggests, it relies on bidirectional LSTM units as well as a final CRF layer. It attempts to leverage the hierarchical structure of a dialogue, where words form utterances and utterances constitute the conversation, by using an architecture that analyzes the structure of the conversation at word-level, utterance-level, and conversation-level.

The neural network they propose is as follows. After an initial embedding layer, which is initialized by the pre-trained fastText word embeddings, there is a layer of bidirectional LSTM nodes, processing the word-level information, followed by a pooling layer that outputs a representation of the utterance. The utterance-level information is fed into a conversation-level layer of bidirectional LSTM nodes, on top which the CRF layer is placed. The CRF layer then outputs a label prediction for each utterance. Figure 3.5 demonstrates a visual model of the architecture proposed by the authors.

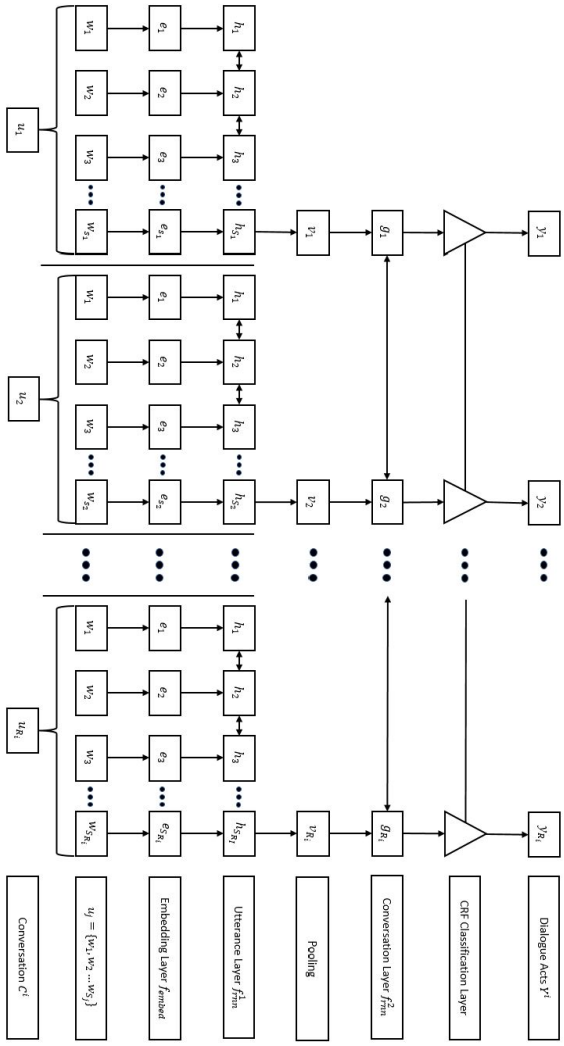

Figure 3.5: An illustration of the proposed hierarchical Bi-LSTM CRF model by Kumar et al. The input is a conversation C^i consisting of R_i utterances u_1, u_2, . . . u_{R_i}, with each utterance u_j itself being a sequence of words w_1, w_2, . . . w_{S_j}. As can be seen, there are four main layers, viz. embedding, utterance encoder, conversation encoder, and CRF classifier. The output is a DA prediction for each utterance in the conversation. (**Figure source:** Kumar et al. [4])

CHAPTER 4

EXPERIMENTS AND RESULTS

4.1 Experiment Setup

Within the scope of research conducted for this thesis, multiple experiments were conducted to observe how dataset translation approach performs in the DA classification task, using configurations proposed by LDAC and UDAC. This section details those experiments and the evaluation process.

The experiments we conduct are as follows. First, we run experiments with LDAC, using different languages, DA classifiers, and datasets. We also explore how the word order within the utterances affect the classification accuracy, by training LDAC with data where words in utterances are shuffled, as well as ordered. We also run experiments on UDAC in the same manner.

As mentioned in Chapter 3, using the same testing method for both training and testing processes may cause a bias in that method to affect the results of the experiments. Therefore, to testing LDAC and UDAC, a different route is taken. Before the training process, for each dataset, each conversation used for testing the accuracy of the classifier is translated to the target languages using Google Translate. In order to obtain a complete translation of the utterance that preserves the sentence-level information as much as possible, as opposed to translating each word separately, each utterance is considered as a single string and is translated as a whole.

During the testing phase of the experiments, the translated version of the testing data is used. In LDAC, the data translated through Google Translate is fed directly to the DA classifier for classification. As the training data was also translated to the target

language by the MT method selected for LDAC, the classifier can already work in that target language. In UDAC, however, the data translated through Google Translate needs to be translated back to the source language (i.e., English) on which the DA classifier trained, using the MT method which was selected to be part of the configuration of UDAC. After any translation phase required by LDAC or UDAC is complete, during the classification step, the word embeddings of each word occurring in the input text are collected from the monolingual vector space of the relevant language, based on the solution being tested.

We acknowledge that using Google Translate is an imperfect way to test the models that adopt the proposed translation methodology. However, the alternative effort of human experts translating all the testing data to all the target languages can not be automated or reused for other target languages. Hence, to explore an automated, reusable way of testing the proposed solution, which may even be extended to other NLP problems than DA classification, we opt for the use of Google Translate in the testing scheme. Due to this preference, we can translate the data to many target languages in an unsupervised manner. As a result, a mechanism that transforms both the training and the testing data is established for LDAC and UDAC.

The subsections in the remainder of this section cover in-depth descriptions of parameters included in the experiments, which apply to both LDAC and UDAC.

4.1.1 Word Embedding

The monolingual word embeddings we use in the experiments are the word vectors pre-trained on Wikipedia using fastText. The vector space has 300 dimensions, and the vectors were obtained by the skip-gram model described in [52]. The choice of using fastText word vectors differs from the word embeddings used in the selected MT or DA classification solutions. The reason this research opted to use fastText is the number of different languages for which a pre-trained monolingual word embedding space is available. While the MT and DA classification solutions we selected use pre-trained word embeddings that are only available in English, monolingual fastText word embeddings are available in 294 languages. The availability of pre-trained fastText word embeddings eliminates the need to train the monolingual word embedding

spaces for multiple languages from scratch, and MT solution can be applied directly to these pre-trained monolingual spaces.

When inputting data into DA classifiers, the relevant monolingual word embedding space is used. For instance, if the experiment is examining LDAC, as LDAC trains and tests data after it is translated to a target language, DA classifier is provided word embeddings from the vector space of the target language. In UDAC, however, MT method translates the data to the source language on which the DA classifier was trained. Therefore, for testing, the word embeddings are provided from the vector space of the source language.

4.1.2 Datasets

The experiments were conducted with two DA classification datasets, both of which are used frequently in previous DA classification research, as we covered in Chapter 2 of this thesis. The first is the Switchboard Dialogue Act (SwDA) corpus by Jurafsky et al. [53]. The corpus contains 1155 human-to-human telephone conversations in English. Each utterance in SwDA is classified by a label from a set of 42 labels, based on DAMSL taxonomy proposed by Core and Allen [54]. (e.g., STATEMENT-OPINION, BACKCHANNEL)

The second dataset used in the experiments is the ICSI Meeting Recorder Dialogue Act (MRDA) corpus by Janin et al. [55], which contains 75 meeting conversations between multiple human parties, in English. The original set of labels used in tagging the utterances in this dataset is called Meeting Recorder Dialogue Act Tagset, and it has 11 general, as well as 39 specific tags. Many previous research efforts on DA classification problem, including the methods used in our experiments, reduce this label set down to five main dialogue acts, namely Statement, Question, Floorgrabber, Backchannel, and Disruption. [3, 4] We follow the precedent set by them and use those five labels for utterances. Additionally, note that the labels do not represent solely syntactical constructs, and set of labels in each dataset features labels with semantic value. (e.g., SUMMARIZE/REFORMULATE label in SwDA, Floorgrabber label in MRDA)

Table 4.1: |C| is the number of Dialogue Act classes, |V| is the vocabulary size. Training, Validation and Testing indicate the number of conversations (number of utterances) in the respective splits.

Dataset	|C|	|V|	Training	Validation	Testing
MRDA	5	10K	51 (76K)	11 (15K)	11 (15K)
SwDA	42	19K	1003 (173K)	112 (22K)	19 (4K)

Note that, similar to the % label reserved for uninterpretable utterances in SwDA dataset, there is also a *z* label for purposefully uninterpreted utterances in MRDA dataset. Even though the previous research on the DA classification task uses five labels for MRDA, they do not mention how utterances with label *z* are treated. Due to the lack of precedent, we considered the best way to handle such utterances. Excluding the utterances with label *z* from the data being fed to the classifier alters the continuity of the dialogues and can undermine the accuracy of the classifier. As a result, in the experiments conducted on MRDA dataset, utterances with label *z* are included in our training and testing process, including the computation of overall accuracy. This decision was made to preserve the entirety of the data, as well as considering the ability of a classifier to differentiate between meaningful and meaningless utterances to be significant. However, as the label *z* has no particular semantic value, it is excluded from further analysis.

Another critical detail in how we process the utterances in the MRDA dataset is regarding the utterances with multiple labels. There are utterances in MRDA datasets are divided into short texts, for each of which, a separate label is given. To be able to differentiate between those, each of these utterance segments are separated in-place, and each such segment is considered as a different utterance. Considering how the datasets already include instances of incomplete sentences with proper labels, due to being recorded in a meeting setting, we hypothesize that this approach is not going to affect the classification accuracy dramatically.

4.1 shows the training, validation, and testing data available for the datasets. For both MRDA and SwDA datasets, we adhere to the data splits used by Lee and Dernoncourt [3] for training, validation, and testing sets, which were made public by the authors.

Table 4.2: Choices of hyperparameters for the model by Lee and Dernoncourt.

Hyperparameter	**Value**
LSTM output dim. (n)	100
LSTM pooling	max
LSTM direction	unidirectional
Dropout rate	0.5
Word vector dim. (m)	300

Table 4.3: Choices of hyperparameters for the model by Kumar et al.

Hyperparameter	**Value**
Pooling	Last
Word Embedding	300D fastText
Dropout	0.2
Bidirectional	True
Hidden Size	300
Learning Rate	1.0
Stacked LSTM Layers	1

[56] It is also noteworthy to point out that neither of the datasets is homogeneous in terms of the frequency of each label observed in the dataset. In SwDA, more than 50% of utterances have either a NON-OPINION or a BACKCHANNEL label. Similarly, in MRDA, the number of utterances that are assigned a STATEMENT label constitutes more than 50% of all the utterances. [4]

4.1.3 DA Method Experiment Specifications

The model proposed by Lee and Dernoncourt is used with its default loss, optimizer, and early stopping parameters. Namely, negative log-likelihood is minimized, Adadelta optimizer by Zeiler [57] is used, and an early stopping with the patience of 10 epochs is set. Also, the choice of values for hyperparameters by the authors, which is shown on 4.2, is followed.

In the experiments conducted with both datasets on the model by Kumar et al., as preprocessing, characters were converted to lowercase, and dots, commas, question marks, and exclamation marks were removed while preprocessing the input. Additionally, the words not found in the word embedding space of the target language were removed from the utterance, and the utterance is comprised of the remaining words, for which there are word embeddings.

The training and validation processes were conducted in batches that contained conversations with the same number of utterances, with a threshold size of 64. L2 regularization of 1e-4 as weight decay and Adadelta optimizer is used. Dropout was applied after every bidirectional LSTM encoding layer. The initial learning rate was set to 1.0, and it was halved every five epochs. Early stopping was used with the patience of five epochs. The preferences of the authors were followed when picking the values for the rest of the hyperparameters used in the architecture, and those values can be seen on 4.3.

4.1.4 Languages

One of the most critical parameters to set is the languages to which the datasets are to be translated. The Spanish language is the first language that is chosen, as it has been a language studied by previous MT researchers such as Mikolov et al. [2]. Secondly, as results of the earlier MT research covered by Vogel et al. [30] demonstrated that there are models that may favor languages with many compound words such as German. Hence, German is selected, as collecting data from a different language with different grammar and linguistic structure is deemed significant. As a third language, Turkish is selected to provide insight to the performance of the approach with a language that is not entirely European, with an alphabet that does not strictly adhere to the Latin alphabet used in English. Additionally, to have a set of results with which we can compare the bilingual experiments we conduct, a monolingual set of experiments are conducted where no translation takes place, and both the training and the testing processes are done in English.

4.1.5 Word Order

To observe if the ordering of the words in each utterance affects the resulting accuracy, we trained every experiment variation with a version of the dataset where the words in each utterance were ordered and with a version of the dataset where word order in each utterance was randomly shuffled. Given that the translation modifications that are implemented meant that most, if not all, sentence-level information could be lost in the bilingual experiments, where the training and testing were not conducted in the same language, we considered investigating how the word order of utterances changed the accuracy of the method to be necessary.

4.1.6 Implementation

Both the translation and the DA classification solutions referred to above are implemented in Python 3.6.8 programming language. [58] Libraries used to implement them are Keras, Tensorflow, and Numpy. [59, 60, 61] Additionally, to train the translation matrices, fastText library was used. [62]

4.2 Results and Discussion

LDAC and UDAC are tested with different languages, datasets, and DA classifiers, as covered in the previous section. The results obtained for LDAC and UDAC are presented in 4.4 and 4.5, respectively.

This section analyzes the results obtained from the experiments using four main methods. Firstly, we cover and review the overall accuracy obtained in the experiments. Secondly, to get a better sense of how accurately each class is learned, we analyze the confusion matrices obtained in some of the experiments. Thirdly, we provide a sample excerpt from each dataset and specifically analyze how it is translated, as well as how it is labeled. Finally, to assess whether utterance-based MT is a better fit for LDAC configuration, which typically uses a word-based MT method, we rerun a subset of the experiments we initially conducted, using an LDAC configuration that uses utterance-based translation. The final subsection analyzes the results of those

experiments.

4.2.1 Accuracy Analysis

The results reflect a loss of accuracy in all the experiments where the translation approach was applied, compared to the experiments where both the training and the testing were done in the original language of the datasets.

The experiments conducted on LDAC with SwDA dataset show a 16.67% decrease in accuracy of the Lee-Dernoncourt model when translating the data into Turkish, and a 27.69% decrease in accuracy of the BiLSTM-CRF model when translating the data into German, marking the highest losses recorded in our experiments for that dataset. Similarly, with LDAC on MRDA corpus, the highest accuracy losses observed for Lee-Dernoncourt and BiLSTM-CRF models were, 17.5% and 4.39%, respectively.

Results obtained from experiments using UDAC show a similar decrease in accuracy, compared to a typical monolingual DA classifier in English. On SwDA dataset, UDAC loses as much as 25.66% and 71.83% accuracy when translating to Turkish using Lee-Dernoncourt and BiLSTM-CRF DA classifiers, respectively. On MRDA corpus, the loss of accuracy recorded for Lee-Dernoncourt and BiLSTM-CRF models are 18.09% and 65.73%, respectively.

Another crucial detail in the results obtained is the accuracy rates obtained for the BiLSTM-CRF classifier trained on SwDA dataset without any translation. While BiLSTM-CRF model performs within a 1% range of the accuracy originally claimed by Kumar et al. [4], on SwDA, it performs with 84.35% accuracy, which is a significant increase from the 79.2% found by the authors. The percentage of agreement over the labels of the SwDA dataset by its annotators is 84%. [26] Considering this fact, despite strictly following the model architecture described by the authors, we argue that overfitting may have occurred in that particular experiment.

The results are also revealing as to the comparative performances of LDAC and UDAC. When Lee-Dernoncourt DA classifier is used, the accuracies obtained by LDAC is higher than the ones obtained by UDAC in all of the cases but one. However, when BiLSTM-CRF classifier is used, although LDAC still outperforms UDAC,

UDAC shows a significant loss of accuracy, especially in the experiments conducted on SwDA corpus.

There may multiple causes of the accuracy drop observed in UDAC when BiLSTM-CRF classifier is used. As shown in more detail below in the Excerpt Analysis subsection, due to the imperfections in both the translations obtained from Google Translate and from the MT method used in UDAC, translating the testing data twice (i.e. first to a target language and then back to the source language) causes a significant loss of contextual information. This loss of information may be the reason a hierarchical model such as the BiLSTM-CRF classifier may fail to learn DA classification properly, even though BiLSTM-CRF outperforms Lee-Dernoncourt model in all the monolingual experiments.

The loss of contextual information due to translating the data twice is significant enough to question the validity of the evaluation method chosen for UDAC. However, we leave the discussion of the validity of the testing methodology of UDAC to Excerpt Analysis subsection below. In the rest of this subsection, we focus mainly on the results of LDAC, as well as the results obtained for UDAC using Lee-Dernoncourt DA classifier. The results of UDAC using BiLSTM-CRF classifier requires a more in-depth analysis, which is provided in the following subsections.

As one may observe, among both LDAC and UDAC experiments, the losses are higher in the experiments conducted using SwDA dataset, than the ones that used MRDA dataset. We hypothesize that this difference stems from the number of labels used in tagging each dataset. In MRDA, the more complex label set was reduced to five labels due to the precedent set by the previous work, whereas in SwDA dataset, all of the 42 labels were used. Labeling elements of any type would be more challenging as the set of labels increase. The process of learning the intricacies between those classes would require more information. It is only natural that the classifiers perform worse on the dataset with more labels because some of the sentence-level information that is lost in translation is more critical to be able to differentiate between 42 classes, compared to the five classes in MRDA.

There is an apparent cause that may affect the loss of accuracy in the experiments with a target language different from the source language. We hypothesize that part

Table 4.4: Accuracies obtained with LDAC configuration. Leftmost column indicates the target languages, while *en* implies that no translation was conducted.

	SwDA				***MRDA***			
	Lee-Dernoncourt		**BiLSTM-CRF**		**Lee-Dernoncourt**		**BiLSTM-CRF**	
	ordered	*shuffled*	*ordered*	*shuffled*	*ordered*	*shuffled*	*ordered*	*shuffled*
en	**60.19%**	58.57%	**84.35%**	83.65%	76.91%	**77.51%**	89.46%	**89.68%**
de	53.08%	**53.48%**	56.66%	**56.73%**	**69.70%**	69.10%	**86.63%**	85.82%
es	**48.14%**	47.51%	**57.11%**	56.75%	66.93%	**67.99%**	**87.69%**	86.38%
tr	43.52%	**49.12%**	**57.26%**	56.93%	**60.35%**	60.01%	**86.68%**	85.29%

Table 4.5: Accuracies obtained with UDAC configuration. Leftmost column indicates the target languages, while *en* implies that no translation was conducted.

	SwDA				***MRDA***			
	Lee-Dernoncourt		**BiLSTM-CRF**		**Lee-Dernoncourt**		**BiLSTM-CRF**	
	ordered	*shuffled*	*ordered*	*shuffled*	*ordered*	*shuffled*	*ordered*	*shuffled*
en	**60.19%**	58.57%	**84.35%**	83.65%	76.91%	**77.51%**	89.46%	**89.68%**
de	**52.02%**	51.85%	12.99%	**13.13%**	**66.46%**	66.42%	**24.32%**	23.95%
es	44.89%	**46.63%**	12.57%	**12.96%**	**68.00%**	65.46%	**83.69%**	24.47%
tr	34.53%	**38.44%**	12.52%	**13.23%**	58.95%	**59.42%**	**83.03%**	83.42%

of the reason for the decrease in the accuracy of the model, compared to its counterpart where both the training and testing were done without any translation, is the loss of utterance-level information. The sample excerpts featured in Excerpt Analysis subsection below show that the contextual information is mostly preserved after translating the testing data once. However, information relevant to the grammatical structure of a sentence (e.g., conjugations) can be largely lost. In other words, upon translation, the utterances in a dialogue may become an array of contextually related, yet grammatically incompatible group of words. As a result, training mainly on the contextual information and not being able to infer the complete grammatical information is a compelling factor in the loss of accuracy in the experiments where the translation of the dataset takes place. Conversely, as explained below, it is also an important factor in why changing the word order does not affect the accuracy of the model.

The comparison of the accuracies obtained in each language is useful as well. In MRDA dataset, among the LDAC experiments where translation was conducted, Lee-Dernoncourt model performed with the highest accuracy when the dataset was translated into German word embedding space, followed by Spanish and Turkish, respectively. In comparison, LDAC with BiLSTM-CRF classifier performed best with Spanish word embeddings, followed by German and Turkish, respectively. Likewise, UDAC experiments with Lee-Dernoncourt classifier performed best with Spanish, German, and Turkish, respectively. It is important to note that, both for LDAC and UDAC, the accuracies of different languages are quite close (i.e., less than 3%) to one another for BiLSTM-CRF model, while the difference of accuracies in different languages in Lee-Dernoncourt model is over 9%.

Similar observations about the application of the translation approach in different languages can be made about the accuracies obtained from the experiments using SwDA dataset. In those experiments, LDAC using BiLSTM-CRF classifier performed best with Turkish, followed by Spanish and German while Lee-Dernoncourt model performed best with German, followed by Turkish and Spanish. With UDAC and Lee-Dernoncourt classifier, the decreasing order of DA classification accuracy of different languages is German, Spanish and Turkish, respectively. Similar to the experiments with MRDA corpus, the difference of accuracies in different languages reach up to

14% with configurations using Lee-Dernoncourt classifier, while it is less than 1% when BiLSTM-CRF classifier is used.

We attribute the higher difference of accuracies in different languages in Lee-Dernoncourt model to it not being able to capture the utterance-level and conversation-level information in different languages. Whereas in BiLSTM-CRF model, the design of which reflects the hierarchical structure of a conversation better, this contextual information is more successfully recognized, even when utterance-level information is partially lost in translation. Kumar et al. state BiLSTM-CRF being better at capturing the hierarchical structure of a conversation as one of the main factors in their model outperforming Lee-Dernoncourt classifier. (i.e., in the case where both training and testing is conducted in English, without any translation)

As seen in both 4.4 and 4.5, there were two experiments conducted on the same dataset, using the same model, and using the same target language. In one of them, the word order of each utterance in the training set of dialogues was preserved while in the other, the word order of utterances included in the training and validation split of the datasets was randomly shuffled as part of the preprocessing steps, before the training. We wanted to investigate the effect not preserving the word order of utterances on the accuracy of the trained model. Neither ordered nor shuffled option is proven to be better than another for all cases, as there are experiments supporting both the options. For instance, when both the training and the testing were done in English, the results of the LDAC experiment using SwDA dataset yielded higher accuracies for utterances with ordered words, while the ones using MRDA dataset yielded higher accuracies with utterances with shuffled word order. Furthermore, the results show that for most cases, shuffling the word order of an utterance alters the resulting accuracy by less than 2%.

The close accuracies of the corresponding experiment pairs with ordered and shuffled words indicate that word order does not significantly affect the DA classification accuracy. There may be multiple causes of this result. The first possible reason is the hypothesis we presented above, stating that the grammatical sentence-level information is, partially if not entirely, lost upon translating the dataset. Therefore, an utterance then becomes a list of words that present the same contextual information

without necessarily abiding by the grammar of the target language. This loss of information results in models that are learned mainly from the contextual information presented by the word embeddings, and not the grammar of the language. Based on the results, such contextual information is unaffected by the order in which the words were input to the network. However, this hypothesis alone is not sufficient to explain this phenomenon. Because, as exemplified above, the experiments that were conducted monolingually (i.e., solely in English) favor different options on this issue, based on the dataset that was used in the experiment. One straightforward explanation of seeing different results even on English-only experiments may be rooted in the ability of neural networks to handle noisy data. As detailed in Chapter 3, both DA classification models have a layer to form an utterance representation using the word-level data. The results we obtained are evidence that the robustness of deep learning models with noisy or unordered data helped both architectures encode the relevant information within their nodes, regardless of the order of the words within the utterances.

4.2.2 Confusion Matrices

To better analyze the trained models, we present confusion matrices of the experiments conducted on LDAC with BiLSTM-CRF classifier, as well as the ones on UDAC with Lee-Dernoncourt classifier. In all the experiments confusion matrices of which are presented, the word order of the utterances is kept intact. (i.e., ordered)

4.2.2.1 LDAC Confusion Matrices

The confusion matrices of the experiments conducted with LDAC configuration on MRDA dataset are presented in Table 4.6 and Table 4.7. The confusion matrices of the monolingual experiments are also included in the same tables for comparison.

LDAC with Lee-Dernoncourt classifier works mostly successfully for Backchannel(B), Floorgrabber(F) and Statement(S) labels in the monolingual model. However, Disruption(D) label is mostly misclassified either as F or S, while Question(Q) label is misclassified as S in more than 92% of the cases. Despite the success of the Lee-

Table 4.6: Confusion matrices for LDAC experiment on MRDA dataset with Lee-Dernoncourt model, using word-ordered utterances. TL is True Label and P is Prediction.

	TL \P	***(2028) B***	***(628) D***	***(1931) F***	***(1341) Q***	***(10287) S***
en	***B***	79.44%	00.00%	01.23%	00.00%	19.33%
	D	01.27%	00.00%	51.11%	00.00%	47.61%
	F	13.41%	00.00%	74.88%	00.00%	11.70%
	Q	03.88%	00.00%	03.43%	00.00%	92.69%
	S	07.00%	00.00%	01.28%	00.00%	91.72%
de	***B***	42.90%	00.00%	00.35%	00.00%	56.76%
	D	16.08%	00.00%	21.02%	00.00%	62.90%
	F	09.01%	00.00%	21.18%	00.00%	69.81%
	Q	02.46%	00.00%	03.28%	00.00%	94.26%
	S	01.40%	00.00%	00.54%	00.00%	98.06%
es	***B***	32.74%	00.00%	07.54%	00.00%	59.71%
	D	20.06%	00.00%	16.08%	00.00%	63.85%
	F	06.11%	00.00%	28.02%	00.00%	65.87%
	Q	03.50%	00.00%	15.36%	00.00%	81.13%
	S	03.09%	00.00%	01.79%	00.00%	95.12%
tr	***B***	08.38%	00.00%	00.00%	00.00%	91.62%
	D	28.34%	00.00%	05.25%	00.00%	66.40%
	F	02.74%	00.00%	06.94%	00.00%	90.32%
	Q	19.24%	00.00%	00.15%	00.00%	80.61%
	S	06.56%	00.00%	00.63%	00.00%	92.81%

Table 4.7: Confusion matrices for LDAC experiment on MRDA dataset with BiLSTM-CRF model, using word-ordered utterances. TL is True Label and P is Prediction.

	TL \P	***(2028) B***	***(628) D***	***(1931) F***	***(1341) Q***	***(10287) S***
en	***B***	43.34%	00.00%	01.82%	00.00%	54.49%
	D	13.22%	00.00%	61.15%	00.00%	25.16%
	F	03.63%	00.00%	83.01%	00.00%	13.21%
	Q	02.68%	00.00%	00.00%	00.00%	97.02%
	S	03.49%	00.00%	02.92%	00.00%	93.09%
de	***B***	50.99%	00.00%	30.52%	00.00%	17.11%
	D	24.36%	00.00%	35.19%	00.00%	38.22%
	F	27.29%	00.00%	44.02%	00.00%	27.34%
	Q	04.25%	00.00%	03.95%	00.00%	88.59%
	S	06.69%	00.00%	03.83%	00.00%	87.53%
es	***B***	53.94%	00.30%	28.65%	00.00%	16.96%
	D	24.04%	00.32%	49.84%	00.00%	25.48%
	F	21.80%	00.16%	64.47%	00.00%	13.52%
	Q	05.15%	00.00%	05.00%	00.00%	89.78%
	S	08.82%	00.10%	03.54%	00.00%	87.41%
tr	***B***	58.63%	00.00%	23.03%	00.00%	17.95%
	D	21.66%	00.00%	49.84%	00.00%	28.03%
	F	25.22%	00.00%	42.83%	00.00%	31.85%
	Q	03.88%	00.00%	05.29%	00.00%	90.23%
	S	09.22%	00.00%	03.90%	00.00%	86.54%

Table 4.8: Confusion matrices for LDAC experiment on SwDA dataset with Lee-Dernoncourt model, using word-ordered utterances. TL is True Label and P is Prediction.

	TL\P	***(208) aa***	***(765) b***	***(76) ba***	***(81) fc***	***(73) ny***	***(55) qw***	***(84) qy***	***(1317) sd***	***(718) sv***	***(94) x***
en	***aa***	32.69%	50.96%	00.48%	00.00%	00.00%	00.00%	00.00%	07.69%	05.77%	00.00%
	b	01.44%	96.08%	00.26%	00.00%	00.00%	00.00%	00.00%	00.78%	00.00%	00.00%
	ba	03.95%	02.63%	57.89%	00.00%	00.00%	00.00%	00.00%	13.16%	14.47%	00.00%
	fc	18.52%	22.22%	03.70%	00.00%	00.00%	00.00%	00.00%	24.69%	18.52%	00.00%
	ny	09.59%	90.41%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
	qw	01.82%	00.00%	03.64%	00.00%	00.00%	00.00%	21.82%	25.45%	14.55%	00.00%
	qy	02.38%	01.19%	04.76%	00.00%	00.00%	00.00%	34.52%	22.62%	21.43%	00.00%
	sd	00.38%	00.30%	00.23%	00.00%	00.00%	00.00%	00.08%	80.56%	11.69%	00.00%
	sv	00.84%	00.97%	01.25%	00.00%	00.00%	00.00%	00.14%	33.15%	57.52%	00.00%
	x	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
de	***aa***	18.27%	47.12%	00.00%	00.00%	00.00%	00.00%	00.00%	26.44%	03.37%	00.00%
	b	03.79%	86.54%	00.00%	00.00%	00.00%	00.00%	00.00%	07.19%	00.92%	00.00%
	ba	07.89%	07.89%	02.63%	00.00%	00.00%	00.00%	00.00%	36.84%	35.53%	00.00%
	fc	11.11%	30.86%	00.00%	00.00%	00.00%	00.00%	00.00%	43.21%	07.41%	00.00%
	ny	00.00%	94.52%	00.00%	00.00%	00.00%	00.00%	00.00%	01.37%	02.74%	00.00%
	qw	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	47.27%	20.00%	00.00%
	qy	04.76%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	53.57%	11.90%	00.00%
	sd	00.99%	00.38%	00.00%	00.00%	00.00%	00.00%	00.00%	86.71%	05.69%	00.00%
	sv	00.42%	00.28%	00.00%	00.00%	00.00%	00.00%	00.00%	56.41%	35.93%	00.00%
	x	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
es	***aa***	24.04%	40.38%	00.48%	00.00%	00.00%	00.00%	00.00%	20.67%	06.25%	00.00%
	b	03.79%	62.88%	00.78%	00.00%	00.00%	00.00%	00.00%	29.80%	00.00%	00.00%
	ba	06.58%	14.47%	17.11%	00.00%	00.00%	00.00%	00.00%	30.26%	26.32%	00.00%
	fc	13.58%	22.22%	00.00%	00.00%	00.00%	00.00%	00.00%	44.44%	07.41%	00.00%
	ny	02.74%	86.30%	00.00%	00.00%	00.00%	00.00%	00.00%	09.59%	00.00%	00.00%
	qw	00.00%	01.82%	00.00%	00.00%	00.00%	00.00%	00.00%	56.36%	21.82%	00.00%
	qy	02.38%	01.19%	04.76%	00.00%	00.00%	00.00%	00.00%	66.67%	07.14%	00.00%
	sd	00.46%	00.46%	00.23%	00.00%	00.00%	00.00%	00.00%	79.12%	12.83%	00.00%
	sv	00.42%	00.14%	00.00%	00.00%	00.00%	00.00%	00.00%	49.03%	43.87%	00.00%
	x	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
tr	***aa***	38.94%	23.56%	00.96%	00.00%	00.00%	00.00%	00.00%	15.38%	11.06%	00.00%
	b	50.98%	41.44%	00.26%	00.00%	00.00%	00.00%	00.00%	03.53%	00.65%	00.00%
	ba	02.63%	13.16%	39.47%	00.00%	00.00%	00.00%	00.00%	21.05%	15.79%	00.00%
	fc	01.23%	27.16%	09.88%	00.00%	00.00%	00.00%	00.00%	50.62%	06.17%	00.00%
	ny	84.93%	13.70%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
	qw	01.82%	00.00%	03.64%	00.00%	00.00%	00.00%	00.00%	70.91%	16.36%	00.00%
	qy	01.19%	04.76%	03.57%	00.00%	00.00%	00.00%	00.00%	60.71%	20.24%	00.00%
	sd	00.76%	03.64%	00.76%	00.00%	00.00%	00.00%	00.00%	77.15%	07.74%	00.00%
	sv	00.70%	03.34%	00.70%	00.00%	00.00%	00.00%	00.00%	53.76%	29.81%	00.00%
	x	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%

Table 4.9: Confusion matrices for LDAC experiment on SwDA dataset with BiLSTM-CRF model, using word-ordered utterances. TL is True Label and P is Prediction.

	TL \P	***(208) aa***	***(765) b***	***(76) ba***	***(81) fc***	***(73) ny***	***(55) qw***	***(84) qy***	***(1317) sd***	***(718) sv***	***(94) x***
en	***aa***	35.58%	40.38%	08.65%	01.44%	01.44%	00.00%	00.00%	04.33%	03.85%	00.48%
	b	03.79%	90.72%	00.65%	00.39%	01.05%	00.00%	00.13%	00.39%	00.13%	00.13%
	ba	07.89%	01.32%	64.47%	00.00%	00.00%	00.00%	01.32%	10.53%	05.26%	02.63%
	fc	00.00%	01.23%	00.00%	92.59%	00.00%	00.00%	00.00%	03.70%	02.47%	00.00%
	ny	13.70%	68.49%	00.00%	00.00%	17.81%	00.00%	00.00%	00.00%	00.00%	00.00%
	qw	00.00%	00.00%	00.00%	01.82%	00.00%	30.91%	16.36%	12.73%	09.09%	00.00%
	qy	01.19%	03.57%	02.38%	00.00%	00.00%	03.57%	29.76%	19.05%	19.05%	00.00%
	sd	00.30%	00.00%	00.38%	00.38%	00.00%	00.00%	00.15%	85.57%	07.74%	00.46%
	sv	01.11%	00.00%	01.11%	00.28%	00.00%	00.00%	00.70%	54.46%	34.82%	00.00%
	x	01.06%	00.00%	00.00%	01.06%	00.00%	00.00%	00.00%	00.00%	00.00%	94.68%
de	***aa***	00.00%	04.33%	13.94%	03.85%	00.00%	00.00%	01.44%	18.27%	00.00%	34.13%
	b	01.57%	04.44%	15.03%	02.09%	00.00%	00.00%	00.00%	03.92%	00.00%	57.65%
	ba	00.00%	03.95%	27.63%	02.63%	00.00%	00.00%	06.58%	34.21%	00.00%	11.84%
	fc	00.00%	03.70%	00.00%	80.25%	00.00%	00.00%	00.00%	16.05%	00.00%	00.00%
	ny	00.00%	01.37%	10.96%	02.74%	00.00%	00.00%	00.00%	02.74%	00.00%	38.36%
	qw	00.00%	00.00%	01.82%	01.82%	00.00%	03.64%	01.82%	70.91%	00.00%	00.00%
	qy	00.00%	00.00%	05.95%	00.00%	00.00%	00.00%	04.76%	72.62%	00.00%	00.00%
	sd	00.30%	00.38%	00.38%	00.68%	00.00%	00.00%	00.53%	91.42%	00.08%	00.53%
	sv	00.14%	00.00%	00.70%	00.97%	00.00%	00.42%	02.92%	84.96%	01.25%	00.14%
	x	00.00%	00.00%	00.00%	01.06%	00.00%	00.00%	00.00%	00.00%	00.00%	93.62%
es	***aa***	00.96%	05.77%	06.73%	11.06%	00.00%	00.96%	01.44%	25.00%	00.00%	24.52%
	b	00.13%	05.62%	06.14%	04.05%	00.26%	00.00%	00.13%	29.54%	00.00%	40.26%
	ba	00.00%	02.63%	30.26%	02.63%	00.00%	00.00%	01.32%	22.37%	00.00%	15.79%
	fc	00.00%	00.00%	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
	ny	00.00%	05.48%	04.11%	05.48%	00.00%	00.00%	00.00%	16.44%	00.00%	45.21%
	qw	00.00%	00.00%	00.00%	05.45%	00.00%	05.45%	03.64%	69.09%	00.00%	00.00%
	qy	00.00%	00.00%	04.76%	03.57%	00.00%	00.00%	17.86%	55.95%	01.19%	00.00%
	sd	00.00%	00.08%	01.14%	03.26%	00.00%	03.80%	03.49%	77.52%	00.53%	00.46%
	sv	00.00%	00.14%	01.53%	05.29%	00.00%	04.46%	09.19%	64.35%	04.46%	00.14%
	x	00.00%	00.00%	00.00%	07.45%	00.00%	00.00%	00.00%	00.00%	00.00%	80.85%
tr	***aa***	04.33%	27.88%	12.50%	07.21%	04.81%	00.00%	00.48%	23.08%	00.00%	05.29%
	b	02.88%	45.23%	22.35%	03.27%	02.75%	00.00%	00.00%	08.10%	00.00%	06.80%
	ba	00.00%	01.32%	22.37%	01.32%	00.00%	01.32%	03.95%	42.11%	00.00%	03.95%
	fc	00.00%	00.00%	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
	ny	17.81%	54.79%	04.11%	05.48%	08.22%	00.00%	00.00%	02.74%	00.00%	02.74%
	qw	00.00%	00.00%	01.82%	03.64%	00.00%	00.00%	00.00%	76.36%	00.00%	03.64%
	qy	00.00%	00.00%	02.38%	04.76%	00.00%	00.00%	07.14%	67.86%	01.19%	01.19%
	sd	00.00%	00.00%	00.91%	02.51%	00.00%	00.23%	01.44%	81.70%	00.30%	00.46%
	sv	00.14%	00.14%	00.14%	05.01%	00.00%	00.70%	02.23%	78.97%	00.14%	00.00%
	x	00.00%	00.00%	04.26%	07.45%	00.00%	00.00%	00.00%	01.06%	00.00%	82.98%

Dernoncourt classifier with three of the five labels in the monolingual experiment, in the cases where the translation of data to a target language takes place, the confusion matrices show a significant drop in the accurate predictions for each label except *S*. More than 50% of the cases for each of the other labels are misclassified as *S*. Overall, as seen in Table 4.6, LDAC with Lee-Dernoncourt classifier fails to capture the relevant information it needs when translation takes place. Consequently, partly due to the unbalanced distribution of labels in the dataset (i.e., utterances with label *S* comprising more than 50% of the MRDA dataset) mentioned in the previous section, the classifier learns a bias towards classifying utterances with the label *S*.

Experiments with LDAC using a BiLSTM-CRF classifier yield better results. The model trained on the dataset without conducting any translation learns to classify utterances of type *F* and *S* with success for more than half of the cases. However, more than half of the utterances of type *B* are misclassified as *S*. Similarly, all of the *Q* utterances are misclassified, mostly as *S*, while all the *D* utterances are misclassified, mostly as *F*.

The confusion matrices of the experiments with a translation step are similar to the confusion matrix of the monolingual case. Utterances with label *S* are classified correctly in more than 85% of cases in each experiment. While more than half of the utterances with label *F* are misclassified for target languages German and Turkish, more than 40% of them are classified correctly. Interestingly, while utterances with label *B* were misclassified in more than half of the cases by the model without the translation, the correct classification of Backchannel utterances rose above 50% of the test cases in all the target languages with which experiments were conducted. Unfortunately, the incorrect classification of Disruption(*D*) and Question(*Q*) utterances persisted in the experiments with a target language as well. In each of the LDAC models with BiLSTM-CRF classifier trained on a translated MRDA dataset, while Question(Q) utterances were misclassified mostly as *S*, Disruption utterances were misclassified relatively evenly as *B*, *F* or *S*.

The confusion matrices of the LDAC experiments conducted on SwDA dataset are presented in Table 4.8 and Table 4.9. As SwDA dataset uses a label classification with 42 labels, we demonstrate the confusion matrices for the ten labels which have

the highest frequency in the test data.

Experiments conducted on SwDA dataset using LDAC with Lee-Dernoncourt classifier, shown in Table 4.8, reveal that the classifier can not handle the noise that is introduced by translation. In the monolingual experiment, the highest percentage of prediction for utterances with labels Backchannel(*b*), Appreciation(*ba*), Yes-no-question(*qy*), Statement-non-opinion(*sd*) and Statement-opinion(*sv*) are their corresponding true labels. Yet, utterances with Agree/Accept(*aa*) label are misclassified as *b* in almost 51% of the cases while utterances with labels Conventional-closing(*fc*), Yes-answers(*y*), Wh-question(*qw*) and Non-verbal(*x*) are never correctly predicted. When translation is introduced, utterances with labels *b* and *sd* are still correctly classified. However, for the rest of the labels, correct classification fails in majority of the cases, with labels *fc*, *ny*, *qw*, *qy* and *x* never being correctly predicted.

In the experiments on SwDA that tested LDAC with BiLSTM-CRF classifier, for the monolingual case, classification of labels *aa*, *ny* and *sv* are incorrect in more than 50% of the cases for each label. Utterances with labels *b*, *ba*, *fc*, *sd* and *x* are classified correctly in most of the cases for those labels. Although more than half of the cases for the labels *qw* and *qy* are misclassified, the percentage of their correct classification is higher than the percentage of misclassification as any other individual label.

The confusion matrices of the experiments involving translation of SwDA dataset show that labels *fc*, *sd* and *x* are still correctly classified in the majority of the cases. However, in many cases labels *aa*, *ba*, *qw*, *qy* and *sv* are misclassified as *sd*. This can be observed clearly by observing the background colors of the table cells on the column of *sd* label.

When the semantics of the labels are considered, a striking pattern that spans both datasets can be seen. Regardless of the DA classifier being used, in all the experiments run with LDAC where the relevant dataset was translated to a target language, utterances with a question (i.e. *qw* and *qy* labels in SwDA; *Q* label in MRDA) were misclassified as statements (i.e. *sd* and *sv* in SwDA; *S* in MRDA) in most of the cases. The same problem occurred in the monolingual counterparts of the experiments as well, but the translation process caused a higher number of utterances to be misclassified as statements.

The most important factor contributing to these misclassifications is related to the composition of the datasets. As presented in the previous section, neither of the datasets are homogeneous in terms of the frequency of each label. In fact, in both datasets, half of the utterances included have labels indicating a statement. These frequencies are even more unbalanced when the testing split of the data is considered. For instance, as seen in the confusion matrices for MRDA dataset, the number of statements is almost ten times greater than the number of questions seen in the test data. Similarly, in the testing split of the SwDA dataset, the number of utterances with labels *qw* and *qy* is less than 100, while there are 1317 utterances with the label *sd*. Although the noise added by translation is visible when the confusion matrices are compared to their monolingual counterpart, without seeing sufficient amount of data for each label, it is natural that the classifiers can not learn correctly classifying the labels with smaller frequencies.

In addition to the heterogeneity of the datasets, there are other factors affecting the quality of the classification, especially in the models trained with translated data. The effect of translation approach on the label-based classification accuracy is perhaps most visible in the LDAC experiments with BiLSTM-CRF classifier conducted on SwDA. In the confusion matrix of the monolingual experiment, a diagonal line of gray background colors is visible. However, in the experiment where the target language is German, a clear misclassification of many labels as *sd* is visible. We attribute this behavior to the loss of sentence-level information upon translation. As stated above, many grammatical constructs are at least partially lost upon translation, and the sentences become a loosely tied array of words, which are contextually but not grammatically related. Consequently, some of the nuanced patterns that the model can learn from the dataset without any translation are lost, and misclassifications increase.

4.2.2.2 UDAC Confusion Matrices

The resulting confusion matrices of the experiments on MRDA dataset with UDAC are presented in Table 4.10 and Table 4.11. Confusion matrices of the monolingual experiments are also included.

Table 4.10: Confusion matrices for UDAC experiment on MRDA dataset with Lee-Dernoncourt model, using word-ordered utterances. TL is True Label and P is Prediction.

	TL \P	***(2028) B***	***(628) D***	***(1931) F***	***(1341) Q***	***(10287) S***
en	***B***	79.44%	00.00%	01.23%	00.00%	19.33%
	D	01.27%	00.00%	51.11%	00.00%	47.61%
	F	13.41%	00.00%	74.88%	00.00%	11.70%
	Q	03.88%	00.00%	03.43%	00.00%	92.69%
	S	07.00%	00.00%	01.28%	00.00%	91.72%
de	***B***	11.34%	00.00%	32.20%	00.00%	56.46%
	D	00.80%	00.00%	34.39%	00.00%	64.81%
	F	01.86%	00.00%	31.28%	00.00%	66.86%
	Q	02.09%	00.00%	03.73%	00.00%	94.18%
	S	00.19%	00.00%	01.91%	00.00%	97.90%
es	***B***	10.85%	00.00%	34.17%	00.00%	54.98%
	D	00.64%	00.00%	45.38%	00.00%	53.98%
	F	01.66%	00.00%	56.71%	00.00%	41.64%
	Q	02.98%	00.00%	06.26%	00.00%	90.75%
	S	00.11%	00.00%	04.30%	00.00%	95.60%
tr	***B***	00.15%	00.00%	08.53%	00.00%	91.32%
	D	00.80%	00.00%	41.88%	00.00%	57.32%
	F	04.19%	00.00%	16.93%	00.00%	78.87%
	Q	00.15%	00.00%	23.86%	00.00%	75.99%
	S	00.31%	00.00%	08.87%	00.00%	90.82%

Table 4.11: Confusion matrices for UDAC experiment on MRDA dataset with BiLSTM-CRF model, using word-ordered utterances. TL is True Label and P is Prediction.

	TL \P	***(2028) B***	***(628) D***	***(1931) F***	***(1341) Q***	***(10287) S***
en	***B***	43.34%	00.00%	01.82%	00.00%	54.49%
	D	13.22%	00.00%	61.15%	00.00%	25.16%
	F	03.63%	00.00%	83.01%	00.00%	13.21%
	Q	02.68%	00.00%	00.00%	00.00%	97.02%
	S	03.49%	00.00%	02.92%	00.00%	93.09%
de	***B***	20.46%	00.00%	61.64%	00.00%	17.90%
	D	26.11%	00.00%	01.11%	00.00%	72.77%
	F	23.30%	00.00%	23.98%	00.00%	52.72%
	Q	05.15%	00.00%	03.65%	00.00%	91.20%
	S	03.24%	00.00%	08.08%	00.00%	88.68%
es	***B***	18.74%	00.00%	63.76%	00.00%	17.50%
	D	29.14%	00.00%	01.11%	00.00%	69.75%
	F	24.91%	00.00%	24.91%	00.00%	50.18%
	Q	04.10%	00.00%	04.18%	00.00%	91.72%
	S	03.01%	00.00%	08.84%	00.00%	88.15%
tr	***B***	20.86%	00.00%	62.77%	00.00%	16.37%
	D	30.73%	00.00%	04.78%	00.00%	64.49%
	F	25.48%	00.00%	25.69%	00.00%	48.83%
	Q	06.34%	00.00%	04.18%	00.00%	89.49%
	S	05.11%	00.00%	09.20%	00.00%	85.69%

Table 4.12: Confusion matrices for UDAC experiment on SwDA dataset with Lee-Dernoncourt model, using word-ordered utterances. TL is True Label and P is Prediction.

	TL \P	***(208) aa***	***(765) b***	***(76) ba***	***(81) fc***	***(73) ny***	***(55) qw***	***(84) qy***	***(1317) sd***	***(718) sv***	***(94) x***
en	***aa***	32.69%	50.96%	00.48%	00.00%	00.00%	00.00%	00.00%	07.69%	05.77%	00.00%
	b	01.44%	96.08%	00.26%	00.00%	00.00%	00.00%	00.00%	00.78%	00.00%	00.00%
	ba	03.95%	02.63%	57.89%	00.00%	00.00%	00.00%	00.00%	13.16%	14.47%	00.00%
	fc	18.52%	22.22%	03.70%	00.00%	00.00%	00.00%	00.00%	24.69%	18.52%	00.00%
	ny	09.59%	90.41%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
	qw	01.82%	00.00%	03.64%	00.00%	00.00%	00.00%	21.82%	25.45%	14.55%	00.00%
	qy	02.38%	01.19%	04.76%	00.00%	00.00%	00.00%	34.52%	22.62%	21.43%	00.00%
	sd	00.38%	00.30%	00.23%	00.00%	00.00%	00.00%	00.08%	80.56%	11.69%	00.00%
	sv	00.84%	00.97%	01.25%	00.00%	00.00%	00.00%	00.14%	33.15%	57.52%	00.00%
	x	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
de	***aa***	19.71%	50.00%	00.48%	00.00%	00.00%	00.00%	00.00%	16.35%	01.44%	00.00%
	b	01.05%	87.19%	00.65%	00.00%	00.00%	00.00%	00.00%	01.31%	00.00%	00.00%
	ba	01.32%	07.89%	38.16%	00.00%	00.00%	00.00%	01.32%	31.58%	03.95%	00.00%
	fc	00.00%	32.10%	12.35%	00.00%	00.00%	00.00%	00.00%	32.10%	09.88%	00.00%
	ny	01.37%	94.52%	02.74%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
	qw	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	63.64%	09.09%	00.00%
	qy	01.19%	00.00%	04.76%	00.00%	00.00%	00.00%	00.00%	69.05%	08.33%	00.00%
	sd	00.23%	00.53%	00.46%	00.00%	00.00%	00.00%	00.08%	87.55%	04.40%	00.00%
	sv	00.14%	00.28%	00.00%	00.00%	00.00%	00.00%	00.28%	62.81%	26.88%	00.00%
	x	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
es	***aa***	13.94%	43.27%	05.77%	00.00%	00.00%	00.00%	00.00%	12.98%	05.77%	00.00%
	b	01.18%	58.04%	05.62%	00.00%	00.00%	00.00%	00.00%	01.44%	00.00%	00.00%
	ba	02.63%	06.58%	26.32%	00.00%	00.00%	00.00%	00.00%	15.79%	25.00%	00.00%
	fc	01.23%	23.46%	13.58%	00.00%	00.00%	00.00%	00.00%	44.44%	09.88%	00.00%
	ny	01.37%	75.34%	10.96%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
	qw	00.00%	10.91%	00.00%	00.00%	00.00%	00.00%	00.00%	34.55%	12.73%	00.00%
	qy	01.19%	04.76%	04.76%	00.00%	00.00%	00.00%	00.00%	35.71%	10.71%	00.00%
	sd	00.68%	00.46%	00.53%	00.00%	00.00%	00.00%	00.00%	68.56%	12.45%	00.00%
	sv	01.11%	00.84%	00.28%	00.00%	00.00%	00.00%	00.00%	35.93%	45.68%	00.00%
	x	00.00%	100.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
tr	***aa***	34.13%	31.25%	03.85%	00.00%	00.00%	00.00%	00.48%	14.90%	00.96%	00.00%
	b	46.67%	41.31%	03.27%	00.00%	00.00%	00.00%	00.00%	03.27%	00.00%	00.00%
	ba	01.32%	25.00%	21.05%	00.00%	00.00%	00.00%	00.00%	22.37%	10.53%	00.00%
	fc	08.64%	51.85%	02.47%	00.00%	00.00%	00.00%	00.00%	19.75%	04.94%	00.00%
	ny	76.71%	16.44%	04.11%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
	qw	03.64%	05.45%	05.45%	00.00%	00.00%	00.00%	00.00%	49.09%	07.27%	00.00%
	qy	00.00%	14.29%	01.19%	00.00%	00.00%	00.00%	00.00%	39.29%	07.14%	00.00%
	sd	00.68%	07.21%	00.61%	00.00%	00.00%	00.00%	00.00%	56.34%	04.56%	00.00%
	sv	00.56%	07.66%	01.11%	00.00%	00.00%	00.00%	00.00%	44.71%	15.46%	00.00%
	x	00.00%	86.17%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%

Table 4.13: Confusion matrices for UDAC experiment on SwDA dataset with BiLSTM-CRF model, using word-ordered utterances. TL is True Label and P is Prediction.

	TL \P	***(208) aa***	***(765) b***	***(76) ba***	***(81) fc***	***(73) ny***	***(55) qw***	***(84) qy***	***(1317) sd***	***(718) sv***	***(94) x***
en	***aa***	35.58%	40.38%	08.65%	01.44%	01.44%	00.00%	00.00%	04.33%	03.85%	00.48%
	b	03.79%	90.72%	00.65%	00.39%	01.05%	00.00%	00.13%	00.39%	00.13%	00.13%
	ba	07.89%	01.32%	64.47%	00.00%	00.00%	00.00%	01.32%	10.53%	05.26%	02.63%
	fc	00.00%	01.23%	00.00%	92.59%	00.00%	00.00%	00.00%	03.70%	02.47%	00.00%
	ny	13.70%	68.49%	00.00%	00.00%	17.81%	00.00%	00.00%	00.00%	00.00%	00.00%
	qw	00.00%	00.00%	00.00%	01.82%	00.00%	30.91%	16.36%	12.73%	09.09%	00.00%
	qy	01.19%	03.57%	02.38%	00.00%	00.00%	03.57%	29.76%	19.05%	19.05%	00.00%
	sd	00.30%	00.00%	00.38%	00.38%	00.00%	00.00%	00.15%	85.57%	07.74%	00.46%
	sv	01.11%	00.00%	01.11%	00.28%	00.00%	00.00%	00.70%	54.46%	34.82%	00.00%
	x	01.06%	00.00%	00.00%	01.06%	00.00%	00.00%	00.00%	00.00%	00.00%	94.68%
de	***aa***	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	61.06%	00.00%	25.48%
	b	00.00%	00.00%	01.70%	00.00%	00.00%	00.00%	00.00%	65.75%	00.00%	26.14%
	ba	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	01.32%	81.58%	00.00%	07.89%
	fc	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	62.96%	00.00%	17.28%
	ny	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	43.84%	00.00%	43.84%
	qw	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	100.00%	00.00%	00.00%
	qy	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	01.19%	97.62%	00.00%	00.00%
	sd	00.00%	00.00%	00.08%	00.00%	00.00%	00.00%	00.08%	96.74%	00.00%	00.53%
	sv	00.00%	00.00%	00.14%	00.00%	00.00%	00.00%	00.56%	95.54%	00.00%	00.28%
	x	00.00%	01.06%	42.55%	00.00%	00.00%	00.00%	00.00%	34.04%	00.00%	02.13%
es	***aa***	00.00%	00.00%	00.48%	00.00%	00.00%	00.00%	00.00%	74.52%	00.00%	15.87%
	b	00.00%	00.00%	01.57%	00.00%	00.00%	00.00%	00.00%	83.92%	00.00%	10.98%
	ba	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	85.53%	00.00%	03.95%
	fc	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	58.02%	00.00%	17.28%
	ny	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	64.38%	00.00%	26.03%
	qw	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	96.36%	00.00%	00.00%
	qy	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	01.19%	96.43%	00.00%	00.00%
	sd	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.23%	95.90%	00.00%	00.61%
	sv	00.00%	00.00%	00.28%	00.00%	00.00%	00.00%	00.14%	97.21%	00.00%	00.70%
	x	00.00%	00.00%	10.64%	00.00%	00.00%	00.00%	00.00%	72.34%	00.00%	00.00%
tr	***aa***	00.00%	00.00%	01.92%	00.00%	00.00%	00.00%	00.48%	62.98%	00.00%	19.71%
	b	00.00%	00.00%	03.01%	00.00%	00.00%	00.00%	00.39%	69.15%	00.00%	18.95%
	ba	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	01.32%	78.95%	00.00%	02.63%
	fc	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	69.14%	00.00%	12.35%
	ny	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	38.36%	00.00%	36.99%
	qw	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	90.91%	00.00%	01.82%
	qy	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	02.38%	91.67%	00.00%	01.19%
	sd	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%	01.14%	92.79%	00.00%	00.99%
	sv	00.00%	00.00%	00.14%	00.00%	00.00%	00.00%	00.84%	93.31%	00.00%	00.42%
	x	00.00%	01.06%	34.04%	00.00%	00.00%	00.00%	00.00%	27.66%	00.00%	02.13%

As covered in the section LDAC Confusion Matrices, in the monolingual setting, Lee-Dernoncourt classifier works most successfully for *B*, *F*, and *S* labels, but it fails to classify any of the utterances with label *D* or *Q*. Unfortunately, the results show that, in the experiments involving a translation process, UDAC with Lee-Dernoncourt classifier misclassifies every other label but *S* in more than 50% of the cases for each label. Compared to LDAC using Lee-Dernoncourt classifier, however, UDAC performs somewhat better, as the percentage misclassifications as *S* for each label is lower in most cases, most notable of which is the label *F* in the confusion matrix of the experiment where the target language is Spanish.

Experiments with UDAC using a BiLSTM-CRF classifier yield better results. While all the utterances with label *D* and *Q* are misclassified, the utterances with labels *F* and *S* are classified correctly in more than 80% of the cases and label *B* is predicted correctly in more than 40% of the cases. When translation occurs, though, the accuracy of correct classification for each label except *S* reduces significantly. Unlike the UDAC experiments using Lee-Dernoncourt classifier, not all labels are misclassified as *S*. Yet, neither the label *B* nor the label *F* is has a correct prediction accuracy higher than 25%, and the 0.00% correct prediction percentage persists for labels *D* and *Q*. Compared to LDAC experiments using BiLSTM-CRF classifier, UDAC performs worse, as the result of LDAC experiments show that the models trained with translated datasets achieve higher percentages of correct classification for labels *B* and *F*, as well as having lower percentages of misclassification of utterances as *S*.

Another critical insight as to the performance issues of UDAC with BiLSTM-CRF classifier is gained through the anomalies seen in the overall accuracies observed in MRDA dataset results on Table 4.5. Even though UDAC with BiLSTM-CRF classifier performs poorly on SwDA dataset, in three of the experiments on MRDA dataset, it works within 7% of the accuracy of the monolingual experiment. We found that this irregularity is caused by the label *z* we mentioned in the previous section. Even though in the rest of the MRDA experiments, BiLSTM-CRF classifier failed in learning to classify utterances with label *z*, in those three experiments, it managed to classify them with more than 98% accuracy. As the utterances with label *z* in the test split of the dataset constitutes more than 60% of the utterances, while the classifier has failed in learning to predict the labels of the utterances which have semantic value, the over-

all accuracy that includes the utterances with label *z* makes it seem as if the classifier works successfully. Therefore, upon a more in-depth investigation of the results of the experiments conducted on MRDA dataset, regardless of the high overall accuracies it achieved in some of them, we find that UDAC with BiLSTM-CRF classifier is not successful in differentiating between the semantically significant labels.

The confusion matrices of the UDAC experiments conducted on SwDA dataset are presented in Table 4.12 and Table 4.13. As with the confusion matrices of the LDAC experiments, we demonstrate the confusion matrices for the ten labels with the highest frequency in the test data.

Experiments conducted on SwDA dataset using UDAC with Lee-Dernoncourt classifier, shown in Table 4.12, reveal that the classifier can not handle the noise that is introduced by translation.

As noted above, in the monolingual experiment, utterances with labels *b*, *ba*, *sd* and *sv* are predicted correctly in more than half of the cases, and the utterances with the label *qy* is classified correctly in 34.52% of the cases, which is higher than the percentage of any other misclassification for that label. Yet, the label *aa* is classified incorrectly in 67.31% of the cases, and the labels *fc*, *ny*, *qw* and *x* are never correctly classified.

In the experiments involving a target language, the only labels that are correctly predicted in more than half of the cases are *sd* and *b*, with the exception of Turkish, where the label *b* is misclassified in 58.69% of the cases. Apart from those two labels, the highest percentage of correct classification of a label observed is 38.16%, 45.68% and 34.13% for German, Spanish and Turkish, respectively. Additionally, in the UDAC models involving translation, none of the labels *fc*, *ny*, *qw*, *qy* and *x* are ever correctly predicted. Based on the confusion matrices Table 4.8 and Table 4.12, LDAC and UDAC configurations have comparable performances on SwDA dataset, when they both utilize the Lee-Dernoncourt classifier. However, note that the overall accuracies obtained and displayed on Table 4.4 and Table 4.5 show that LDAC performs considerably better, especially when the target language is Turkish.

The experiments testing UDAC with BiLSTM-CRF classifier on SwDA dataset yielded the worst results. While the monolingual BiLSTM-CRF classifier achieves the cor-

rect classification of five labels in more than 50% of the cases, in the experiments with a target language, none of the labels but *S* is correctly classified at that rate. 7 of the most frequent ten labels are incorrectly classified in all cases, while the correct classification of the remaining two labels *qy* and *x* have accuracies less than 2.5%. Also, almost all of the labels are misclassified as *sd* in more than 50% of the cases, as can be seen from the gray vertical line drawn by the background colors of the cells in the column for *sd* label. Considering the LDAC and UDAC experiments with BiLSTM-CRF classifier on SwDA, LDAC outperforms its alternative significantly, which is visible on the overall accuracies shown in Table 4.4 and Table 4.5.

Regardless of which DA classifier is used, UDAC experiments show that many other labels are incorrectly predicted to be statements. (i.e. *sd* and *sv* in SwDA; *S* in MRDA) Most visible in Table 4.10 and Table 4.13, this is an issue also seen in the confusion matrices of LDAC, and is attributed to the heterogeneity of the datasets as well as the loss of sentence-level information in translation.

The most compelling observation to be made regarding the confusion matrices of UDAC is the failure of BiLSTM-CRF classifier in predicting any label other than statements correctly. As presented by Kumar et al. [4], BiLSTM-CRF classifier has higher accuracy than the Lee-Dernoncourt classifier in the monolingual case. However, when the confusion matrices of UDAC are examined, unlike the results of LDAC experiment, it is seen that this comparative relationship between the accuracies of the classifiers does not hold for UDAC when translation occurs.

We believe that the reason for the translation of data crippling the performance of UDAC with BiLSTM-CRF classifier is rooted in the evaluation method of UDAC. In our experiments with UDAC, since we did not have labeled datasets in the target language with which we may test the trained models, we opted to translate the test split of the original datasets into the target languages using Google Translate, and treated those translated test data as if it is an authentic labeled test data in that target language. Note that, since UDAC works with a single trained model in the original (i.e., source) language of the dataset, during testing, we translated the data back to the source language using the MT method selected by UDAC. Based on these results, however, we hypothesize that translating a dataset twice may cause too much loss of

sentence-level and contextual data. As a result, fine-grained, hierarchical DA classifiers such as BiLSTM-CRF may fail to learn to classify the data correctly as good as Lee-Dernoncourt classifier. This hypothesis is empirically examined in the following Excerpt Analysis subsection, using excerpts from dialogues used to test LDAC and UDAC.

4.2.3 Excerpt Analysis

There is merit in analyzing how the trained models perform on individual dialogue excerpts, just as there is merit in observing the overall performances of the models. This subsection observes a sample excerpt from the testing portion of each dataset, and how the relevant LDAC and UDAC models label the utterances of that excerpt. We start by examining the testing data used to evaluate LDAC experiments, and then we investigate the possible issue with the testing data used in evaluating UDAC models.

Table 4.14 and Table 4.18 show two selected excerpts from MRDA and SwDA datasets in their original language, respectively. Table 4.15 and Table 4.19 display the German translation of the same dialogue excerpts, obtained automatically using Google Translate. Similarly, Table 4.16 and Table 4.20 show the Spanish translations while Table 4.17 and Table 4.21 contains the Turkish translations of the same excerpts. Each table also contains columns, each of which corresponds to a model trained in labeling dialogues on the relevant dataset, and in that particular target language. Those columns present the labels predicted by the relevant model, laid against another column displaying the true labels of each utterance.

The first thing to note here is the imperfections of the translations used for testing the trained models. As native speakers of the German, Spanish, or Turkish languages may observe, the translations contain several errors. One instance of such a mistake is the first utterance in the excerpt taken from SwDA dataset, which is given in Table 4.18. As shown in Table 4.21, it is translated to Turkish as "How long did you play [with it?]", as opposed to the correct translation "How long did you play [the instrument]?". These errors in the Google Translate translations of the testing portion of the datasets may have affected the accuracy of the translated models. As mentioned in the Ex-

periment Setup section above, this possibility was recognized before the experiments were run. Nonetheless, the method was still preferred due to not requiring any expert human effort, which can neither be reused nor automated for other language pairs.

A second point to make is regarding the misclassifications made by the LDAC models. In Table 4.14, some utterances feature a true label of *B* or *F*. However, the utterances are labeled as *S*. This issue becomes severe in Lee-Dernoncourt models trained using a dataset which was translated to German, Spanish, or Turkish. As can be seen in 4.15, Table 4.16, and most visibly in Table 4.17, the Lee-Dernoncourt models regularly misclassify other classes with the label *S*. Similar misclassifications can be observed in the excerpt that was taken from SwDA dataset. Table 4.18 shows instances of the label *b^m* being labeled as *sd*, which is a misclassification that persists even when the models are trained using the translated versions of the dataset. As hypothesized in the previous section, we attribute this behavior to the bias caused by having an unbalanced amount of statement utterances (i.e., *S* for MRDA, *sd* and *sv* for SwDA) as well as the loss of information when the utterances are translated into another target language.

In addition to incorrectly predicting utterances to have a statement label, many of the LDAC models trained on SwDA dataset demonstrate the observation we make in the previous subsection, regarding the misclassification of the question labels. The first utterance in 4.18 is classified correctly only by the BiLSTM-CRF model which was trained without translation and with utterances word order of which were not shuffled. And even that model setting fails to recognize the question when it is trained on a translated version of the dataset, as can be seen in 4.19, 4.20 and 4.21. Part of the issue with failing to recognize the questions may be rooted in the preprocessing step involved in feeding the training data. To use only the word embeddings of words that have no other character but the alphabetical characters, the dots, commas, exclamation marks, and question marks were all removed in the preprocessing step. In reality, most of the word embedding spaces, including the fastText word embeddings used in this research effort, feature embeddings for common punctuation marks such as dots and question marks. Including those fundamental punctuation marks may improve the capability of recognizing a question dramatically, and should be investigated further.

Table 4.14: Dialogue excerpt from test data of MRDA dataset in English, along with how each utterance was labeled by the models trained in relevant experiments. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	LD-o	LD-s	CRF-o	CRF-s
519	*what worked best is the hand labeled data .*	S	S	S	S	S
520	*uhhuh .*	B	B	B	B	B
521	*um ==*	F	F	F	F	F
522	*uh - so yeah .*	F	F	S	S	S
523	*i don't know if we can get some hand labeled data from other languages .*	S	S	S	S	S
524	*yeah .*	B	B	S	S	S
525	*it's not so easy to find .*	S	S	S	S	S
526	*right .*	B	S	S	S	S
527	*but that would be something interesting t- - to - to see .*	S	S	S	S	S
528	*yeah .*	B	B	B	B	B
529	*yeah .*	B	B	B	B	B
530	*yeah .*	B	B	B	S	S
531	*also uh - | i mean there was just the whole notion of having multiple nets that were trained on different data*	F | S	F | S	F | S	F | S	F | S

Table 4.15: German translation (via Google Translate) of a dialogue excerpt from test data of MRDA dataset, along with how each utterance was labeled by the models trained in relevant experiments. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	LDAC LD-o	LDAC LD-s	LDAC CRF-o	LDAC CRF-s
519	*Was am besten funktioniert hat, sind die handbeschrifteten Daten.*	S	S	S	S	S
520	*uhhuh.*	B	S	S	F	B
521	*eine Eins*	F	S	S	S	S
522	*äh - also ja.*	F	S	S	S	S
523	*Ich weiß nicht, ob wir handbeschriftete Daten aus anderen Sprachen erhalten können.*	S	S	S	S	S
524	*ja*	B	S	S	B	B
525	*es ist nicht so leicht zu finden.*	S	S	S	S	S
526	*Recht .*	B	B	B	S	S
527	*aber das wäre etwas interessantes zu sehen.*	S	S	S	S	S
528	*ja*	B	S	S	B	F
529	*ja*	B	S	S	B	F
530	*ja*	B	S	S	F	F
531	*auch äh | Ich meine, es gab nur den ganzen Gedanken, mehrere Netze zu haben, die auf unterschiedlichen Daten trainiert wurden.*	F | S	S | S	S | S	S | S	F | S

Table 4.16: Spanish translation (via Google Translate) of a dialogue excerpt from test data of MRDA dataset, along with how each utterance was labeled by the models trained in relevant experiments. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	LDAC LD-o	LDAC LD-s	LDAC CRF-o	LDAC CRF-s
519	*Lo que funcionó mejor son los datos etiquetados a mano.*	S	S	S	S	S
520	*uhhuh*	B	B	B	F	F
521	*un ==*	F	S	S	F	F
522	*uh - así que sí.*	F	S	F	S	S
523	*No sé si podemos obtener algunos datos etiquetados a mano de otros idiomas.*	S	S	S	S	S
524	*si*	B	S	S	B	F
525	*No es tan fácil de encontrar.*	S	S	S	S	S
526	*derecho .*	B	F	S	S	S
527	*pero eso sería algo interesante para ver.*	S	S	S	S	S
528	*si*	B	S	S	B	B
529	*si*	B	S	S	B	B
530	*si*	B	S	S	F	F
531	*también uh | Quiero decir que solo existía la noción de tener múltiples redes que fueron entrenadas en diferentes datos.*	F | S	F | S	F | S	F | S	F | S

Table 4.17: Turkish translation (via Google Translate) of a dialogue excerpt from test data of MRDA dataset, along with how each utterance was labeled by the models trained in relevant experiments. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	LDAC LD-o	LDAC LD-s	LDAC CRF-o	LDAC CRF-s
519	*En iyi sonuç veren el etiketli verilerdir.*	S	S	S	S	S
520	*paparazzi.*	B	S	S	F	F
521	*bir tane*	F	S	S	S	S
522	*eh - evet.*	F	S	S	S	S
523	*Diğer dillerden bazı el etiketli veriler alabilir miyiz bilmiyorum.*	S	S	S	S	S
524	*evet.*	B	S	S	B	F
525	*Bulması o kadar kolay değil.*	S	S	S	S	S
526	*sağ .*	B	B	B	S	S
527	*ama bu görmek için ilginç bir şey olurdu.*	S	S	S	S	S
528	*evet.*	B	S	S	B	B
529	*evet.*	B	S	S	B	B
530	*evet.*	B	S	S	B	F
531	*ayrıca - | demek istediğim, sadece farklı veriler üzerinde eğitilmiş birden fazla ağa sahip olma kavramı vardı.*	F | S	S | S	S | S	F | S	F | S

Table 4.18: Dialogue excerpt from test data of SwDA dataset in English, along with how each utterance was labeled by the models trained in relevant experiments. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	LD-o	LD-s	CRF-o	CRF-s
41	*So, how long did you play?*	qw	sd	sd	qw	+
42	*Only for about three months.*	sd	sd	sd	sd	sd
43	*Three months.*	b^m	sd	sd	sd	sd
44	*Yeah,*	b	b	b	b	b
45	*me and my brother both took the classes*	sd	sd	sd	sd	sd
46	*and we got pretty bored quick <laughter>.*	sd	sd	sd	sd	sd
47	*I was going to say, y-, y-, you got as far as the, uh, chop sticks, huh.*	sd	sd	sd	sd	sd
48	*Um, well, I could play, uh, the wood chuck song.*	sd	sd	sd	sd	sd
49	*Oh, the wood chuck song <laughter>.*	b^m	sd	sd	sd	sd
50	*And I still can to this date.*	sd	sd	sd	sd	sd
51	*<Laughter>.*	x	b	b	x	x
52	*<Laughter>.*	x	b	b	x	x

Table 4.19: German translation (via Google Translate) of a dialogue excerpt from test data of SwDA dataset, along with how each utterance was labeled by the models trained in relevant experiments. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	LDAC LD-o	LDAC LD-s	LDAC CRF-o	LDAC CRF-s
41	*Also, wie lange hast du gespielt?*	qw	sv	sv	+	+
42	*Nur für etwa drei Monate.*	sd	sd	sd	sd	sd
43	*Drei Monate.*	b^m	sd	sd	sd	sd
44	*Ja,*	b	b	b	x	ba
45	*Ich und mein Bruder nahmen beide am Unterricht teil*	sd	sd	sd	sd	sd
46	*und wir haben uns ziemlich schnell gelangweilt <laughter>.*	sd	sd	sd	sd	sd
47	*Ich wollte sagen, y-, y-, du bist so weit wie die Hackenstäbchen gekommen, hm.*	sd	sd	sd	sd	sd
48	*Ähm, ich könnte den Holz-Chuck-Song spielen.*	sd	sd	sd	sd	sd
49	*Oh, das Holzlied <laughter>.*	b^m	b	ba	ba	ba
50	*Und ich kann es bis heute noch.*	sd	sd	sd	sd	sd
51	*<Laughter>.*	x	b	b	x	x
52	*<Laughter>.*	x	b	b	x	x

Table 4.20: Spanish translation (via Google Translate) of a dialogue excerpt from test data of SwDA dataset, along with how each utterance was labeled by the models trained in relevant experiments. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	LDAC LD-o	LDAC LD-s	LDAC CRF-o	LDAC CRF-s
41	*Entonces, ¿cuánto tiempo jugaste?*	qw	sd	sd	sd	+
42	*Sólo por unos tres meses.*	sd	sd	sd	sd	sd
43	*Tres meses.*	b^m	sd	sd	sd	sd
44	*Sí,*	b	b	b	x	sd
45	*mi hermano y yo tomamos las clases*	sd	sd	sd	sd	sd
46	*y nos aburrimos bastante rapido <laughter>.*	sd	sd	sv	sd	sd
47	*Iba a decir, y-, y-, llegaste hasta los palos, eh.*	sd	sd	sd	sd	sd
48	*Um, bueno, podría tocar la canción de madera.*	sd	sd	sd	sd	sd
49	*Oh, el canto de madera <laughter>.*	b^m	sd	sd	sd	qy
50	*Y todavía puedo hasta esta fecha.*	sd	sd	sd	sd	qy
51	*<Laughter>.*	x	b	b	x	x
52	*<Laughter>.*	x	b	b	x	x

Table 4.21: Turkish translation (via Google Translate) of a dialogue excerpt from test data of SwDA dataset, along with how each utterance was labeled by the models trained in relevant experiments. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	LDAC LD-o	LDAC LD-s	LDAC CRF-o	LDAC CRF-s
41	*Ne zamandır oynadın?*	qw	sd	sd	+	+
42	*Sadece yaklaşık üç ay.*	sd	sd	sd	sd	sd
43	*Üç ay.*	b^m	+	sd	sd	sd
44	*Evet,*	b	aa	b	b	b
45	*ben ve erkek kardeşim ikimiz de dersleri aldık.*	sd	sd	sd	sd	sd
46	*ve çok çabuk sıkıldık <laughter>.*	sd	sd	sd	sd	sd
47	*Söyleyecektim, ee, pirzola çubukları kadarıyla aldın, ha.*	sd	sd	sd	sd	sd
48	*Şey, ben de, ağaçlıklı şarkıyı çalabilirim.*	sd	sd	sd	sd	sd
49	*Ah, ağaç aynası şarkısı <laughter>.*	b^m	b	%	sd	sd
50	*Ve hala bu tarihe kadar yapabilirim.*	sd	sd	sd	sd	sd
51	*<Laughter>.*	x	b	b	x	x
52	*<Laughter>.*	x	b	b	x	x

As part of the excerpt analysis, the process of translating the data twice should be investigated. As explained in Experiment Setup section above, translating twice was essential to evaluate the performance of the UDAC model, but the overall results obtained and displayed in Table 4.5 show that this evaluation method should be analyzed.

The MT method used as part of UDAC translates each word separately, and any word for which a translation could not be found is extracted from the utterance. As a result, while evaluating the trained models, the DA classifier of UDAC is given dialogues with utterances which are not grammatically proper sentences but contain contextually related information. The MT method used in UDAC is utilized to translate the excerpts from MRDA test data given in Table 4.15, Table 4.16 and Table 4.17 from German, Spanish and Turkish, respectively, back to English so that the test data can be inputted to the DA classifier. Table 4.22, Table 4.23 and Table 4.24 demonstrate the translated excerpts of the data before it is given to the DA classifier as input. The original version (i.e., before any translation takes place) of the excerpts featured in all three tables is the same and is provided above in Table 4.14. Similarly, Table 4.25, Table 4.26 and Table 4.27 feature the translated versions of the excerpts provided in Table 4.19, Table 4.20 and Table 4.21, respectively. The initial English version of the excerpt is given in Table 4.18.

The predictions featured in the UDAC tables show a parallel with the analysis conducted on the confusion matrices. In most cases, every trained model misclassifies other labels as *S* in MRDA and *sd* in SwDA. Still, the vital insight about the evaluation method comes not from the predictions featured in the tables, but the translations. We observed above that translating once caused some loss of grammatical and contextual information. The tables reflecting the UDAC testing data, however, show an even more significant loss of information after the second translation. For instance, as can be seen from tables 4.18 and 4.27 the 49^{th} utterance from the original excerpt, which was initially "Oh the wood chuck song <laughter>.", became "yah" after being translated to Turkish by Google Translate, and back to English by the MT method of UDAC. This and similar examples from the excerpt tables show the severity of the loss of contextual information. As UDAC trains a single DA classifier in a single source language (i.e., English) using training data that is much richer in contextual

and grammatical information, it is natural that the classifications made by UDAC suffer due to the loss of most of the information included in the utterances after two translations.

The most critical information loss occurs when translating the dataset from the target language back to the source language. The MT method used in UDAC fails to find some of the words in the test data within the monolingual word embedding space of the target language. The reason for this is the fact that Google Translate and the fastText word embeddings are trained on different source data. As fastText monolingual word embedding spaces are limited (i.e., contains fewer words than used by Google Translate) some of the words are not found. As shown in Table 4.28, the percentage of words found in the word embedding space while testing UDAC is lower than that of LDAC, in almost all of the cases.

Combined with the empirical analysis of the dialogue excerpts, results obtained support the argument that the current evaluation method of UDAC is not sufficiently reliable, and should be improved. Currently, due to the relatively limited coverage of the monolingual word embedding spaces, the resulting utterances lose too much contextual information. Therefore, despite the acceptable accuracies obtained with Lee-Dernoncourt classifier, the results obtained while testing UDAC should be considered as inconclusive.

4.2.4 Comparison with Utterance-based Translation

Throughout our examination of the experiment results, we noted the substantial loss of grammatical and contextual information. This loss is partly expected when an MT method is used. Nonetheless, the usage of a word-based translation method may be a factor that increased the loss. To investigate how a word-based MT method affects the accuracy of our solutions, we conducted additional experiments with LDAC using an utterance-based MT method. Specifically, we used Google Translate to conduct utterance-based translation of the entire MRDA dataset. Then we used these translated versions of the dataset as training data, as opposed to the word-based translation LDAC used in the rest of the experiments. We used both of the DA classifiers used in the rest of the experiments and tried to compare the utterance-based and word-based

Table 4.22: English translation of an MRDA dialogue excerpt obtained using the MT method of UDAC, from the German translation (via Google Translate) of the original dialogue. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	UDAC LD-o	UDAC LD-s	UDAC CRF-o	UDAC CRF-s
519	*"nothing on best disable has are the data"*	S	S	S	S	S
520	*""*	B	F	D	F	F
521	*"a peaked"*	F	S	S	S	B
522	*"hence thats"*	F	S	S	S	S
523	*"i neither whether we data from other languages receive"*	S	S	S	S	S
524	*"thats"*	B	S	S	F	F
525	*"there is neither so slightly to find"*	S	S	S	S	S
526	*"law"*	B	S	S	S	S
527	*"but ese slightly interesting to seen"*	S	S	S	S	S
528	*"thats"*	B	S	S	B	F
529	*"thats"*	B	S	S	B	B
530	*"thats"*	B	S	S	F	F
531	*"also"* | *"i my there had only both whole ideas several networks to have the on different data coached were"*	F | S	F | S	D | S	S | S	S | S

Table 4.23: English translation of an MRDA dialogue excerpt obtained using the MT method of UDAC, from the Spanish translation (via Google Translate) of the original dialogue. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	UDAC LD-o	UDAC LD-s	UDAC CRF-o	UDAC CRF-s
519	*"that that best are parians data labeled to hand"*	S	S	S	S	S
520	*""*	B	F	D	F	F
521	*"a"*	F	S	S	S	S
522	*"uh that"*	F	F	F	S	S
523	*"neither if can obtain many data labeled to hand of other languages"*	S	S	S	S	S
524	*"if"*	B	S	S	F	F
525	*"neither is exceedingly of find"*	S	S	S	S	S
526	*"law"*	B	S	S	S	S
527	*"but really something interesting for see"*	S	S	S	S	S
528	*"if"*	B	S	S	F	F
529	*"if"*	B	S	S	B	B
530	*"if"*	B	S	S	F	F
531	*"uh"* | *"me mean that only una of give networks that were trained in different data"*	F | S	F | S	F | S	S | S	S | S

Table 4.24: English translation of an MRDA dialogue excerpt obtained using the MT method of UDAC, from the Turkish translation (via Google Translate) of the original dialogue. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	UDAC LD-o	UDAC LD-s	UDAC CRF-o	UDAC CRF-s
519	*"most best giving el labels tabulates"*	S	S	S	S	S
520	*"gaga"*	B	S	S	F	F
521	*"a two"*	F	S	S	S	B
522	*"nyaah yes"*	F	S	S	S	S
523	*"languages el labels data obtains diffrent anyway"*	S	S	S	S	S
524	*"yes"*	B	S	S	F	F
525	*"o until easier"*	S	S	S	S	S
526	*""*	B	F	D	S	S
527	*"but which a should"*	S	S	S	S	S
528	*"yes"*	B	S	S	F	B
529	*"yes"*	B	S	S	B	B
530	*"yes"*	B	S	S	F	F
531	*""* | *"say only data multiple fewer possesses becoming"*	F | S	F | S	D | S	S | S	S | S

Table 4.25: English translation of an SwDA dialogue excerpt obtained using the MT method of UDAC, from the German translation (via Google Translate) of the original dialogue. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	UDAC LD-o	UDAC LD-s	UDAC CRF-o	UDAC CRF-s
41	*"hence like long you respondant played"*	qw	sd	sd	sd	sd
42	*"only approximately four months"*	sd	sd	sd	sd	sd
43	*"four months"*	b^m	sd	sd	sd	sd
44	*"thats"*	b	b	b	x	sd
45	*"i and my brother took both on lessons part"*	sd	sd	sd	sd	sd
46	*"and we have yourselves quite quickly bored"*	sd	sd	sd	sd	sd
47	*"i wanted say y y respondant me so far like the brought hmmm"*	sd	sd	sd	sd	sd
48	*"i both games"*	sd	sd	sd	sd	sd
49	*"oh ese <laughter>"*	b^m	b	b	sd	sd
50	*"and i can there until today still"*	sd	sd	sd	sd	sd
51	*""*	x	b	b	sd	x
52	*""*	x	b	b	ba	+

Table 4.26: English translation of an SwDA dialogue excerpt obtained using the MT method of UDAC, from the Spanish translation (via Google Translate) of the original dialogue. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	UDAC LD-o	UDAC LD-s	UDAC CRF-o	UDAC CRF-s
41	*"then time love/you"*	qw	+	sd	sd	sd
42	*"by hundred three months"*	sd	+	sd	sd	sd
43	*"three months"*	b^m	sd	sd	sd	sd
44	*""*	b	b	b	sd	sd
45	*"my brother and me then, these classes"*	sd	sd	sd	sd	sd
46	*"and we awful very"*	sd	sv	sv	sd	sd
47	*"went to mean and and you/i until parians horseshoes eh"*	sd	+	sd	sd	sd
48	*"um good ukulele una of wooden"*	sd	sd	sd	sd	sd
49	*"oh the singing of wooden"*	b^m	sd	sd	sd	sd
50	*"and you until itself date"*	sd	sd	sd	sd	sd
51	*""*	x	b	b	ba	x
52	*""*	x	b	b	+	+

Table 4.27: English translation of an SwDA dialogue excerpt obtained using the MT method of UDAC, from the Turkish translation (via Google Translate) of the original dialogue. TL denotes True Label, LD-o denotes Lee-Dernoncourt (ordered), LD-s denotes Lee-Dernoncourt (shuffled), CRF-o denotes BiLSTM-CRF (ordered) and CRF-s denotes BiLSTM-CRF (shuffled).

Utt Num	Utterance	TL	UDAC LD-o	UDAC LD-s	UDAC CRF-o	UDAC CRF-s
41	*"ne"*	qw	sd	sd	sd	sd
42	*"only months"*	sd	sd	sd	sd	sd
43	*"months"*	b^m	+	sd	sd	sd
44	*"yes"*	b	aa	aa	sd	sd
45	*"say and male everybody de taught"*	sd	sd	sd	sd	sd
46	*"and"*	sd	sd	sd	sd	sd
47	*"ee marinated ha"*	sd	sd	sd	sd	sd
48	*"say de"*	sd	sd	sd	sd	sd
49	*"yah "*	b^m	b	b	sd	sd
50	*"and still which history an until thats"*	sd	sv	sv	sd	sd
51	*""*	x	b	b	ba	x
52	*""*	x	b	b	+	+

Table 4.28: Ratio of words in the testing data which are found in the corresponding monolingual word embedding spaces

	MRDA		**SwDA**	
	LDAC	***UDAC***	***LDAC***	***UDAC***
de	72.48%	49.11%	47.41%	61.29%
es	74.23%	52.68%	76.18%	61.95%
tr	36.30%	13.86%	40.78%	19.39%

Table 4.29: Comparison of utterance-based and word-based translation methods on MRDA dataset, with Turkish as target language

	Lee-Dernoncourt		**BiLSTM-CRF**	
	ordered	***shuffled***	***ordered***	***shuffled***
Word-based	**60.35%**	60.01%	**86.68%**	85.29%
Utterance-based	64.19%	**64.97%**	**86.31%**	85.01%

translation approaches with a single target language, namely Turkish.

Table 4.29 shows the results of the experiments conducted. The results show that the accuracy of LDAC models increase more than 4% with utterance-based translation when LDAC uses Lee-Dernoncourt classifier. On the other hand, when BiLSTM-CRF model is used as the DA classifier, the accuracy of LDAC is higher with the word-based translation method. Based on these results, we can not definitively state that an utterance-based translation method is a better fit for LDAC and UDAC.

CHAPTER 5

CONCLUSION

In this thesis, we presented a problem faced commonly in many NLP tasks. Many researchers trying to devise a solution to any NLP task focusing on a language other than a select few, such as English or Spanish, face the challenge of not having enough textual data that they can use to devise a solution. Considering that lack of unlabeled data can be remedied in a relatively straightforward manner with the help of recent advents such as the Internet, we focused specifically on the DA classification task, a task that requires data labeled by human experts.

There are two simple solutions to the lack of language-specific data in DA classification tasks. The first method consists of compiling a new dataset in a specific language and getting it labeled manually by human experts. The second method a researcher can use is to get a dataset available in a language translated to the target language by a human expert. Due to needing to be replicated for each specific language to be studied, neither of these efforts can be automated, as the human effort needs to be replicated for each new language in both techniques.

We chose to focus on the DA classification task and offered two DA classification solutions that are making use of MT methods so that the need for large amounts of labeled data can be eliminated when implementing a DA classification solution in a new target language. Our first solution, called Localized Dialogue Act Classifier (LDAC), is based on translating the dataset in a source language to the target language in an automated manner so that the DA classification can be learned from the translated dataset. The second solution, named Universal Dialogue Act Classifier (UDAC), trains a single DA classifier in the source language of the dataset being used and automatically translates existing dialogues in target languages into the source

language so that the trained DA classifier can predict the labels of the utterances.

To analyze how our accurately solutions perform, we tested LDAC and UDAC with two existing solutions to DA classification problem and a partially automated MT method. We experimented with two datasets SwDA and MRDA, which are frequently used in DA classification. We used German, Spanish, and Turkish as the target languages with which LDAC and UDAC are tested. We also investigated how the word order within an utterance affects the final accuracy achieved in the DA classification by training a duplicate of each setting with which we experimented, where the word order of each utterance was shuffled randomly.

Presenting the results we obtained, we showed that the accuracies obtained by both LDAC and UDAC are worse than their monolingual counterparts where no translation is conducted. We hypothesized possible causes, including loss of grammatical and contextual information upon translation and the biased, heterogeneous structure of the datasets in terms of the label frequencies. We also argued that the testing methodology for UDAC is not reliable and should be improved.

Our proposed solution performed considerably worse than the state-of-the-art in monolingual DA classification solutions in most cases. However, with MRDA dataset, when a state-of-the-art DA classifier by Kumar et al. [4] is used, the accuracies of LDAC solution in all the target languages is better than the monolingual DA classification solution proposed by Lee and Dernoncourt [3]. We conclude that this is an indication of how promising our approach is, and state that this technique should be explored further.

5.1 Future Work

Considering the issues of the solutions as well as the possible uses of them, the research conducted in this thesis is an initial effort which can be expanded to a whole family of a solution approach.

The first step to take is remedying the possible issues we observed in our solutions, and about which we made hypotheses. One crucial exploration that needs to be done

is searching for a replacement to the Google Translate translations we used in evaluating the accuracy of the trained models. If a bilingual, expert-labeled dialogue dataset is compiled, then it can be used as a more reliable source of testing our translation approach, as opposed to our current method, which we showed to contain contextually and grammatically erroneous translations for utterances. Additionally, further experiments should be conducted with word vector spaces that include a more significant amount of word embeddings to see if the evaluation method of UDAC can be made more reliable.

Another recognized issue is regarding the misclassification of some semantically significant DA labels by both LDAC and UDAC, such as the labels assigned to questions. In Chapter 4, we recognize that using the word embeddings of only the words with solely alphabetic characters in them may not be the best way to translate the utterance. We consider the idea of including the word embeddings for fundamental punctuation marks such as question marks and dots, embeddings for which exist in many pre-trained word embedding spaces. As a result, this modification should be tested on both LDAC and UDAC as well.

A different category of future work involves analyzing how the individual components used affect the overall approach. Experimenting with different MT methods and DA classification solutions are possible works that can be included in this effort. For instance, an alternative word-based architecture can be achieved by using the linear transformation to map the monolingual word embedding spaces of the target languages to the vector space of the source language. Then, the projected word embeddings of the words in a target language can be used directly, as opposed to finding a semantically closest word in the source language. This approach may alleviate some of the accuracy loss caused by the failure in finding a relevant translation.

Naturally, training LDAC and UDAC to work on various other target languages is also worth examining. An analysis of how LDAC and UDAC perform when as the target language becomes linguistically less similar to the source language (e.g., uses a different alphabet) is a particularly intriguing exploration to be made.

Finally, once LDAC and UDAC are well-studied within the boundaries of the DA classification problem, they can be modified and expanded to be used in other NLP prob-

lems. If further studied and proven to be useful, their configuration can be adopted by any NLP task, where the goal is to learn a solution in a particular language, but the sufficient data or dataset to learn that solution in that target language does not exist.

REFERENCES

[1] C. Bothe, C. Weber, S. Magg, and S. Wermter, "A context-based approach for dialogue act recognition using simple recurrent neural networks," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, (Miyazaki, Japan), European Languages Resources Association (ELRA), May 2018.

[2] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *ArXiv*, vol. abs/1309.4168, 2013.

[3] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 515–520, Association for Computational Linguistics, June 2016.

[4] H. Kumar, A. Agarwal, R. Dasgupta, S. Joshi, and A. Kumar, "Dialogue act sequence labeling using hierarchical encoder with crf," in *AAAI*, 2018.

[5] K. Papineni, S. Roukos, T. Ward, and W.-j. Zhu, "Bleu: a method for automatic evaluation of machine translation," 10 2002.

[6] M. Mast, H. Niemann, E. Nöth, and E. G. Schukat-Talamazzini, "Automatic classification of dialog acts with semantic classification trees and polygrams," in *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, (London, UK, UK), pp. 217–229, Springer-Verlag, 1996.

[7] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *EUROSPEECH*, 1997.

[8] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic deter-

mination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, pp. 100–107, July 1968.

[9] N. Reithinger and R. Engel, "Robust content extraction for translation and dialog processing," 01 2000.

[10] N. Reithinger and M. Rumpler, ""dialogue act classification using language models."," 01 1997.

[11] W. S. Choi, J.-M. Cho, and J. Seo, "Analysis system of speech acts and discourse structures using maximum entropy model," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, (Stroudsburg, PA, USA), pp. 230–237, Association for Computational Linguistics, 1999.

[12] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," *Computational Linguistics*, vol. 21, pp. 543–565, 1995.

[13] K. Samuel, S. Carberry, and K. Vijay-Shanker, "Computing dialogue acts from features with transformation-based learning," 07 1998.

[14] M. Kipp, "The neural path to dialogue acts," pp. 175–179, 01 1998.

[15] E. Shriberg, D. Jurafsky, and K. Ries, "Dialog act modeling for conversational speech," 1998.

[16] J. J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," 1992.

[17] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, pp. I/1061–I/1064 Vol. 1, 2005.

[18] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proceedings of the 5th SIGdial*

Workshop on Discourse and Dialogue at HLT-NAACL 2004, (Cambridge, Massachusetts, USA), pp. 97–100, Association for Computational Linguistics, Apr. 30 - May 1 2004.

[19] J. D. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[20] T. Nakagawa, K. Inui, and S. Kurohashi, "Dependency tree-based sentiment classification using crfs with hidden variables.," pp. 786–794, 01 2010.

[21] J. a. Silva, L. Coheur, A. C. Mendes, and A. Wichert, "From symbolic to sub-symbolic information in question classification," *Artif. Intell. Rev.*, vol. 35, pp. 137–154, Feb. 2011.

[22] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," pp. 1–12, 01 2013.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Neurocomputing: Foundations of research," ch. Learning Representations by Back-propagating Errors, pp. 696–699, Cambridge, MA, USA: MIT Press, 1988.

[24] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, Oct. 2014.

[25] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *NEURAL NETWORKS*, pp. 5–6, 2005.

[26] Z. Chen, R. Yang, Z. Zhao, D. Cai, and X. He, "Dialogue act recognition via crf-attentive structured network," pp. 225–234, 06 2018.

[27] B. Krause, L. Lu, I. Murray, and S. Renals, "Multiplicative lstm for sequence modelling," 09 2016.

[28] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical*

Translation, (Doha, Qatar), pp. 103–111, Association for Computational Linguistics, Oct. 2014.

[29] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguist.*, vol. 19, pp. 263–311, June 1993.

[30] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996.

[31] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (Philadelphia, Pennsylvania, USA), pp. 295–302, Association for Computational Linguistics, July 2002.

[32] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, (Sapporo, Japan), pp. 160–167, Association for Computational Linguistics, July 2003.

[33] D. Chiang, "Hierarchical phrase-based translation," *Comput. Linguist.*, vol. 33, pp. 201–228, June 2007.

[34] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Comput. Linguist.*, vol. 30, pp. 417–449, Dec. 2004.

[35] H. Mi and L. Huang, "Forest-based translation rule extraction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, (Stroudsburg, PA, USA), pp. 206–214, Association for Computational Linguistics, 2008.

[36] A. Klementiev, I. Titov, and B. Bhattarai, "Inducing crosslingual distributed representations of words," in *COLING*, 2012.

[37] G. Dinu and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," 12 2014.

[38] M. Artetxe, G. Labaka, and E. 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)*, (Vancouver, Canada), pp. 451–462, Association for Computational Linguistics, July 2017.

[39] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *ArXiv*, vol. abs/1702.03859, 2017.

[40] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *ArXiv*, vol. abs/1710.04087, 2018.

[41] M. Artetxe, G. Labaka, and E. 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)*, (Melbourne, Australia), pp. 789–798, Association for Computational Linguistics, July 2018.

[42] K. Duh, A. Fujino, and M. Nagata, "Is machine translation ripe for cross-lingual sentiment classification?," in *ACL*, 2011.

[43] S. Mohammad, M. Salameh, and S. Kiritchenko, "How translation alters sentiment," *J. Artif. Intell. Res.*, vol. 55, pp. 95–130, 2016.

[44] J. Barnes, P. Lambert, and T. Badia, "Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification," in *COLING*, 2016.

[45] A. Balahur and M. Turchi, "Comparative experiments for multilingual sentiment analysis using machine translation," in *SDAD@ECML/PKDD*, 2012.

[46] C.-H. Lee and H. yi Lee, "Cross-lingual transfer learning for question answering," *ArXiv*, vol. abs/1907.06042, 2019.

[47] J. Martínek, P. Král, L. Lenc, and C. Cerisara, "Multi-lingual dialogue act recognition with deep learning methods," *ArXiv*, vol. abs/1904.05606, 2019.

[48] T. Brychcin, "Linear transformations for cross-lingual semantic textual similarity," *ArXiv*, vol. abs/1807.04172, 2018.

[49] S. Jekat, A. L. Klein, E. Maier, M. Mast, and B. Schmitz, "Dialogue acts in verbmobil," 1995.

[50] N. Samet, S. Hiçsönmez, P. Duygulu, and E. Akbas, "Could we create a training set for image captioning using automatic translation?," pp. 1–4, 05 2017.

[51] N. Hammerla, "fasttext multilingual." `http://www.mathworks.com/videos/optimization-modeling-2-converting-to-solver-form-101560.html`, Jun 2017. Online, accessed 17 August 2019.

[52] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[53] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual," Tech. Rep. Draft 13, University of Colorado, Institute of Cognitive Science, 1997.

[54] M. Core and J. F. Allen, "Coding dialogs with the damsl annotation scheme," 01 2001.

[55] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," pp. 364–367, 2003.

[56] F. Dernoncourt, "Data splits for the naacl 2016 paper." `https://github.com/Franck-Dernoncourt/naacl2016`, Mar 2016. Online, accessed 17 August 2019.

[57] M. D. Zeiler, "Adadelta: An adaptive learning rate method," vol. 1212, 12 2012.

[58] Python Software Foundation, "Python.org." `https://www.python.org`. Online, accessed 18 August 2019.

[59] "Keras documentation." `https://keras.io`. Online, accessed 18 August 2019.

[60] "Tensorflow." https://www.tensorflow.org. Online, accessed 18 August 2019.

[61] "Numpy." https://www.numpy.org. Online, accessed 18 August 2019.

[62] Facebook Inc., "fasttext." https://fasttext.cc. Online, accessed 18 August 2019.