

IDENTIFICATION OF DISCOURSE RELATIONS IN TURKISH DISCOURSE BANK

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

FERHAT KUTLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

JANUARY 2023

Identification of Discourse Relations in Turkish Discourse Bank

submitted by **FERHAT KUTLU** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Banu GÜNEL KILIÇ
Dean, **Graduate School of Informatics**

Dr. Ceyhan TEMÜRCÜ
Head of Department, **Cognitive Science**

Prof. Dr. Deniz ZEYREK BOZŞAHİN
Supervisor, **Cognitive Science**

Dr. Murathan KURFALI
Co-supervisor, **Linguistics Department, Stockholm University**

Examining Committee Members:

Assoc. Prof. Dr. Barbaros YET
Cognitive Science, METU

Prof. Dr. Deniz ZEYREK BOZŞAHİN
Cognitive Science, METU

Prof. Dr. İsmail Sengör ALTINGÖVDE
Computer Engineering, METU

Assoc. Prof. Dr. Savaş YILDIRIM
Computer Engineering, Bilgi University

Prof. Dr. Olcay Taner YILDIZ
Computer Engineering, Özyeğin University

Date: 25.01.2023

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: Ferhat KUTLU

Signature :

ABSTRACT

IDENTIFICATION OF DISCOURSE RELATIONS IN TURKISH DISCOURSE BANK

KUTLU, Ferhat

Ph.D., Department of Cognitive Science

Supervisor: Prof. Dr. Deniz ZEYREK BOZŞAHİN

Co-Supervisor: Dr. Murathan KURFALI

January 2023, 87 pages

Discourse is the level of language where linguistic units are organized in a structured and coherent way. One of the major problems in the field of discourse in particular, and NLU in general is how to build better models to sense the way constitutive units of discourse stick together to form a coherent whole. The discourse would be coherent if it had meaningful connections between its parts. Discourse relations, i.e., semantic or pragmatic relations between discourse units (clauses or sentences), are one of the most important aspects of discourse structure. Discourse relations can be realized explicitly (i.e. through connectives), or without them, known as implicit relations. The task that automatically reveals these aspects of texts has been known as ‘discourse parsing’, and in the last two decades, the problem has turned into how to make machines a better discourse detector. Most of the existing studies target the automatic extraction of discourse structure by detecting explicit and implicit relations and the constitutive parts of the relation (i.e., arguments). Focusing on a relatively less studied language, Turkish, this thesis is designated to reveal its discourse structure by focusing on two sub-tasks of shallow discourse parsing, namely, identification of discourse relation realization types and the sense classification of explicit and implicit relations. In this way, a better model which learns discourse structure in a supervised fashion is searched. Such models have been highly needed in the enhancement of tasks such as text summarization, dialogue systems and machine translation that need information above the clause level. Working on Turkish Discourse Bank 1.2, the thesis develops the most thorough pipeline towards shallow discourse parsing. The series of experiments starts with a classification model based on linguistic features fed into legacy machine learning algorithms and ends with fine-tuning a pre-trained language model as an encoder and classifying the encoded data with neural network-based classifiers. Expressed in terms of F1-Scores, this effort has resulted in: (i) an increase from 0.36 to 0.77 in classifying discourse relation realization types, (ii) achieved 0.82 in the classification of the Level-1

senses of explicit relations and 0.54 of implicit relations. The Level-2 Senses of discourse relations are so many that it becomes impossible to end up with a sound classification performance by training with the less number of samples available in the discourse bank. Thus, the study of Level-2 Senses is left to future works, potentially supported with bigger size of discourse bank. We further explore the effect of multilingual data aggregation on the classification of discourse relation realization type through Cross-lingual Transfer Learning experiments practiced with the advantage of the BERT multilingual base model (cased) with Turkish, Chinese and English datasets. We believe that the findings are important both in providing insights regarding the performance of modern language models in Turkish and in the low-resource scenario.

Keywords: Discourse Relation, Classification, Pre-trained Language Model, Encoding, Cross-lingual Transfer Learning

ÖZ

TÜRKÇE SÖYLEM BANKASINDA SÖYLEM BAĞINTILARININ BELİRLLENMESİ

KUTLU, Ferhat

Doktora, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Prof. Dr. Deniz ZEYREK BOZŞAHİN

Ortak Tez Yöneticisi: Dr. Murathan KURFALI

Ocak 2023, 87 sayfa

Söylem, dilbilimsel birimlerin yapılandırılmış ve tutarlı bir şekilde düzenlendiği dil düzeyidir. Özellikle, söylem alanındaki ve genel olarak makinenin doğal dili anlamasındaki en büyük sorunlardan biri, söylemin kurucu birimlerinin tutarlı bir bütün oluşturan yapısını algılamaya yönelik daha iyi modellerin nasıl inşa edileceğidir. Eğer parçaları arasında anlamlı bağlantılar varsa, söylem tutarlı olacaktır. Söylem bağintıları, yani söylem birimleri (tümceler veya tümcecikler) arasındaki anlamsal veya edim bilimsel ilişkiler, söylem yapısının en önemli yönlerinden biridir. Söylem bağintıları, açık bir şekilde (yani bağlayıcılar aracılığıyla) veya bunlar olmadan algılanabilen örtük bağintılar olarak gerçekleştirilebilirler. Metinlerin bu yönlerini otomatik olarak ortaya çıkaran görev "söylem ayrıştırma" olarak bilinmekte olup son yirmi yılın çalışmaları, makinelerin nasıl daha iyi bir söylem algılayıcısı haline getirileceği konusuna odaklanmıştır. Mevcut çalışmaların çoğu, açık ve örtük bağintıları ve bağintının kurucu kısımlarını (yani üyelerini) tespit ederek söylem yapısının otomatik olarak çıkarılmasını hedefler. Nispeten daha az çalışılan bir dil olan Türkçe'ye odaklanan bu tez çalışması ise, sık söylem ayrıştırması yönteminin iki alt görevine, yani söylem bağintısı gerçekleştirme türlerinin ayrıştırılması ile açık ve örtük sınıflarının 1. Seviye anlamlarının sınıflandırmasına odaklanarak söylem yapısını tespit etmeyi amaçlamıştır. Böylece denetimli bir şekilde söylem yapısını öğrenebilen daha iyi bir modelin geliştirilmesi amaçlanmıştır. Bu tür modellere, cümle seviyesinin üzerinde bilgi gerektiren metin özetleme, diyalog sistemleri ve makine çevirisi gibi görevlerin geliştirilmesinde oldukça ihtiyaç duyulmaktadır. Türkçe Söylem Bankası 1.2 versiyonu üzerinde gerçekleştirilen tez çalışması, sık söylem çözümlemesine yönelik mevcut teknoloji ile olabilecek en yüksek faydayı sağlayan bir sistemin bileşenlerini hayata geçirmeye yöneliktir. Dilbilimsel özelliklerden faydalanılarak çıkarılan verinin, klasik makine öğrenimi algoritmalarıyla sınıflandırılmasına yönelik model geliştirilmesiyle başlayan bu tez çalışması, önceden eğitilmiş bir dil modelinin göreve yönelik tadil edilmesi ve sayısallaştırılmış verinin sinir ağı tabanlı sınıflandırıcılarla ayrıştırılabilmesi ile sona ermiştir. Sınıflandırma deney

sonuçlarını F1-Puanları cinsinden ifade edersek, tez çalışmasında geliştirilen modeller: (i) söylem bağıntısı gerçekleşme tiplerini 0,36'dan başlayıp 0,77'ye yükselen bir başarı ile sınıflandırabilmiş, (ii) açık ve örtük sınıflarının 1. Seviye anlamları için sırasıyla 0,82 ve 0,54 başarı ile sınıflandırabilmiştir. Söylem bağıntısı tiplerinin 2. Seviye anlamlarının sınıflandırılması gereken kategori sayısını yüksek bir düzeye çıkardığı için, Türkçe Söylem Bankasında bulunan işaretleme sayısı ile sağlıklı bir sınıflandırma performansı elde etmenin imkânsız olduğu görülmüştür. Çalışmada son olarak, değişik dillerin veri kümeleri birleştirilerek söylem bağıntısı türlerinin sınıflandırılması üzerindeki etkisi araştırılmış, bu amaçla Türkçe, Çince ve İngilizce veri kümelerinin BERT (büyük küçük harf duyarlı) çok dilli temel modeli ile Diller Arası Transfer Öğrenme deneyleri gerçekleştirilmiştir. Bulguların, modern dil modellerinin Türkçe gibi az kaynaklı diller üzerinde yapılacak çalışmaların performansına etkilerine ilişkin fikir vermesi açısından önemli olduğu değerlendirilmektedir.

Anahtar Kelimeler: Söylem Bağıntısı, Ön Eğitimli Dil Modeli, Sayısallaştırma, Diller Arası Aktarımlı Öğrenme

To My Family

ACKNOWLEDGMENTS

Sincere thanks to Prof. Dr. Deniz ZEYREK BOZŞAHİN , Prof. Dr. Cem BOZŞAHİN, Prof. Dr. İsmail Sengör ALTINGÖVDE, Assoc. Prof. Dr. Barbaros YET, Assist. Prof. Dr. Umut ÖZGE and Dr. Murathan KURFALI for their invaluable contributions during various stages of the thesis work.

TABLE OF CONTENTS

ABSTRACT.....	iv
ÖZ.....	vi
DEDICATION.....	viii
ACKNOWLEDGMENTS	ix
TABLE OF CONTENTS	x
LIST OF TABLES	xiii
LIST OF FIGURES	xv
LIST OF ABBREVIATIONS	xvii
CHAPTERS	
1 INTRODUCTION	1
1.1 Background.....	1
1.2 Research Questions	3
1.3 The Approach towards Solutions	4
1.4 Contributions of the Study.....	5
1.5 Organization of the Thesis.....	7
2 RELATED WORK	9
2.1 Shallow Discourse Parsing	9

2.1.1	The Penn Discourse Bank Annotation Principles	9
2.1.2	Shallow Discourse Parsing Sub-tasks	12
2.2	A Brief Survey of Related Work	13
2.2.1	Discourse Parser Development and Discourse Relation Realization Type Identification Methods	13
2.2.2	Pre-Trained Language Models: The Paradigm Shift in NLP Domain	16
2.2.3	Cross-lingual Transfer Learning	20
3	METHODOLOGY	23
3.1	Datasets	23
3.1.1	Turkish Discourse Bank 1.2	23
3.1.2	Chinese Discourse Bank	25
3.1.3	Penn Discourse Treebank 3.0	25
3.2	A Snapshot of the Methodology of the Three Phases of the Thesis	25
3.2.1	The Multilingual Dataset	25
3.3	The Conduct of Experiments	26
4	STUDY I: EARLY EXPERIMENTS ON LINGUISTIC FEATURE SELECTION	31
4.1	Analysis of the Best Features	33
5	STUDY II: ENCODING TDB DATASET WITH USE AND MONOLINGUAL BERT FOR DR IDENTIFICATION AND SENSE IDENTIFICATION EXPERIMENTS	39
5.1	Encoding TDB 1.2 with USE for Testing Multiple Classification Algorithms	40
5.2	Experiments Based on Encoding TDB 1.2 with Turkish BERT PLM	41
5.2.1	Experimental Setup	41
5.2.2	Discourse Relation Realization Type Classification	44
5.2.3	Semantic Similarity Analysis of Discourse Relation Realizations	46

5.2.4	Discourse Relation Sense Classification	46
5.2.5	A Cross-Domain Experiment	49
6	STUDY III: ENCODING MULTILINGUAL DATASET WITH THE MULTILINGUAL BERT FOR CROSS-LINGUAL TRANSFER LEARNING EXPERIMENTS	51
6.1	Reasoning for Cross-lingual Transfer Learning Experiments	51
6.2	Cross-lingual Transfer Learning Experiments and the Results	52
6.3	Semantic Similarity Analysis of Discourse Relation Realizations	56
7	CONCLUSION AND FUTURE WORK	61
7.1	A Discussion of the Research Questions	62
7.2	Highlights of the Contributions and Notes on Further Research	64
	REFERENCES	67
	APPENDICES	
A	PDTB 3.0 SENSE HIERARCHY	75
B	THE CLASSIFICATION MODEL, METHOD OF FINE-TUNING AND TESTS	77
C	KAPPA ANALYSIS OVER CONFUSION MATRICES	81
	CURRICULUM VITAE	85

LIST OF TABLES

Table 1	Annotation statistics of DR Realization Types in TDB 1.2 and Level-1 senses	24
Table 2	Annotation statistics of DR Realization Types in TDB 1.2, TED-CDB, and PDTB-3 (<i>the Multilingual Dataset of the work</i>)	26
Table 3	Annotation statistics of Level-1 senses in TDB 1.2, CDB, PDTB-3 (<i>the Multilingual Dataset of the work</i>)	26
Table 4	DR Annotations in TDB 1.1: A Summary	31
Table 5	Annotation statistics of Level-2 senses	32
Table 6	Evaluation of Explicit DR Identification on TDB 1.1 by Linear Regression	36
Table 7	Overview of the Experiments	39
Table 8	Accuracies of the Classification Algorithms on the TDB 1.2	40
Table 9	The distribution of labels in the DR realization type classification experiments	44
Table 10	The distribution of labels in the DR sense classification experiments of Explicit and Implicit type of DR realizations	44
Table 11	DR Classification Evaluation of TDB 1.2	44
Table 12	DR realization type classification results over the TDB 1.2	45
Table 13	Explicit and implicit sense classification results over the TDB 1.2	49
Table 14	Cross-domain DR realization type classification results over the T-TED-MDB	50
Table 15	Cross-domain four-way sense classification results over the T-TED-MDB	50
Table 16	DR token statistics of the three datasets in the joined multilingual dataset formations .	52
Table 17	F1-Scores of Cross-lingual Transfer experiments that classify DR realization types in TDB 1.2.	53
Table 18	F1-Scores of DR realization type classification experiments of: (i) TDB 1.2, CDB and PDTB 3.0, encoded with their respective monolingual BERT PLM, (ii) The combina- tion of three languages, encoded with the multilingual BERT PLM.	54
Table 19	F1-Scores of the same DR realization type classification experiments with equal number of tokens in each DR realization type category of each language	55

Table 20	Overall DR Realization Type Classification Evaluation for TDB	63
Table 21	Overall DR Realization Sense Classification Evaluation for TDB	63
Table 22	BERT_MultiClass TensorFlow Model	79
Table 23	κ Coefficients of the Confusion Matrices in the Figures 22, 23 and 24 Calculated by the Formula 5	82

LIST OF FIGURES

Figure 1	The research workflow of Study 2 and 3 in the thesis.	6
Figure 2	The Transformer – Model Architecture [1]	18
Figure 3	The TensorBoard view of the Turkish BERT encodings for the noun “burnunu” (his/her nose).	28
Figure 4	The TensorBoard view of the Turkish BERT encodings for adjective “iğrenç” (disgusting).	29
Figure 5	Number of DR Class Tokens in TDB 1.1	32
Figure 6	Supervised Learning Approach	33
Figure 7	POS Tag Samples	34
Figure 8	Correlation levels of DCs to Being an Explicit DR (<i>The magnitude represented by the thickness of the line.</i>)	35
Figure 9	POS Tag Data Bag of Words Data Matrix Sample Portion	35
Figure 10	Absolute Values of the Linear Regression Coefficients of Selected Features	36
Figure 11	BERT Classification Architecture	41
Figure 12	The plotting of loss and accuracy values during the experiments for fine-tuning of Turkish BERT Model over TDB 1.2	42
Figure 13	The Workflow of the Experiments Conducted by Encoding with BERT Models ..	43
Figure 14	The bar-chart showing the semantic similarity analysis of DR types in the TDB 1.2 (encoded with BERT Turkish PLM)	47
Figure 15	The box-plot showing the semantic similarity analysis of DR types in the TDB 1.2 (encoded with BERT Turkish PLM)	47
Figure 16	The confusion matrix of DR realization Level-1 sense classification, where the model is trained over TDB 1.2 (encoded with BERT Turkish PLM)	48
Figure 17	The box-plot of the semantic similarity scores and bar-chart of the semantic sim- ilarity averages of DR realization text elements per each realization type category be- tween CDB and TDB 1.2 (encoded with the Multilingual BERT)	57

Figure 18	The box-plot of the semantic similarity scores and bar-chart of the semantic similarity averages of DR realization text elements per each realization type category between PDTB 3.0 and TDB 1.2 (encoded with the Multilingual BERT)	58
Figure 19	PDTB 3.0 Sense Hierarchy [2]	76
Figure 20	The Symbolic Representation of the BERT MultiClass TensorFlow Model	77
Figure 21	The bar chart of the number of words in the textual elements of a DR in TDB 1.2 listed in Table 1.	78
Figure 22	Confusion Matrices of Classification of DR Types in TDB 1.2 (Turkish) Encoded with BERT Turkish PLM (left) and the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages (right).	83
Figure 23	Confusion Matrices of Classification of DR Types in CDB (Chinese) Encoded with BERT Chinese PLM (left) and the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages (right).	83
Figure 24	Confusion Matrices of Classification of DR Types in PDTB-3 (English) Encoded with BERT English PLM (left) and the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages (right).	84

LIST OF ABBREVIATIONS

Arg	Argument (of a discourse relation)
AltLex	Alternative Lexicalization type of DR realization
BERT	Bidirectional Encoder Representations from Transformers
CDB	Chinese Discourse Bank
CTL	Cross-lingual Transfer Learning
CLS	Class tag used by BERT
DC	Discourse connectives
DB	Discourse Bank
DR	Discourse Relation
EntRel	Entity Relations type of DR realization
GPT-3	Generative Pre-trained Transformer 3
GPU	Graphical Processing Unit
κ	Cohen's Kappa Association Coefficient
METU	Middle East Technical University
MDB	Multilingual Discourse Bank
NoRel	No Relations type of DR realization
NN	Neural Network
NLP	Natural Language Processing
NLG	Natural Language Generation
NLU	Natural Language Understanding
PLM	Pre-trained Language Model
PDTB	Penn Discourse Treebank

POS	Part of Speech
R&D	Research and Development
SDP	Shallow Discourse Parsing
SE	Sentence Encoding
SEP	Separator tag used by BERT
Seq2Seq	Sequence-to-Sequence
SOTA	State-of-the-art
RNN	Recurrent Neural Network
TDB	Turkish Discourse Bank
TED	Technology, Entertainment, Design
TF-IDF	Term Frequency-Inverse Document Frequency
USE	Universal Sentence Encoder

CHAPTER 1

INTRODUCTION

In the search of discourse parsing methods in a low-resource scenario, this study is based on two main principles. One is to work with a discourse framework (the Penn Discourse Treebank, or the PDTB framework) that focuses on local discourse relations between two individual discourse units rather than the frameworks based on learning a more hierarchical structure. The other is to search for an enhancement to low resourced languages by focusing on Turkish and working with a multilingual corpus, annotated by the rules and principles of the PDTB style datasets in Turkish, English and Chinese.

1.1 Background

The primary target language for the solution of the research questions of the thesis is Turkish, which has a complex morphology where suffixation is a major tool of both derivation and inflection, and hence poses several challenges for Natural Language Processing (NLP) [3]. Turkish is a language of more than 80M speakers and belongs to the Turkic sub-family of the Altaic language family. Despite its large number of speakers, its entrance to the NLP field is rather recent, and language technology tools have been attempted only in the last few decades [4]. The interest in Turkish NLP and language technology tools have been increasing with the research and improvements in sentence-level tasks such as named entity recognition [5, 6] as well as semantics [7, 8], suggesting new solutions for Natural Language Understanding (NLU), similar to most of other NLP topics.

NLU means the perception of the text content or speech by machines at a certain accuracy level and it has always been one of the the most interesting fields of NLP. While the level of human-computer interaction is increasing, one of the most rapidly developing fields in this area is NLU. Understanding natural language texts not only requires the knowledge of sentence structure and meaning, but also the knowledge of how texts cohere. Although naturally easy for human language users, understanding how texts cohere is still a challenge for NLU because the task requires to go beyond words and clauses.

Recently, to enable research on linguistic structures above the sentence level, and to enhance language technology applications that exploit such structures (such as text summarization, dialogue systems, information retrieval and machine translation), there has been rigorous attempts to create linguistic corpora annotated for semantics or discourse, e.g., Framenet [9], Propbank [10], Groningen Meaning Bank [11], and the PDTB [12].

Being one of the most interesting aspects of coherence, discourse relations (DR) are well worth analysing. The task of the analysis of DRs may be defined as the process of determining various aspects of contextual information surrounding certain lexical anchors, namely, discourse connectives, such as their sense, or the length of the spans linked by them. DR analysis is also very helpful for the downstream NLP tasks that need information above the clause level, such as text summarization, dialogue systems, information retrieval and machine translation.

What are Discourse Relations, How does the PDTB capture them, and How are they Realized in Texts?

Briefly, discourse relations are semantic (e.g., addition, cause-effect, conditional, similarity) or pragmatic relations (e.g., speech act relations) that hold between two adjacent discourse units (clauses or sentences). They may be realized explicitly, with a connective (*and*, *but*, *however*, etc.), or implicitly without any connectives. To date, the largest annotated discourse corpus for English is the PDTB containing over 40,600 DR annotations on Wall Street Journal texts [13].

We have worked on annotated corpora that follow the rules and principles of the PDTB framework and the PDTB team’s best annotation practices. The PDTB framework is aimed to be theory-neutral and gathers years of discourse knowledge in a single corpus. The discourse-annotated Turkish corpora used in this thesis show the validity of this knowledge in a typologically different language. From the annotation perspective, the PDTB has a simple labelling template that captures discourse phenomena at the local level; that is, the annotators are only asked to recognize discourse relations and other phenomena at a local level without keeping in mind the global, hierarchical structure of the texts. Thus, a good agreement level could be obtained between annotators and sustained across texts. Likewise, our tests for the labelling accuracy of the Turkish discourse corpora, produced by the METU research team, have ended up with good performance levels.

In the PDTB framework, connectives are considered as lexico-syntactic devices that signal the presence of a discourse relation. This is referred to as the lexical approach to discourse. Relations that are made salient by discourse connectives (DC) are referred to as explicit relations. Discourse relations may also be instantiated without any discourse connective, known as implicit relations. Even in these cases, the semantic relation between text segments (referred to as the arguments of a relation) can be easily inferred by humans. For example, the conjunction *and* in Example 1.1.1 makes the additive relation between two propositions explicit, while Example 1.1.2 conveys an implicit comparison relation, more specifically, a contrast relation between two sentences:

Example 1.1.1 (Explicit DR) *Jack is interested in computer science **and** he’s planning to apply for a Ph.D. in this field.*

Example 1.1.2 (Implicit DR) *Jack loves classic novels. His brother is an ardent reader of science fiction.*

The PDTB treats discourse connectives as discourse-level predicates that take two abstract objects as arguments (events, states, and propositions [14] and annotates explicitly and implicitly conveyed relations, their arguments, and senses. This annotation style has triggered an active line of research in discourse parsing. Particularly targeting English, the PDTB has triggered an active line of research on the topics of automatic DC detection, as well as DR sense prediction, and has even led to end-to-end discourse parsers, detecting all components of annotated DRs.

However, many languages still lag behind such developments presenting a challenge to universal end-to-end NLU pipelines. One could start with DR realization identification (i.e. whether a relation is explicit or not) that could serve to construct such pipelines and to the best of our knowledge it is a novel task for Turkish. We will also attempt to classify the Level-1 senses of DRs, as explained in the upcoming chapters.

1.2 Research Questions

Given that discourse relations are semantic (or pragmatic) objects instantiating how different segments of text are related to each other, DR realization detection followed by DR type identification is one of the main aspects of the task required by natural language text comprehension. This has been an open problem in natural language applications and most studies have focused on distinguishing explicit DRs from implicit ones. Until recently, most of the efforts have attempted to accomplish this task via linguistic feature engineering, which has become an integral part of NLP.

Because the free text nature of languages is meaningless to machines, both the academia and the industry have been doing research in search for computational models to translate text into low-dimensional real-valued vectors. Linguistic feature engineering is applied to extract the features according to a set of linguistic rules represented in the syntax and grammatical structure of languages. Although they are useful to analyze the grammatical correctness of texts, they cannot always help to develop fundamental techniques to understand any natural language such as its semantics. Thus, the popularity of feature engineering, especially in the academic setting, is criticized due to their limited capacity [15] and this critic has formed the departure point of this study, leading to a series of experiments that incorporate state-of-the-art language models as well as feature engineering. The questions that have motivated the present thesis were as follows:

- **What is the best performance level that could be reached by linguistic feature engineering and syntactic parsers in the classification of DR realization identification? Considering that feature engineering might work for a scenario where the number of annotated data is limited, how cost effective could be the deep learning models which require a lot of training?**

As we will see in the next chapter, although Turkish is missing in the discourse parsing literature, feature engineering methods (namely, methods that use syntax especially to identify explicit discourse connectives) have led to very good results in English. Eventually, end-to-end discourse parsers that join multiple components of DRs in a sequential pipeline architecture [16] have been designed. Given the absence of works on Turkish discourse parsing, the main motivation of the current thesis is the need to design new models for data preprocessing by applying sentence-level text encoding mechanisms utilizing Pre-trained Language Models (PLM) and to test the efficiency of these models in the classification of DR realization (Explicit, Implicit, etc.) types as well predicting their sense. Thus, our main goal was also to answer the following research question:

- **What is the impact of PLM based sentence-level text encoding mechanisms over DR realization identification (predicting the labels over DRs such as Explicit, Implicit, etc.) and their senses?**

In a nutshell, the current study turns the DR identification problem into a multi-class classification task and presents a series of experiments that use sentence encoding with state-of-the-art models and Neural Network-based classifiers. We also asked:

- **What could be the architecture of a yet another Neural Network-based classifier that might possibly reach the best performance in predicting the labels of DR types (Explicit, Implicit, etc.) and their senses separately?**

A secondary but a much-needed task in the discourse parsing field is to compensate the data scarcity problem. (Since discourse involves various forms of expressions with a variety of senses that are often ambiguous, building consistently annotated, large discourse datasets is not a trivial task and it is very expensive.) This aim could be achieved not only by the help of multilingual PLMs and encoding multilingual datasets, but also by the Cross-lingual Transfer Learning approach. Thus, other research questions of the thesis were:

- **Will a broad Cross-lingual Transfer Learning classification experiment that uses a multilingual dataset encoded by a multilingual PLM for text encoding circumvent the data scarcity problem?**
- **What kind of an effect could be produced by the model, fine-tuned within such an experiment, over the performance of the DR realization identification task?**

1.3 The Approach towards Solutions

This thesis is aimed to be the first work towards an end-to-end Turkish discourse parser which aims to uncover all the underlying DRs, along with their arguments and senses, if any, in a given text. The work involves various sub-tasks, each targeting different components of discourse relations, e.g., identifying explicit versus non-explicit DRs, identifying the textual spans that are related in each DR, identifying the sense of DRs (see Section 2.1). Our current pipeline sidesteps the problem of argument span extraction by performing DR realization identification directly on text by modeling the task as multi-class classification and presenting a series of experiments that use the modern PLMs and neural network-based classifiers. Specifically, we train models to perform two tasks through (i) six-way DR realization type identification experiments (explicit, implicit, etc.), and (ii) classification of Level-1 senses (i.e., classifications at a coarse level involving Temporal, Expansion, Comparison, Contingency categories), where the sense classification of implicit relations is known to be the most challenging task in discourse parsing.

We make use of the existing small but reliably annotated Turkish datasets of the discourse banks, and the datasets for Chinese and English, all of which fully comply with the rules and principles of the PDTB:

- METU Turkish Discourse Bank (TDB) 1.2 (a new version of TDB 1.1 explained in [17] and [18]),
- Turkish subpart of the TED Multilingual Discourse Bank (T-TED-MDB) [19]

- TED-Chinese Discourse Bank (TED-CDB) [20],
- The PDTB-3, the latest annotation of PDTB [21].

Besides using the Turkish datasets for our specific purposes, we also discuss whether the Cross-lingual Transfer Learning concept would be helpful if we joined the TDB 1.2 with TED-CDB and PDTB-3 (both of which are larger annotated-corpora than the Turkish datasets) to develop a multilingual DR realization identification model to improve the efficiency of the respective task. The details of the datasets are provided in Section 3.1.

1.4 Contributions of the Study

Complying with the reported experiments on shallow discourse parsing in the literature, namely the tasks of DR realization type identification and sense identification, the major aspect of the present work is to build classification models to predict DR types (Explicit, Implicit, etc.) and their senses. The contribution of the thesis is potentially important not only for the discourse perspective at a low-resourced scenario, but also towards developing solutions for the needs arising from the NLU development side of AI.

We perform the most thorough analysis on discourse parsing on Turkish through three phases named as Study 1, Study 2 and Study 3, as detailed in further chapters. Study 1 involves DR identification and classification via linguistic feature engineering. Here, the DR identification task is not separated as DR type and sense identification, but DR types and their senses are combined into a Class. The thesis continues with Study 2, where DR types and senses are handled separately, and the effectiveness of monolingual BERT is explored for both in-domain (Turkish Discourse Bank) and out-domain datasets (TED-Multilingual Discourse Bank). The thesis culminates in Study 3, where the effectiveness of multilingual BERT encoding is explored through multilingual datasets, and a DR realization type classifier and Level-1 sense classifiers are built for explicit and implicit discourse relations. These are depicted in Figure 1, and the tasks mentioned are novel tasks for Turkish, to the best of our knowledge, at the time when the thesis was conceived.

Our major contributions, then, can be summarized as follows:

- As an aspect of NLU, shallow discourse parsing in an understudied (non-English) context has been attempted by performing various sub-tasks of the prospective Turkish shallow discourse parsing pipeline. It specifically focuses on a rather overlooked task (the classification of DR realization types). Starting with an attempt to incorporate linguistic feature selection for DR Class classification (Study 1), and exploring the effectiveness of PLMs over DR type classification and DR sense classification separately (Study 2, 3), it constitutes the most exhaustive study of Turkish discourse parsing, as it also incorporates the difficult task of the sense classification of DRs.
- Different from Study 1, in Study 2, the classification of DR types and their senses were taken separately and experiments with USE and a monolingual BERT were run. The negative effect caused by a highly unbalanced number of DR type samples in TDB is investigated with a cosine similarity analysis of DR realization types in this corpus. Also, with the monolingual BERT

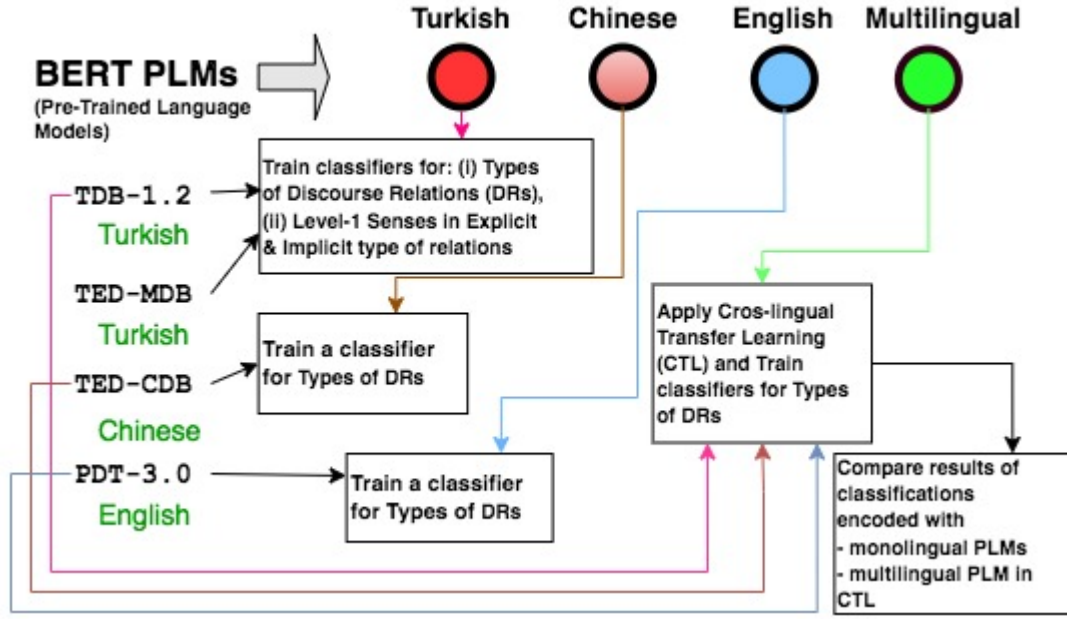


Figure 1: The research workflow of Study 2 and 3 in the thesis.

PLM, a cross-domain experiment that classifies DR realization types are done over TED-MDB, and the results are compared to the results over TDB. Study 3, detailed below, was exhaustively devoted to multilingual investigations.

- Overall, the results of the experiments conducted with PLMs (Study 2, 3) are quite satisfactory and show that the models can detect the facts that are easily inferred by humans in the written text at levels compatible with the levels reported for other languages in the literature.
- In its entirety, the thesis can be considered an efficiency analysis done with baseline classification algorithms for the automatic recognition of Turkish DR realization types and senses.

Beyond an attempt to automatically recognize Turkish DRs realization types and senses, this thesis explores cross-lingual datasets as a remedy for the training data scarcity problem that arises due to the cost of annotation in discourse. Thus, the most optimal models are built for a language whose discourse structure is less known, aiming to make multilingual NLU able to keep up with such languages.

- In an attempt to reach a high performance level of the tasks at hand, in the final phase (Study 3), the following have been accomplished: (i) investigation of the addition of an additional language to BERT by preparing a custom multilingual dataset consisting of three languages (English, Chinese and Turkish) and run experiments ending up with new multilingual PLMs fine-tuned for the tasks. (ii) the efficiency assessment of the Cross-lingual Transfer Learning technique by comparing the results of the experiments over the monolingual (Turkish) PLM with those of multilingual PLM.
- By aggregating the TDB 1.2 with the PDTB 3.0 and the CDB in Study 3, to the best of our knowledge, we have performed the first Cross-lingual Transfer Learning investigation on the DR realization type identification task in Study 3.

- The experiment with a multilingual PLM (Study 3) reached a performance very close to that of the monolingual one (Study 2). The unexpected difference between the effects of Chinese versus English over Turkish to DR realization type classification performance is investigated with another cosine similarity analysis.
- Finally, in Study 3, the scientific validity of multilingual test results over DR realization types are tested by conducting a Kappa Analysis over confusion matrices.

1.5 Organization of the Thesis

The rest of thesis proceeds in the following manner: A literature review is provided in Chapter 2 (p. 9) which starts with the section on Shallow Discourse Parsing (SDP) and describes the PDTB annotation principles, focusing on DR realization types and the sub-task descriptions of our SDP pipeline undertaken in the current research. Section 2.2 of this chapter is written to summarize our survey about the related work by providing the literature review on discourse parsing, as well as the necessary background to understand the models and techniques employed in our experiments. The chapter also contains an overview of DR analysis that led to the baseline classification algorithms in the present work, namely, Neural Network (NN) models, Transfer Learning, PLMs for text encoding and Cross-lingual Transfer Learning.

The Literature Review is followed by the Methodology Chapter 3 (p. 23) which describes the data sources of the work, and overviews the methods used in the experiments.

In Chapter 4 (p. 31), the first group of experiments (Study 1) is detailed. This Study approaches the discourse relation identification task by merging DR types and Level-1 senses into Classes and uses machine learning methods. The experiments are enriched with linguistic features used to find the best machine learning method for the discourse relation Class classification.

As introduced in Chapter 5 (p. 39), in the second group of experiments, the task is separated into two, as a six-way DR type identification and a four-way sense classification task, where the TDB is encoded with PLMs towards solving our tasks (Study 2). Firstly, USE is used for encoding the TDB, and the DR Class classification and sense classification experiments are run. The outputs are tested with various legacy classification algorithms. Secondly, DRs are encoded with the BERT Turkish PLM (also referred to as the monolingual BERT in the thesis), where we fine-tuned the best performing models. A cross-domain experiment over TED-MDB has been realized as well.

Chapter 6 (p. 51) details the third group of experiments (Study 3) in which multilingual dataset variations are encoded with the multilingual BERT for Cross-lingual Transfer Learning experiments. This experiment involves the task of discourse relation type classification only and has the aim of exploring the extent at which this task can be leveraged by multilingual datasets. At the end, a category-wise semantic similarity analysis is conducted over the arguments of discourse relations in order to assess the performance.

Finally, Chapter 7 (p. 61) concludes the thesis by evaluating the findings obtained in the experiments conducted in the thesis. The results of our experiments carried within the scope of the thesis are discussed along with their implications and some future directions are suggested.

The thesis has three appendices: Appendix A presents the PDTB 3.0 Sense Hierarchy mentioned throughout the thesis. Appendix B overviews the details of the experimental setup and the classification model used in all fine-tuning and tests, in order to facilitate the replicability of our results. Appendix C includes the proof of multilingual test results' scientific validity by conducting a Kappa Analysis over confusion matrices.

The next chapter will present a review of the background to the present thesis.

CHAPTER 2

RELATED WORK

One of the major, yet unresolved problems in NLP for NLU is discourse processing. There have been so many theories and approaches developed so far for building models of how utterances stick together to form coherent discourses. In such a broad area of study, this chapter focuses on the development of discourse parsers. The chapter involves the literature review on Shallow Discourse Parsing (SDP) (Section 2.1), the methods in discourse parser development (Section 2.2.1), the new methods that have appeared only in the last decade via the NN based developments and the PLMs (Section 2.2.2). The chapter ends in Cross-lingual Transfer Learning (Section 2.2.3) as a potentially useful task for the identification of DR realization.

2.1 Shallow Discourse Parsing

SDP refers to uncovering local discourse relations in a text as they are defined according to the PDTB. Since all the data of the current work is based on the PDTB Annotation Principles, they are explained in detail with a focus on DR realization types, before the details of SDP. Then, we describe the sub-tasks of the SDP pipeline undertaken in the current research. Finally, the attempts towards the identification of DR realization and DR senses as well as the methods for discourse parser development are explained in this section.

2.1.1 The Penn Discourse Bank Annotation Principles

The current thesis is based upon Turkish (and as it will be discussed in the following chapters, Chinese) discourse corpora that follow the PDTB 3.0, which is the most recent version of the PDTB corpus. While version 3.0 keeps its original rules and annotation principles in place, there have been important extensions that mainly involve the following: (i) it introduced a new, flatter sense hierarchy (provided in Appendix A) of the current thesis), (ii) it increased the number of annotated DRs (see Section 7.2), (iii) in addition to inter-sentential implicits annotated in earlier versions, it annotated implicit relations also at the intra-sentential level, and (iv) annotated the multiple senses of both explicit and implicit relations.

The PDTB 3.0 recognizes and annotates six DR types listed below. In addition to relation types, the binary arguments (Arg1, Arg2) to a relation, and the sense of DRs chosen from the hierarchy of senses

are annotated. Adjacent sentence pairs between which annotators found no implicit relation are further distinguished.

The arguments of a DR are constitutive units of the DR that have an abstract object interpretation (propositions, eventualities, etc.) which can be clauses, sentences, or multi-sentence segments [14]. The PDTB DR realization types are summarized below with examples taken from [22]. Discourse connectives, where available, are underlined for clarity.

Explicit DRs are those relations that are explicitly signaled through lexico-syntactic elements such as coordinating conjunctions (*and, so, but*) subordinating conjunctions (*because, after, while*), adverbials (*however, additionally, consequently*) or prepositional phrases (*in summary, on the contrary*) (see Example 2.1.1). These markers are referred to as explicit discourse connectives.

Example 2.1.1 *The city's Campaign Finance Board has refused to pay Mr. Dinkins \$95,142 in matching funds because his campaign records are incomplete.*

Implicit DRs are not only inferred from the linguistic context but they can also be rephrased by an overt marker. In cases where a discourse relation is not overtly marked by a connective, human language users can still infer the sense of the relation and can also rephrase the discourse relation with an overt marker. The PDTB asks annotators to insert an explicit marker that best conveys the sense of an implicit relation. These are called implicit connectives (see Example 2.1.2 for an intra-sentential implicit relation, and example 2.1.3 for an inter-sentential implicit).

Example 2.1.2 . . . the government should encourage home ownership, (Implicit = by means of) including issuing bonds that guarantee holders the right to purchase an apartment.

Example 2.1.3 *So much of the stuff poured into its Austin, Texas, offices that its mail rooms there simply stopped delivering it. (Implicit = so) Now, thousands of mailers, catalogs and sales pitches go straight into the trash.*

Implicit discourse relations are further realized by alternatively lexicalizing a relation, entity-based coherence, 'no relation', and hypophora.

Alternative Lexicalization (AltLex): When an implicit discourse relation is inferred but the insertion of an implicit connective in the relation is perceived redundant, the relation is referred to as being alternatively lexicalized. Expressions that are inferred to confirm the presence of a discourse relation are annotated as AltLexes (see Example 2.1.4).

Example 2.1.4 *After trading at an average discount of more than 20% in late 1987 and part of last year, country funds currently trade at an average premium of 6%. The Reason: Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold.*

Entity Relations (EntRels) capture those cases where the implicit relation between adjacent text units is a form of entity-based coherence. When there is no discourse relation to be inferred between adjacent sentences and adjacent sentences form entity-based coherence with the same entity being realized in both sentences either directly or indirectly, the discourse relation is annotated as an EntRel (see Example 2.1.5).

Example 2.1.5 *Pierre Vinken, 61 years old, will join the board as a non-executive director Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.*

No Relations (NoRels) hold between adjacent text spans whose semantic relation cannot be identified (see Example 2.1.6). Where a semantic relation between adjacent text spans cannot be identified, the adjacent pair of sentences are annotated as NoRel.

Example 2.1.6 *Mr. Rapanelli met in August with U.S. Assistant Treasury Secretary David Mulford. Argentine negotiator Carlos Carballo was in Washington and New York this week to meet with banks.*

Hypophora holds when Arg1 expresses a question and Arg2 provides a meaningful response. Hypophora is a new coherence relation, standing for discourse relations involving dialogue acts, which cannot be instantiated through connectives [23, 91] where the relation consists of a dialogue act, which cannot be instantiated as connectives [24], (see Example 2.1.7). In short, if a question is asked and a meaningful response is provided, the relation is annotated as Hypophora.

Example 2.1.7 *Underclass youth are a special concern. Are such expenditures worthwhile, then? Yes, if targeted.*

The PDTB Sense Hierarchy involves four first-level senses, namely Expansion, Comparison, Contingency, and Temporal [12], assigned to Implicit and Explicit DR types and AltLexes annotated in the corpus. EntRels, NoRels, and Hypophora are not assigned a sense tag.

The Level-1 senses are refined to more specific senses at Level-2. For symmetric Level-2 relations such as Expansion:Conjunction (i.e., if the DR holds between Arg1 and Arg2, it also holds between Arg2 and Arg1), the hierarchy stops at this level. Asymmetric relations such as Expansion:Level-of-detail are further refined at Level-3. The Level-3 senses define the semantic contribution of each argument with respect to directionality.

In addition to DR types, the current work tackles sense classification only at the first level. The Level-1 senses are briefly described below with examples.

Temporal stands for the situations described in the arguments related temporally (see example 2.1.8).

Example 2.1.8 *Halil epeyce koştuktan sonra yüksekçe bir duvardan atlayarak mahallenin dışındaki kadınlar hamamının bahçesine girdi, (Implicit = o zaman) çevreyi iyice kolağan ettiğinde hamamın girişinde yanan bir lambayı fark etti. (Level-1 Sense: Temporal; Level-2 Sense: Synchronous)*

Comparison captures the cases where one eventuality is compared to the other in terms of the degree it is similar to or different from the other eventuality [17] (see example 2.1.9).

Example 2.1.9 *İyimser beklentiler sürüyor, ancak parayı tek enstrümana yatırmak riskli olabilir. (Level-1 Sense: Comparison; Level-2 Sense: Concession)*

Contingency holds when one of the situations described in Arg1 and Arg2 causally influences the other (see example 2.1.10). That is, it indicates that the situations described in the arguments influence each other causally. All conditional relations also fall into this category.

Example 2.1.10 *İbrahim Trk aday adayı olarak seilmese bile aktif politikada yer almı bir kiidir. Bu nedenle BDDK’da grev alması doęru olmaz. (Level-1 Sense: Contingency; Level-2 Sense: Cause)*

Expansion works to expand the discourse and move its narrative or exposition forward (see example 2.1.11). The discourse moves the narrative forward in certain ways such as via a paraphrase, by providing additional related information and by listing exceptions to the generalizations put forward in one of the arguments.

Example 2.1.11 *ABD’nin, "Iraklıları da Saddam’dan kurtaracaęız" tezi, pek raębet grmyor. (Implicit = yle ki) Bir ęrenci, bu durumu "Madem bizi kurtarmak istiyorlar, neden tepemize bomba yaędırmayı dnyorlar?" szleriyle zetliyor. (Level-1 Sense: Expansion; Level-2 Sense: Level-of-detail)*

2.1.2 Shallow Discourse Parsing Sub-tasks

A typical shallow discourse parser consists of three main sub-tasks: (i) connective identification, (ii) argument extraction and (iii) sense classification. We have adopted a different approach by merging and converting the first two tasks into the identification task of a discourse relation in a given text piece, referred to as *DR realization type identification*. That is, for the time being, we do not perform any connective and argument extraction; yet, our pipeline can still identify the discourse relations, if there is any, in a given text piece and further disambiguate them in terms of the senses.

DR realization type identification: This is one of the least studied aspects of discourse worth to be considered as an important step because it mimics PDTB annotation guidelines, where the annotators are asked to identify different realizations of discourse as a first step. So, automatically identifying different types of DR realization at the first step facilitates our work on the next sub-task as well.

Here, it is modeled as a six-way classification task which aims to identify possible DR realization types, explained in Section 2.1.1, in the given text, formed by the arguments of relations. Although the usage disambiguation of discourse connectives (distinguishing between the connectives’ discourse and non-discourse role) has been investigated for many languages as discussed in Section 2.2, and implicit relation identification has also been targeted, these have often been considered as tasks on their own. The multi-level identification of relation realization type is a challenging task, which to our knowledge, has not been tackled before.

Sense classification of discourse relations: This is the most popular sub-task of discourse parsing, where the aim is to find the sense conveyed by a given explicit or implicit discourse relation. The implicit classification task is highly challenging due to the lack of an explicit signal, i.e. the discourse connective, and the models must classify the semantics of each argument correctly. Due to its challenging nature, the task is most commonly limited to only the four Level-1 senses in the PDTB sense hierarchy; hence, in the current work, it is modeled as a four-way classification task. To the best of our knowledge, the four-way sense classification of implicit discourse relation is the first attempt over Turkish data.

2.2 A Brief Survey of Related Work

2.2.1 Discourse Parser Development and Discourse Relation Realization Type Identification Methods

End-to-end discourse parsing involves the distinction between discourse and non-discourse use of connectives, the identification of DR realization types, their senses, and their binary arguments. Contemporary research, prior to the development of an end-to-end discourse parser by [16], targeted different sub-tasks for parsing, such as detecting discourse connectives at high levels of accuracy, mostly by utilizing linguistic features and knowledge of the connector’s surrounding context. In an early paper, [25] used a decision tree approach to distinguish between explicit and implicit discourse connectives over the PDTB 2.0, as well as attempting a four-way classification of their senses and a separate four-way sense classification of tokens with explicit connectives. Their model reached a higher performance in the classification of explicit discourse connectives than implicit ones. One of the earliest studies on the identification of discourse vs. non-discourse usage of explicit connectives has been carried out by [26] over PDTB 2.0. Feeding syntactic features extracted from the arguments of discourse connectives into a maximum entropy classifier,¹ the authors reached an F-Score of 0.92 in explicit discourse connective disambiguation and 94% accuracy in the four-way sense classification of explicit discourse connectives [12].

Later work has reached highly successful results in domain-specific applications. [27] used various supervised machine-learning-based algorithms for automatically identifying explicit discourse connectives in BioDRB corpora, and proposed a hybrid classifier based on a conditional random fields-based classifier and a combination of instance pruning, feature augmentation, and domain adaptation techniques. Extracting syntactic features such as the part-of-speech (POS) tags of the tokens, the syntactic labels of the immediate parent of the token’s POS in the parse tree, and the POS tags of the left sibling (the token to the left of the current word inside the innermost constituent), an F-Score of 0.76 is reached. [28] used fewer linguistic features relevant to discourse and employed machine learning models to automatically extract explicit discourse connectives, their senses and arguments. The authors use features produced by text processing techniques and the syntactic classification of connectives (Subordinator, Coordinator, Conjunct Adverb, and Correlative Conjunction) and they reported an F-Score of 0.85 in the classification of explicit discourse connectives’ senses with the Conditional Random Fields technique.

Work on non-English languages such as Arabic discovered features that contribute to the disambiguation of the discourse usage of explicit connectives. [29] worked on an Arabic corpus where explicit discourse connectives are marked. They used syntactic features such as the position of the potential connective (sentence-initial, -medial or -final), lexical features of the surrounding words, the POS tags of the words and the syntactic category of the parent of the potential connective. The authors also discovered the predictive role of the infinitive form of the verb in the second argument of prepositional connectives, and achieved an F-Score of 0.78 in explicit discourse connective recognition with the best feature combination.

However, the presence of a DC does not always help for DR identification because the DC may be ambiguous between several senses. For instance, *since* can be used to signal either a temporal or a cause relation. Assuming that the distribution of the majority of the ambiguous DCs is highly skewed

¹ <https://github.com/mimno/Mallet>

toward certain senses, [25] use a feature set based on the surface levels of DCs in the PDTB 2.0 [12] and try to distinguish each DR type from the rest of the DR types. Among others, adjacency of DRs is used as a feature to enhance the disambiguation of implicit DRs. Working with the TED-CDB corpus, a recent work by [2] shows that there is room for the exploration of more effective disambiguation methods to handle the larger range of DC ambiguities.

While it could be said that recognition of explicit type of DR realization is largely a solved problem, especially for well-studied languages such as English, identification of implicit relations still remains a challenge that cannot be solved through feature-based approaches. A study by [30, 8] predicts the Level-1 senses of Implicit DRs in the PDTB 2.0 in a realistic setting and, "distinguishing a relation of interest from all others, where the relations occur in their natural distributions" achieve an F-Score of 0.76 with the best combination of their features (polarity+Inquirer tags+context) in the disambiguation of the Level-1 senses of the PDTB hierarchy.

[31] have been the first to identify implicit relations by removing discourse connectives to cheaply gather large amounts of training data. They also discovered that word pairs are indicative of implicit relations (e.g., the pair *embargo ... legally* was a good indicator of contrast relations) and used them in extracting large amounts of data. They reached high accuracies with this technique of obtaining artificial implicit relations. Particularly, two of their classifiers were successful in distinguishing between Cause-Explanation-Evidence vs Elaboration (93%) and Cause-Explanation-Evidence vs Contrast (87.3%).

A later paper by [30] predicted the Level-1 senses of implicit relations in the PDTB 2.0 in a realistic setting taking advantage of linguistically informed features as well as lexical pairs from unannotated text. They achieved an F-Score of 0.76 with the best combination of their features (polarity+Inquirer tags+context). [32] took parser production rules as the main source of features and worked on the Level-2 senses of the PDTB hierarchy. They showed that syntactic patterns could contribute to predicting the senses of implicit relations.

In CoNLL-2015 Shared Task, the feature-based work of [33] presented statistical classifiers for identifying the senses of implicit relations in the PDTB 2.0. The authors introduced novel feature sets that exploit distributional similarity and coreference information. They showed that Brown cluster pairs ([34]) work well in implicit relation recognition.

A breakthrough is seen in the field with the development of an end-to-end discourse parser (for English) by [16]. Working on the PDTB 2.0, the authors produced a PDTB-styled parser, constructing a thorough pipeline for parsing the text in seven sequential steps titled as connective classification, argument labeling, argument position classification, argument extraction, explicit relation classification, and sense recognition of explicit as well as non-explicit relations. The authors classified and labeled discourse relations and the attribution spans, where relevant. Their parser can also parse any unrestricted English text into its discourse structure in the PDTB style. The best F1-Scores are reported as 0.87 for the explicit classifier, and 0.4 for the non-explicit classifier, reaching to the highest success level of its period in the classification of DR realizations.

Lately, SDP has been attempted in a series of shared tasks. As in the work of [35], SDP is mostly conducted by using syntactic and semantic features for classification. It is also attempted in CoNLL 2015 Shared Task², to which 16 teams participated by using a piece of newswire text as input and re-

² <https://www.cs.brandeis.edu/~clp/conll15st/index.html>

turning relations in the form of a discourse connective (either explicit or implicit) with two arguments. Each team developed an end-to-end system that could be regarded as variations of [16], detecting and categorizing individual discourse relations and returning a set of relations contained in the text. The best system achieved an F1-Score of 0.24 on the blind test set, reflecting the serious error propagation problem in such a system [36, 14].

The Discourse Relation Parsing and Treebanking (DISRPT) 2019³ shared task event was held on discourse unit segmentation across formalisms, including shallow discourse parsing, aiming to promote the convergence of resources and a joint evaluation of discourse parsing approaches. The corpora included 15 datasets in 10 languages, 12 of which target elementary discourse unit segmentation, and three dedicated to explicit connective annotation [37]. In the overall evaluation, ToNy [38] performed the best on most of the datasets reaching an F-Score of 0.9 in the average of all its tests. Turkish is also represented in the dataset with the TDB 1.0, where only explicit discourse connectives are annotated. On this data, ToNy [38] obtained the best results in discourse connective detection on plain, unannotated data, reaching an F-Score of 0.85. In DISRPT21⁴, the shared task of Discourse Segmentation, Connective and Relation Identification across Formalisms was broadened by proposing the first iteration of a cross-formalism shared task on discourse relation classification. Best achievement for Turkish was reached by the approach of the group DiscoDisco, increased the F1-Score of explicit discourse connective detection sub-task to 0.94.

The methods enhanced with statistical approaches and machine learning techniques such as those summarized so far tried to increase the success rate of discourse parsers by surface level phenomena in addition to syntactic and semantic feature extraction out of annotated texts. One of the latest noteworthy achievements has been reported by [39]. The authors worked on the PDTB 3.0, where implicit relations are annotated both at the inter-sentential and intra-sentential levels. In addition to these stand-alone implicits, the implicit senses of explicit relations are also annotated. In a series of experiments, the authors first recognized the location of implicits, then they recognized their senses, arguing that the data annotated in this way simplifies the difficult problem of sense-labeling of implicits. The authors improved a Long Short-Term Memory (LSTM) classifier that runs for the classification of 20 categories formed by joining the first and second level DR senses. The classification tasks were repeated for each DR class and reached the highest F1-Scores of 0.75 in the intra-sentential classifier and 0.936 in the stand-alone classifier. The classification of inter-sentential and linked DRs remained at very low levels, showing there is a long way to go in Implicit DR Recognition. They also proved the claim that knowing the location of an Implicit DR would benefit sense identification.

To sum up, the supervised classification algorithms described above have been able to classify discourse relations, particularly explicit ones, very successfully in written texts but the results are far from societal impact, as less-studied, low-resource languages are hardly targeted. Moreover, sense-labeling of implicit discourse relations still lags behind that of explicit relations. The next section deals with the impact of deep learning models in the field of discourse understanding as background to the experiments conducted in the current work.

³ <https://sites.google.com/view/disrpt2019/shared-task>

⁴ <https://sites.google.com/georgetown.edu/disrpt2021>

2.2.2 Pre-Trained Language Models: The Paradigm Shift in NLP Domain

Plenty of NLP effort has been spent for the representation of text at the end of pre-processing. In order to capture the contextual information of the text, Count-based Vector Space Models (e.g., Count Vectorization, Tf-IDF) have been replaced by Context-Based Vector Space Models (e.g., ELMo, BERT). There are also Non Context-Based Vector Space Models (e.g., Word2Vec, FastText, Glove) in between but they are not mentioned here because of their little relevance for the current work.

Developments are also making discourse parsing improve well but as in many areas, deep learning methods, leveraged by the NN concept, have been prospering in the field of discourse to build universal, end-to-end models. For example, [40] built a Gated Relevance Network to capture important word pairs; [41] applied a sophisticated multi-layer attention model; and [42] employed a Recurrent NN stacked with convolutional networks. In addition, [43] built a simple word interaction NN model that captures word pair interactions by calculating an interaction score for each word pair and measuring the importance of the interaction between the component words.

The introduction of NNs to NLP has created ways to overcome the challenges that traditional methods couldn't totally solve, by bringing in the ability "to alleviate the feature engineering problem" [44, 1].

Discretely handcrafted features are being replaced by low-dimensional and dense vectors useful to develop various NLP systems able to understand syntactic or semantic features of the language. Methods using NN-based language models try to predict words from their neighboring words looking at word sequences in the corpus, and in the process, they learn distributed representations ending up with dense word embeddings beneficial for a wide variety of tasks, including scenarios with constraints such as lack of adequate data.

One of the most remarkable advances in the field has been Encoder-Decoder Architectures capable of taking inputs, for example sentences (sequences), and mapping them to a high-dimensional representation. The encoder here learns which parts of the inputs are important and passes them to the representation, while the less-important aspects are left out. So, the encoder technology gave birth to Sequence-to-Sequence (Seq2Seq) models, developed to achieve context-awareness, thus reducing ambiguity. Context-awareness is handled by encoding the source text, generating an output sequence, and predicting one word at a time; that is, Seq2Seq models can capture context by looking at a token in terms of previous words/sentences to generate the next words/sentences. The introduction of this representation of context embedded in space have multiple advantages. For example, data sparsity due to similar contexts is prevented, and mapping of sequences with meanings close to each other is avoided. Seq2Seq models also enable the generation of synthetic data. However, the linguistic context is very sophisticated and long range dependencies are needed to achieve context-awareness fully.

Recurrent NNs (RNN) are innovated to fulfill this need and have been used for a decade. In this line of research, [45] studied the quality of the vector representations of words derived through various models on a collection of syntactic and semantic language tasks. The vector representations produced by the available RNNs of the time were detected achieving over 50% increase in Spearman's rank correlation over the previous best result. RNNs are reevaluated by [46] showing that the vectors learned by the standard sigmoidal RNNs (which are highly non-linear) improve significantly as the amount of the training data increases. This result suggests that non-linear models also have a preference for a linear structure of the word representations.

As a generalization of feed-forward NNs to sequences, the RNN takes a sequence of inputs and computes another sequence of outputs by iterating over a sigmoid equation and can map sequences to sequences whenever the alignment between the inputs and the outputs is known ahead of time [47, 3]. Context-awareness is handled by encoding the source text, generating an output sequence, and predicting one word at a time; that is, Seq2Seq models can capture context by looking at a token in terms of previous words/sentences to generate the next words/sentences. The introduction of context representation in this way has multiple advantages: data sparsity is prevented, and mapping of sequences with meanings close to each other is avoided.

The advent of attention mechanism has been a breakthrough in NLP [48, 49] which, consequently, gave rise to the transformer architecture [1]. BERT is undoubtedly the most famous language model that is based on the transformer architecture [50]. Despite such developments, the vanishing gradient problem that occurs when the gradient shrinks during back-propagation through time, is still the hunchback for RNNs. Seq2Seq models utilizing RNN use memory mechanisms to regulate the flow of information when processing sequences to achieve a long term memory. RNN also falls short when trying to process an entire paragraph of a text, because too long a sequence makes it hard to carry information from the earlier time steps to later ones. On the other hand, if a gradient value becomes extremely small, it does not contribute much to learning. Moreover, RNNs' topology is very time-consuming, because for every back-propagation step, the network needs to see the entire sequence of words even though not all the text is equally important to gain an understanding. To address this, the attention mechanism is introduced to Seq2Seq models.

An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The attention mechanism performs the multiplication of these vectors, and depending on the angle of the vector, one can determine the importance of each value and it feeds the architecture giving more contextual information to the decoder. At every decoding step, the decoder is informed how much attention it should give to each input word. If the angles of the vectors are close to 90 degrees, then the dot product will be close to zero, but if the vectors point to the same direction, the dot product will return a greater value.

Despite these significant improvements in context awareness, there was still room for improvement. The issue of computational complexity, as the most significant drawback of these methods, gave birth to transformer models, introduced by Google in [1], which turned out to be a groundbreaking milestone in NLP (Figure 2).

Transformer models do not process an input sequence token by token, rather, they take the entire sequence as input in one go which is a big improvement over RNN based models because now the model can be accelerated by the GPUs. One of the most valuable enhancement of the transformer model is the production of PLMs without needing labeled data. That is, we just have to provide a huge amount of unlabeled text data to train a transformer-based model. We can use this trained model for other NLP tasks like text classification, named entity recognition, text generation, etc.

The transformer model introduces the forgetting mechanism to an already complex Seq2Seq model, simplifying the solution by forgetting about everything else and just focusing on attention. The transformer model removes recurrence by using matrix multiplications only. So, the number of sequential operations is reduced, and the computational complexity is decreased. The model processes all the inputs at once without having to process them in a sequential manner. For every back-propagation step, the network needs to see the entire sequence of words. To avoid losing order, it uses positional

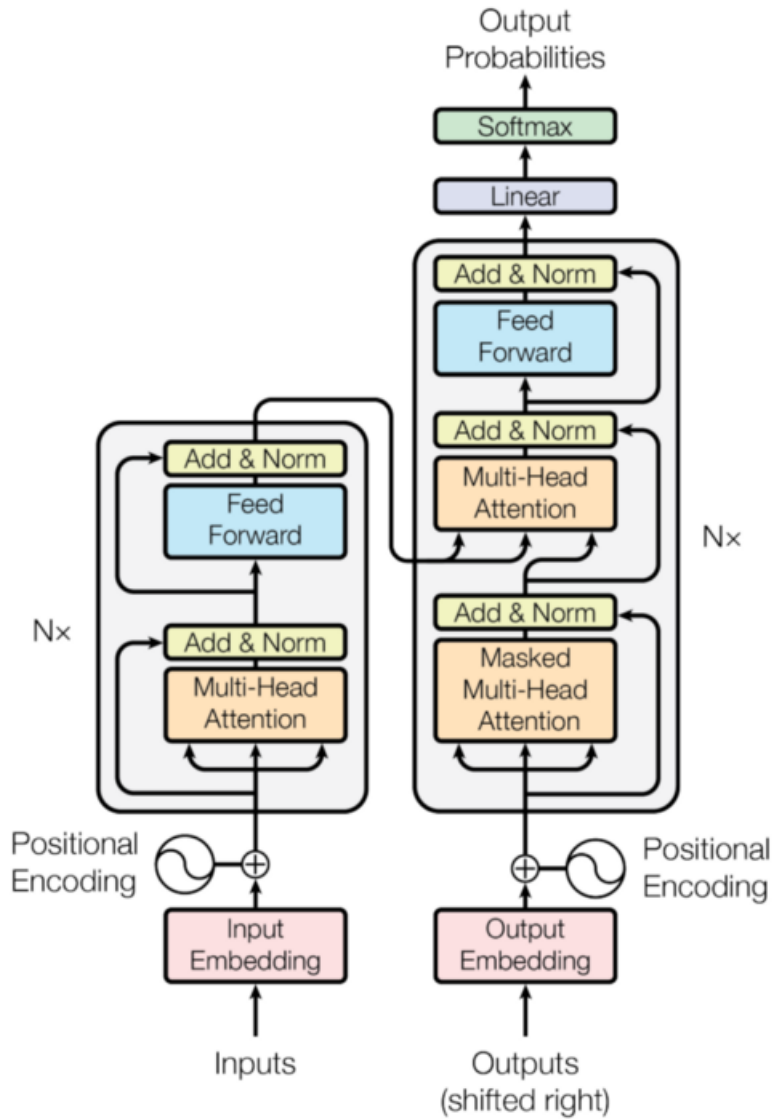


Figure 2: The Transformer – Model Architecture [1]

embeddings thus providing information about the position in the sequence of each element, created using sine and cosine functions with different dimensions. Words are encoded with the pattern created by the combination of these functions, resulting in a continuous binary encoding of positions in a sequence.

Implementing the encoder-decoder paradigm where the source sentence is encoded in a number of stacked encoder blocks, and the target sentence is generated through a number of stacked decoder blocks the Transformer Architecture's each encoder block consists of a multi-head self-attention layer and a feed-forward layer [51]. In multi-head attention, each head learns to pay attention to a specific group of words. That makes it easy to learn to identify short-range and long-range dependencies. This improves context-awareness making it possible to understand, for example, what the terms in a text

refer to. This is highly useful when the referents of terms are not clear, for example, in cases like pronoun resolution.

The issue of computational complexity, the most significant drawback of these methods, gave birth to the transformer model with a forgetting mechanism. A transformer model consists of three building blocks as given in [52] addressing the needs of the present work as well:

- a tokenizer, which converts raw texts to sparse index encodings,
- a transformer, which transforms sparse indices to contextual embeddings,
- a head, which uses contextual embeddings to make a task-specific prediction.

The transformer architecture facilitates the creation of powerful models trained on massive datasets making it possible to take advantage of transfer learning by re-using these PLMs and fine-tuning them for specific tasks.

The resolution of the linguistic feature selection problem and the ability to extract context information at good levels have led to the developments in discourse parsers and automatic DR identification. Deep learning methods leveraged by the NNs and reinforced by PLMs especially favored Implicit DR identification and ambiguity resolution tasks. One of the core models of transformers that sustains a good level of support to non-English languages is the Bidirectional Encoder Representations from Transformers (BERT), designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts in all layers [50].

BERT uses a masked language model which randomly masks 15% of the tokens from the input in order to predict the original vocabulary by feeding the final hidden vectors (corresponding to the masked tokens) into an output softmax over the vocabulary. The BERT output can be fine-tuned without substantial task-specific architecture modifications and, with just one additional output layer, SOTA models can be created for a wide range of tasks, such as Question Answering and Language Inference, which are based on understanding the relationship between two sentences. By virtue of the training done by BERT on the task Next Sentence Prediction, the model possesses valuable information for these tasks.

The computation power costs of transformers and memory limitations have led some researchers to lightweight solutions. Being one of them, Universal Sentence Encoder (USE) [53] follows the same approach, with the encoder that uses a transformer-network trained and optimized on a variety of data sources of sentences, phrases, and short paragraphs. USE can encode texts into high dimensional vectors that can be used for various tasks such as text classification, semantic similarity search and clustering. USE also makes use of a Deep Averaging Network, where input embeddings for words and bi-grams are first averaged together, and then passed through a feed-forward deep NN to produce sentence embeddings, resulting in the proof of the efficiency of transformer approach.

Seq2Seq Models are traditionally used to convert entire sequences from a target format into a source format. There are however more complicated tasks for language generation such as to generate text with a model that is capable of converting sequences, as we simply don't know the full sequence yet. The answer to creating a model that can generate text lies in the class of autoregressive models. Latest research is concerned with taking PLMs to further stages and advanced usages. One of them is "Generative Pre-trained Transformer 3" (GPT-3) [54], being created by the OpenAI Group as an

autoregressive language model, reaching 175 billion parameters at the time of writing. It has been pre-trained on one of the largest text corpora ever, consisting of about 499 billion tokens. The most prominent new method in GPT-3 is tokenizing the words in better resolution than BERT, but less common words like "establishment" are likely to be sliced up into more than one token (e.g., "estab", "lish", "ment") so as to produce more than one "token ids" for a word. On average, each token corresponds to about 4 characters of English text.

Transfer learning is a technique where a deep learning model trained on a large dataset is used to perform similar tasks on another dataset and the transformer architecture facilitates the creation of powerful models trained on massive datasets making it possible to take advantage of transfer learning by re-using these PLMs and fine-tuning them for specific tasks similar to that of this thesis. In the field of discourse, one of the early RNN practices is presented by [55], who build syntactic vector representation trees of arguments reaching the best F-Score of 0.8 in the identification of Expansion DRs. The use of BERT has been gradually increasing in DR classification. For example, [56] show that BERT significantly outperforms previous state-of-the-art models in DR classification. [57] presents a clear best adaptation of the BERT to the task by reaching an F-Score of 0.59.

2.2.3 Cross-lingual Transfer Learning

The main goal of Cross-lingual Transfer is Zero-Shot Learning, an impressive feat that envisions being able to train once and using it in all downstream tasks of all languages, causing it to become an active area of research in favor of low-resource languages. It is a fairly specific way of training models using the data available for languages with ample resources so that it can solve the same task in the target low resource language(s).

Cross-lingual Transfer Learning serves many research domains including the construction of bilingual dictionaries [58], zero-shot translation [59], spoken language understanding [60], semantic utterance classification [61], entity extraction from Web pages [62], fine-grained named entity typing [63], cross-lingual document retrieval [64], relation extraction [65], multilingual task oriented dialog [66], and event detection [67].

As a result of the recent increase in labeled text data, the number of NLP approaches based upon supervised learning has been steadily increasing. These approaches try to optimize their classification power by sensing the connection between sentences and the labels over these connections. Nowadays, these efforts have been supported by Cross-lingual Transfer Learning which has been used to prepare models out of big corpora in order to get better in the sensing capability.

A more specific case of transfer learning is known as *zero-shot learning*, where the classifier is able to classify examples, that have never appeared before, during the training. In the cross-lingual scenario, Zero-Shot Learning often translates into training the classifier in one language and applying it to other languages with a minimal performance loss. The advantage in general, is being able to leverage information from high-resource languages into the low-resource ones.

In the case of discourse parsing, it is highly relevant as almost all of the non-English languages lack large manually annotated datasets. The effects of recent approaches have to be gauged and the best practices have to be applied on less-studied languages. Thus, as detailed in Chapter 6, a task of conducting this approach implemented on BERT's multilingual PLM to produce a Zero-Shot Learning

model, has been included into the work in order to explore the usefulness of Cross-lingual Transfer Learning in DR realization identification.

Having reviewed the major works done in the discourse parsing literature, we will now move on to the Methodology of the thesis in the next chapter.

CHAPTER 3

METHODOLOGY

This chapter begins with an overview of the datasets, used as training data for supervised methods in various research experiments that have eventually allowed us to fine-tune the current PLMs and thereby create new models. The usage of datasets in the exercised state-of-the-art (SOTA) methods for finding better solutions to the research questions of the thesis is explained. The conduct of the methods and the results of the experiments are presented.

3.1 Datasets

In addition to the monolingual experiments with the main data source (TDB 1.2) of this work, multilingual experiments are conducted with the concatenation of three different language datasets for the experiments of the work. The datasets are the following discourse banks, all of which are annotated complying with the rules and principles of the PDTB:

- METU Turkish Discourse Bank (TDB) 1.2 (a new version of TDB 1.1 explained in [17] and [18]), a 40K-word multi-genre corpus of modern, written Turkish manually annotated for DRs, their binary arguments and senses,
- Turkish subpart of the TED Multilingual Discourse Bank (T-TED-MDB) [19],
- TED-Chinese Discourse Bank (TED-CDB) [20] (268.1K-word) transcription of Chinese TED talks,
- The PDTB-3, the latest annotation of PDTB corpus [21].

3.1.1 Turkish Discourse Bank 1.2

The TDB 1.2 is chosen as the main data source. It is a discourse-level resource of Turkish created by manually annotating a corpus of modern Turkish texts written between 1990-2000 [68]. The PDTB and the TDB share many goals, such as annotating the variety of explicit connectives and alternative lexicalizations, and annotating implicit relations. Thus, the annotation principles of the TDB and the annotated categories closely follow those of the PDTB; most notably, all discourse relation realization types described in Section 2.1.1 are spotted and annotated together with their binary arguments and senses, where relevant. The basic principle in annotating explicitly marked relations is the PDTB's

minimality principle, i.e. the annotators are asked to select the shortest text spans (e.g., clauses or sentences) that are necessary and sufficient to interpret a discourse relation encoded by a connective. The datasets over which the two corpora are built differ, however: while the PDTB is built over Wall Street Journals, the TDB is built over multiple genres (newspaper editorials, fiction, popular magazines). Other differences involve (i) the way different types of discourse connectives are annotated (e.g., suffixal connectives are annotated as a type of explicit connectives in the TDB [69]) (ii) a small number of new Level-2 sense tags are spotted and annotated in the TDB [17], and (iii) attribution has not been annotated in the TDB so far.

The current version of the TDB contains 3987 discourse relations (Tables 1, 9 and 10), where discourse senses are annotated on the basis of the PDTB 3.0 sense tag-set (Appendix A) in addition to the realization type of the DRs and their anchors (i.e. discourse connectives) are annotated (if available). Furthermore, as in the PDTB 3.0¹, explicit and implicit relations and AltLexes are annotated both at the inter-sentential and intra-sentential level. EntRels, NoRels and Hypophora (13% of the corpus) are annotated only at the inter-sentential level.

In its annotation stage, [70], any possible discourse relation is searched and annotated by going through the texts sentence by sentence, similar to the incremental processing of discourse. Annotations were performed by a team of trained graduate students, sustaining a minimum value of $\kappa = 0.7$ for inter-annotator agreement [17]. Even though this is below the normal threshold of 0.8, due to the ambiguity of coherence relations, the annotation task is a very hard task, forcing the team to take 0.7 as a satisfactory level as suggested by [71]². The numbers of annotated DR realizations (i.e., type of realization and senses) are given given in Table 1 and Example 3.1.1 presents a sample of annotated DRs in TDB 1.2, where bold fonts show the DR labels used in the present work.

Table 1: Annotation statistics of DR Realization Types in TDB 1.2 and Level-1 senses

Types	Num.& Ratio	Level-1 senses	Num.& Ratio
Explicit	1,524 (38.2%)	Comparison	448 (12.9%)
Implicit	1,791 (44.9%)	Contingency	702 (20.3%)
AltLex	152 (3.8%)	Expansion	1,700 (49.0%)
Hypophora	80 (2%)	Temporal	617 (17.8%)
EntRel	237 (5.9%)		
NoRel	203 (5.1%)		
Total	3,987 (100%)		3,467 (87%)

Example 3.1.1 Arg1: *Kendinden ne kadar uzaklaşabilir ki insan? (How far could a person get away from herself?)*

Arg2: *Nereye gidebilir yaşadıklarını bırakıp? (Where could she go leaving what has been lived?)*

DR Type: *Implicit, Level-1 sense: Expansion, Level-2 sense: Conjunction*

There are 6 DR realization types, 4 Level-1 senses and 24 Level-2 senses selected from the list in Appendix A. All DRs have a realization type label (i.e., Explicit, Implicit, etc.); Explicit, Implicit,

¹ Penn Discourse Treebank Version 3.0 <<https://catalog.ldc.upenn.edu/LDC2019T05>>

² Unlike the PDTB, TDB has not annotated attribution.

and AltLexes are assigned a sense tag, amounting to 3467 tokens. That is, the DRs annotated as NoRel, EntRel and Hypophora, amounting to a total of 520 DRs (13% of TDB 1.2) are not assigned a sense label, as explained in Section 2.1.1.

3.1.2 Chinese Discourse Bank

[20] have created the TED-CDB dataset from a large set of TED talks in Chinese that have been manually annotated according to the goals and principles of the Penn Discourse Treebank and adapted certain features that are not present in English. It is claimed that the TED-CDB can improve the performance of systems being developed for languages other than Chinese and would be helpful for insufficient or unbalanced data in other corpora. The TED-CDB includes a total of 11,975 annotations and the numbers of available tokens in terms of the DR type of realization and senses are given in Tables 2 and 3, respectively.

3.1.3 Penn Discourse Treebank 3.0

Penn Discourse Treebank has been the largest text corpus to date manually annotated for discourse relations [12]. The latest release, the PDTB 3.0, is an enriched version of the PDTB 2.0 with the addition of 13K new annotations, which are mainly intra-sentential relations that were not annotated in the previous edition [2]. The PDTB 3.0 includes a total of 53,631 annotations and the numbers of available tokens in terms of the DR type of realization and senses are given in Tables 2 and 3, respectively.

3.2 A Snapshot of the Methodology of the Three Phases of the Thesis

The present thesis has developed through three major phases, referred to as Study 1 (Chapter 4), Study 2 (Chapter 5) and Study 3 (Chapter 6). While Study 1 involved machine learning experiments, Study 2 involved experiments with PLMs (namely, USE and the BERT Turkish PLM, which will be detailed in the upcoming chapters). In Study 3, in order to reduce the disadvantages caused by the lack of data in Turkish and to observe the effect of augmenting the number of annotated discourse relations on the computational models, a multilingual dataset has been produced with Cross-lingual Transfer Learning and fine-tuned.

3.2.1 The Multilingual Dataset

The research conducted in the literature using modern techniques and various datasets has aimed to produce solutions to alleviate the lack of annotated data. One of the best short term solutions to this issue could be the method of augmenting the size of the training dataset by the concatenation of datasets in other languages. Multilingual datasets have been helpful not only in translation but also in various NLP tasks. It has also been shown that a multilingual implicit DR classifier transfers well across dissimilar languages [72].

The multilingual dataset has been constructed by merging TDB 1.2 (Turkish) (Section 3.1.1), TED-CDB (Chinese) [20] (Section 3.1.2) and the PDTB 3.0³ (English) (Section 3.1.3), as all of these datasets use the PDTB-style annotation.

By merging these three datasets, a big corpus of 67,431 DRs is formed. The total numbers and the distribution of DR type categories of each language are listed in Table 2. English is the largest dataset and it outnumbers the sum of Chinese and Turkish as more than two times. A positive side effect of this unbalanced structure was that it made it possible to test the language independent context-aware encoding capability of multilingual PLMs.⁴

Table 2: Annotation statistics of **DR Realization Types** in TDB 1.2, TED-CDB, and PDTB-3 (*the Multilingual Dataset of the work*)

DR Type	TDB 1.2 (Turkish)	CDB (Chinese)	PDTB 3.0 (English)
Explicit	1,524 (38.2%)	4,309 (36.0%)	24,240 (45.3%)
Implicit	1,791 (44.9%)	5,656 (47.2%)	21,782 (40.7%)
AltLex	152 (3.8%)	749 (6.3%)	1,498 (2.8%)
Hypophora	80 (2.0%)	229 (1.9%)	146 (0.3%)
EntRel	237 (5.9%)	655 (5.5%)	5,538 (10.3%)
NoRel	203 (5.1%)	377 (3.2%)	287 (0.5%)
Total	3,987 (5.9%)	11,975 (17.8%)	51,469 (76.3%)

The distribution of DRs among DR type categories across the multilingual dataset is also highly unbalanced and the ratios of the categories in all languages are very close to each other. The most obvious instances are the Explicit and Implicit categories, which occur in more than 80% of the datasets.

The statistics of 60,095 DRs (out of 67,431 DR realizations) annotated with Level-1 sense categories in the respective languages are listed in Table 3.

Table 3: Annotation statistics of **Level-1 senses** in TDB 1.2, CDB, PDTB-3 (*the Multilingual Dataset of the work*)

Level-1 sense	TDB 1.2 (Turkish)	CDB (Chinese)	PDTB 3.0 (English)
Comparison	448 (12.9%)	1,568 (14.6%)	8,399 (18.3%)
Contingency	702 (20.3%)	3,028 (28.3%)	11,503 (25.0%)
Expansion	1,700 (49.0%)	4,237 (39.5%)	20,266 (44.1%)
Temporal	617 (17.8%)	1,881 (17.6%)	5,874 (12.7%)
Total	3,467 (5.8%)	10,714 (17.8%)	45,914 (76.4%)

3.3 The Conduct of Experiments

The techniques used in the thesis range from classical linguistic feature extraction NLP approaches to text encoding methods based on PLMs, all moving towards demonstrating the state-of-the-art devel-

³ Penn Discourse Treebank Version 3.0 <<https://catalog.ldc.upenn.edu/LDC2019T05>>

⁴ The AltLexC type of DRs annotated in the PDTB 3.0 are eliminated as they are not annotated either in the TDB 1.2 or the TED-CDB.

opments for DR realization type identification and DR sense classification. This section introduces the conduct of experiments in Study 1 as well as Study 2 and Study 3, which use PLMs.

All along the thesis work, research and experiments are based on (i) an approach that involves text processing and feature extraction methods of NLP, (ii) an approach of encoding text data with three different sentence level encoding PLMs (resolving the sparsity problem as well), eventually evolving into (iii) an approach that involves a NN based multiple classification architecture for training and testing.

The present research started with Study 1, with the concept of Count-based Vector Space Models (i.e. Count Vectorization), where we relied heavily on machine learning to extract useful linguistic features from the arguments of DR realizations and apply legacy classification methods to retrieve solid results (see Chapter 4 (p. 31)) so that they can be compared with those of advanced methods.

In Study 2 and Study 3, we have been concerned with encoding of the data with PLMs (see Chapter 5 (p. 39)). So, firstly, USE is used in Study 2 for encoding TDB and the output is tested with various legacy classification algorithms. Secondly, the DRs are encoded with BERT Turkish PLM (both in Study 2 and 3), where we fine-tuned the best performing BERT models for our tasks.

Why BERT?

BERT is used for obtaining the intrinsic representation of each word as the label-specific word representation while preserving the properties of the text (syntactic, contextual, etc.) in order to learn the labels over DRs. That is, the thesis eventually evolved into an approach based on encoding of TDB with BERT PLMs and fine-tuned for the best performance in the identification of DR realization types and senses. Both the classification of DR realization types and Level-1 senses are done in an experiment where TDB 1.2 was encoded with the BERT Turkish PLM (see Chapter 5). A cross-domain experiment has been realized as well.

In order to learn the labels over discourse relations, the BERT PLMs are used, taking advantage of their capacity to obtain the intrinsic representation of each word while preserving the linguistic properties of the text. Visualizing Data using the Embedding Projector in TensorBoard ⁵ is rehearsed for interpreting the embedding by using metadata that allows for visualization of a specific layer of interest in the model. TensorBoard is a tool for providing the measurements and visualizations needed during the machine learning workflow. It enables tracking experiment metrics like loss and accuracy, visualizing the model graph, projecting NLP embeddings to a lower-dimensional space by reading tensors and metadata from the logs of the Tensorflow projects, i.e. setup a 2D tensor that holds TDB embeddings. Taking all the dataset at once, the dashboard allows users to search for specific terms, and highlights words that are adjacent to each other in the embedding (low-dimensional) space, by the help of euclidean distance formula.

As examples for BERT Turkish PLM embedding, see the embeddings plotted for two words from TDB 1.2 in Figures 3 and 4, where we can see that “burnunu” (his/her nose) and “iğrenç” (disgusting) are both rather neutral terms to each other and they are referenced in very different contexts. The noun “burnunu” is a typical sample that demonstrate the referencing with different encodings in different contexts and its semantically closed terms are some other words about body parts. However, the

⁵ https://www.tensorflow.org/tensorboard/tensorboard_projector_plugin?hl=en

adjective “iğrenç” seems to be encoded as a very effective embedding that has semantically various close terms, ranging from nouns to some other adjectives.

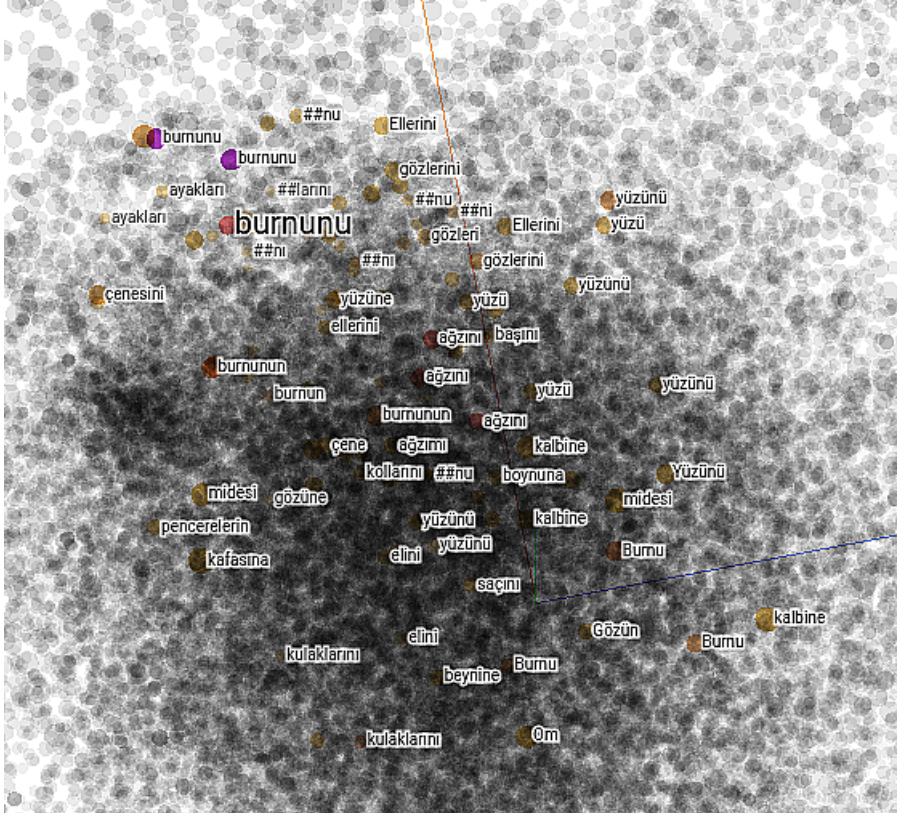


Figure 3: The TensorBoard view of the Turkish BERT encodings for the noun “burnunu” (his/her nose).

In the search of optimal classification method for DR realization type and sense identification problems, all the experiments are modelled as multi-way classification tasks. As we do not assume a separate identification of arguments, each discourse relation is presented as one sentence to BERT; hence, in practice, the tasks are modelled as sentence classification instead of the standard approach of sentence pair classification.

For each dataset, 10% of the data is allocated as the validation set and another 10% as the test set. The distribution of labels in each set is provided in the Tables 9 and 10. For each task, we fine-tune BERT following the standard practice suggested by [50]. Simply, the [CLS] token is used as the representation of the whole sequence and fed into a fully-connected dense layer with a softmax activation. In order to account for the variance due to random initialization and stochastic training during fine-tuning, we fine-tune each model four times and report the average performance in the next section.

Following the previous work in the literature, in BERT, we set the maximum sequence length to 128. We use AdamW optimizer [73] with the learning rate of $5e-5$. We also apply a learning rate warm-up where the learning rate is linearly increased from 0 to $5e-5$ over the first 10% of iterations and then, linearly decreases to 0. The models are fine-tuned for 50 epochs, with the batch sizes of 16. We apply

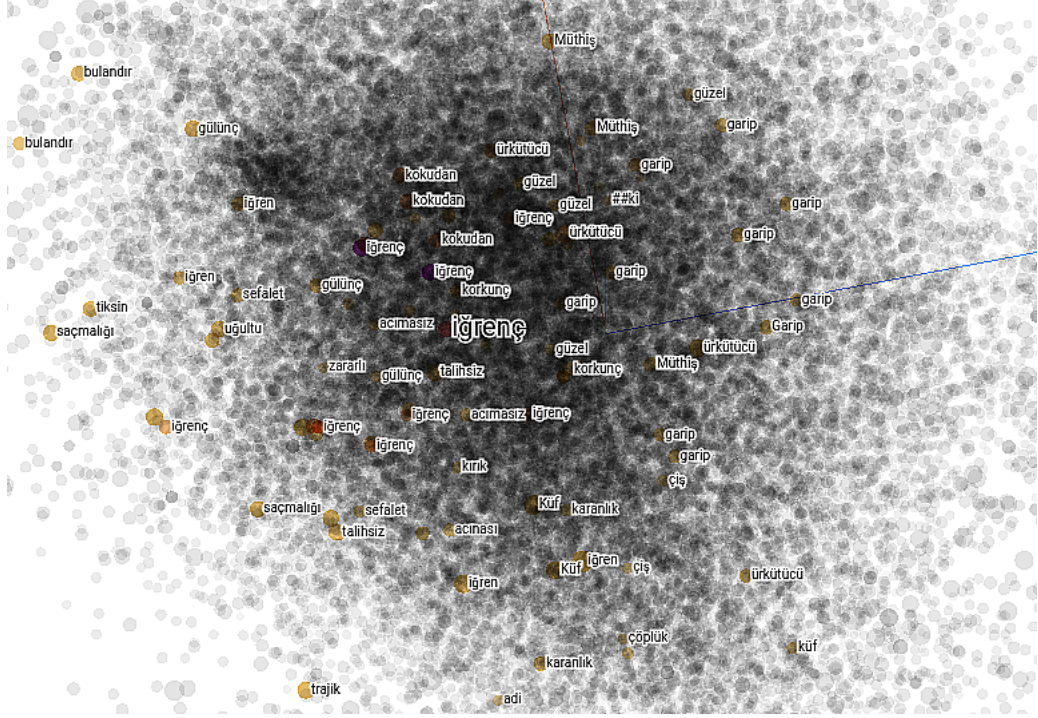


Figure 4: The TensorBoard view of the Turkish BERT encodings for adjective “iğrenç” (disgusting).

EarlyStopping⁶ with patience⁷ of 25 according to the performance on the development set. The models are evaluated four times in each epoch and the one with the best development performance is stored as the final model. For each experiment, we report performance of our models on the set via accuracy, recall, precision and F1-Score⁸, calculated as:

$$Recall = \frac{TP}{TP + FN} \quad Precision = \frac{TP}{TP + FP} \quad (1a)$$

$$F1 = \frac{2 * Recall * Precision}{(Precision + Recall)} \quad (1b)$$

All the training is performed on a single T4 GPU. The experiments for the development of a classification model gave rise to the *Bert_MultiClass* model, built based on the *transformers.TFBertModel* class⁹ (More details about the classification model, the method of fine-tuning and the tests are provided in Appendix B).

The results and the discussions will be introduced in later parts of the thesis. The next chapter explains our early experiments based on linguistic feature selection (Study 1), then Study 2 and 3 will be presented together with the results.

⁶ https://keras.io/api/callbacks/early_stopping/

⁷ Number of epochs with no improvement after which training will be stopped.

⁸ https://scikit-learn.org/stable/modules/model_evaluation.html

⁹ https://huggingface.co/docs/transformers/v4.15.0/en/model_doc/bert#transformers.TFBertModel

CHAPTER 4

STUDY I: EARLY EXPERIMENTS ON LINGUISTIC FEATURE SELECTION

Since the beginning, feature engineering has been an integral part of NLP, where we extract a numerical representation, which we call feature, of text according to a set of linguistic rules. Those representations are fed into a learning algorithm or model, though the popularity of feature engineering, especially in the academic setting, might be declining, due to the emergence of deep learning, which promises to automatically extract suitable representations for an NLP task.

The diversity level of the natural language data is better represented if all the labels in the data are targeted. In the early phases of the research, working on TDB 1.1, we wanted to target the hardest level by creating multilabel classes produced by joining DR realization types and the Level-1 sense labels, ending up with 15 different classes as shown in the Table 4.

Table 4: DR Annotations in TDB 1.1: A Summary

Types	Number	Senses	Number	DR Classes	Number
Explicit	1,155	Comparison	243	Explicit Comparison	243
		Contingency	227	Explicit Contingency	227
		Expansion	402	Explicit Expansion	402
		Temporal	283	Explicit Temporal	283
Implicit	1,540	Comparison	140	Implicit Comparison	140
		Contingency	247	Implicit Contingency	247
		Expansion	1,013	Implicit Expansion	1,013
		Temporal	140	Implicit Temporal	140
AltLex	139	Comparison	16	AltLex Comparison	16
		Contingency	48	AltLex Contingency	48
		Expansion	44	AltLex Expansion	44
		Temporal	31	AltLex Temporal	31
NoRel	218			NoRel	218
EntRel	270	NoSense	567	EntRel	270
Hypophora	79			Hypophora	79

Annotation statistics of 3401 DRs in TDB 1.1. types and senses are listed separately, and joined to form full DR Class labels.

The histogram chart of the frequencies of DR Class tokens in the TDB 1.1 is shown in Figure 5. Even with the concatenation of Level-1 senses with the DR types, the AltLex type of DR realizations are always the group of classes with the least number of tokens in the corpus. The figure also reveals an unbalanced distribution of the DR Class (type and tense) frequencies, which formed one of the main hindrances to improve the performance in the study. In fact, when we targeted multilabel multiclass classification by including 24 Level-2 senses, reported in Table 5, there happened to be no classification performance worth to be reported. The poor performance levels of this exercise over TDB 1.1 have lead us to consider different methods and different pipelines for the discourse parsing of Turkish. Most importantly, it appeared important to do away with the data scarcity problem at hand. Thus, we considered the use of multilingual datasets and Cross-lingual Transfer Learning, as explained below.

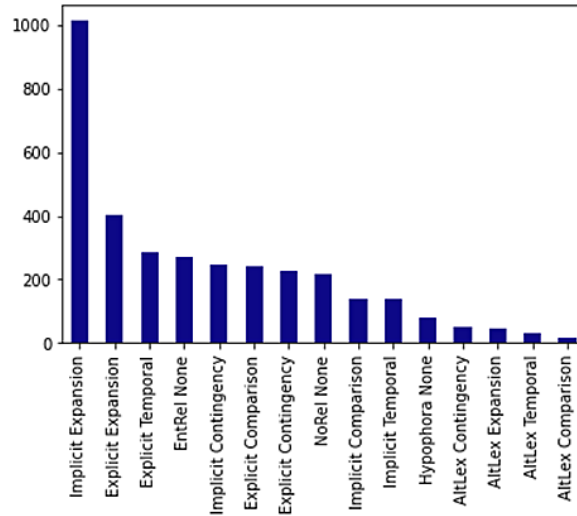


Figure 5: Number of DR Class Tokens in TDB 1.1

Table 5: Annotation statistics of Level-2 senses

Level-2 senses	Num.	Level-2 senses	Num.	Level-2 senses	Num.
Asynchronous	332	Conjunction	962	Instantiation	64
Cause	444	Conjunction+scale	5	Level-of-detail	432
Cause+Belief	57	Contrast	152	Manner	123
Cause+SpeechAct	51	Correction	24	Negative-condition	10
Concession	236	Degree	3	Purpose	86
Concession+SpeechAct	18	Disjunction	18	Similarity	39
Condition	44	Equivalence	44	Substitution	17
Condition+SpeechAct	10	Exception	11	Synchronous	285

In order to assess the effect of linguistic structure incorporation into the model and to test feature engineering, limited number of available (annotated) data scenario is useful. Given the limited data available in Turkish discourse, in Study 1, we started from the concept of Count-based Vector Space Models, where we relied heavily on machine learning to extract useful features from raw data using math, statistics and domain knowledge. Thus, prior to our work with the BERT PLM and Cross-transfer Learning, we applied a kind of feature engineering to the task of DR Class identification by following the steps below in a series of experiments run on an earlier version (1.1) of TDB:

- Process (i.e. tokenization, stemming, lemmatization and part of speech tagging) the arguments of DR realizations with their connectives (if available) and extract linguistic features as explained below,
- Produce DR-Features Matrix (where each DR is given in a row and each feature fill a column with their frequencies in the DRs),
- Train a model by applying the universal supervised learning approach (Figure 6).

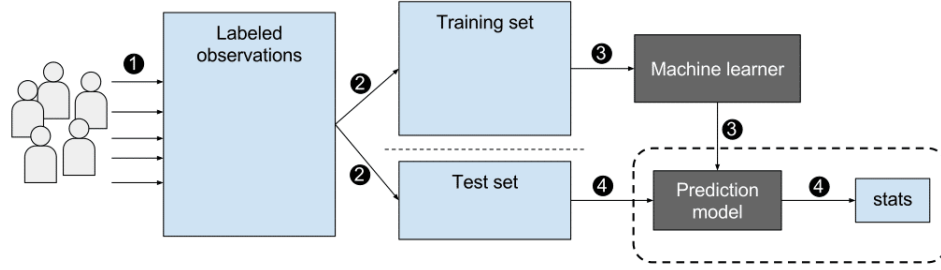


Figure 6: Supervised Learning Approach

4.1 Analysis of the Best Features

The first step of the experiments was the identification of Explicit relations (as they are easily identified) and distinguish them from non-Explicit relations (i.e., Implicit Relations, Entity Relations, Alternative Lexicalizations, No Relations). Thus, given a connective, the automatic determination of the associated Explicit relation is searched in TDB 1.1. Going over statistical techniques, the first step involved a regression analysis to formulate the model and analyze the relationship between the Explicit relations (dependent variable) and the selected features (independent variables) with the help of hypothesis testing.

In order to conduct a multiple regression analysis and to form a linear regression equation three different type of features are retrieved from the DRs:

- Position of the DC (at the beginning, between the arguments and at the end of a relation, labeled as HEAD, MIDDLE, END respectively),
- Number of the arguments,
- Frequencies of the POS tags of words (see Figure 7) in the arguments ¹

Obviously, DCs are the main parts of Explicit DRs and the presence of a DC guarantees the explicitness of a DR. Thus, DCs are highly correlated with the target value of being Explicit DR as shown by the correlation analysis depicted in Figure 8. In order not to cause the linear regression equation over-fit, DCs are not taken as features in the experiments.

¹ POS Tags are retrieved by the ZEMBEREK NLP Library <<https://github.com/ahmetaa/zemberek-nlp>>

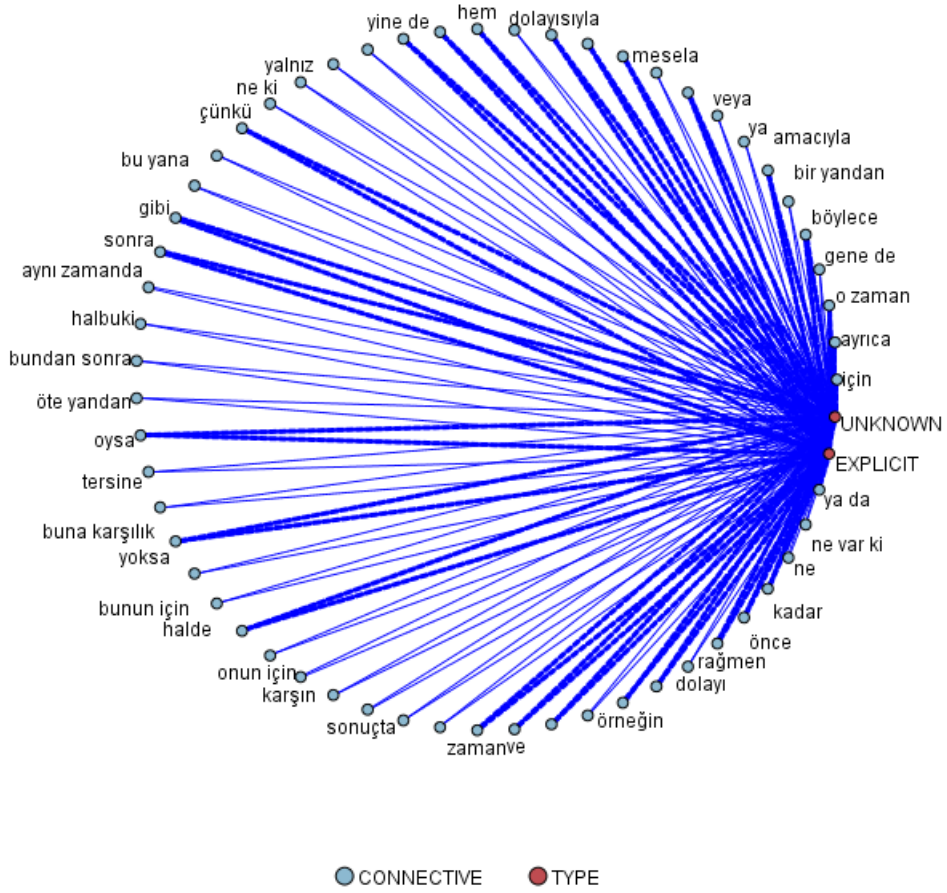


Figure 8: Correlation levels of DCs to Being an Explicit DR
(The magnitude represented by the thickness of the line.)

RELATION	CONNECTIVE	POSITION	RANGE	VERB*DB	INTERJ	ADJ*DB	ADJ	PRON	NUM	CONJ	DATE	NOUN*DB	DUP	PUNC	VERB
EXPLICIT	aksine	MIDDLE	0	3	0	0	1	0	0	0	0	0	0	0	3
EXPLICIT	bunun aksine	MIDDLE	0	2	0	0	2	0	0	0	0	0	0	0	2
EXPLICIT	aksine	MIDDLE	0	2	0	1	5	1	2	1	0	0	0	0	4
EXPLICIT	aksine	MIDDLE	0	0	0	0	2	0	1	0	0	0	0	0	3
EXPLICIT	aksine	MIDDLE	0	0	0	1	1	0	0	0	0	0	0	0	2
EXPLICIT	aksine	MIDDLE	0	0	0	0	4	0	0	0	0	0	0	0	2
EXPLICIT	aksine	MIDDLE	0	1	0	1	4	0	0	0	0	0	0	0	5
EXPLICIT	aksine	MIDDLE	0	2	0	3	4	1	1	1	0	0	0	0	9
EXPLICIT	aksine	MIDDLE	0	0	0	1	0	0	0	0	0	0	0	0	2
EXPLICIT	aksine	MIDDLE	0	1	0	0	0	0	1	0	0	0	0	1	5
EXPLICIT	aksine	MIDDLE	0	0	0	3	2	0	0	0	0	0	0	0	4
EXPLICIT	aksine	MIDDLE	0	1	0	0	0	0	0	0	0	0	0	0	6

Figure 9: POS Tag Data Bag of Words Data Matrix Sample Portion

verbs are the strongest linguistic features for a relation, whereas adjectives, numbers, determiners,

adverbs, conjunctions and pronouns are at the medium level. All other POS tags are detected to be weak in the regression. The two positions of the DC and the number of arguments, have weak and medium level of powers respectively².

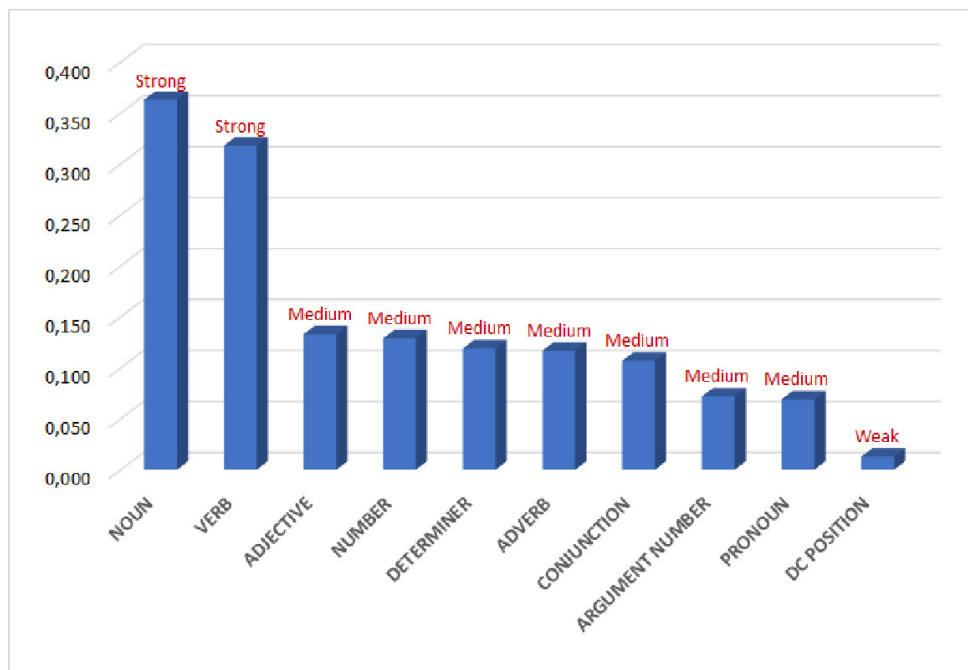


Figure 10: Absolute Values of the Linear Regression Coefficients of Selected Features

Firstly, the regression model is tested in a binary classification model for detecting Explicit DRs in TDB 1.1 and the results are as given in Table 6. Reaching 68.6% as True-Positive (Explicit DRs), the model achieved an accuracy of 76.4% in average. From another view, detecting the 80.4% of Non-Explicit DRs correctly, the model seemed to be stronger in detecting the absence of an Explicit DR. As its Recall is 64.3%, Precision is 68.6% and F1-Score is 66.4%, the model has worked for the Explicit/non-Explicit DR classification task.

Table 6: Evaluation of Explicit DR Identification on TDB 1.1 by Linear Regression

DR Type	Explicit	Non-Explicit	Accuracy
Explicit	792	363	68.6
Non-Explicit	440	1,806	80.4
Average Accuracy			76.4
Recall			64.3
Precision			68.6
F1-Score			66.4

Analysis is done over the data matrix of POS tags counts extracted as features from TDB1.1

² This result is parallel to the annotations and not very surprising, because, during the annotation stage, the DC role of a connective is determined by checking whether the text spans have an "abstract object" interpretation. An abstract object interpretation often correlates with a clause, where the predicate is either nominal or verbal.

Despite the relatively successful result of Explicit/non-Explicit distinction, the 15-way multiclass Class classification experiments (where DR realization type and senses were combined) ended up with very low performance. Thus, we went on classification of DR realization types by utilizing the data matrix of the features (sampled as in Figure 9) with the most appropriate supervised learning multi-class classification algorithms. The best score was reached by the C5 Decision Tree Classifier – e.g., classification of DR realization types, achieved by F1-Score of 0.33 (Accuracy 35%).

In order to increase the success level, an enhancement of the dataset was attempted with new tags with more explanatory information. The agglutinative nature of Turkish was used and the morphological analysis of the words in the arguments were added as features. For that, the morphological analysis outputs produced by the java library of "An Open, Extendible, and Fast Turkish Morphological Analyzer"[74] as shown in the Example 4.1.1.

Example 4.1.1 *Sample Argument: Yarın kar yağacak* (Tomorrow it will snow)

yarın: **Noun+A3sg**

kar: **Noun+A3sg**

yağacak: **Verb+acak:Fut+A3sg**

Classification experiments were run with the same algorithms conducted over the data-matrix and again, the best scores were reached by the C5 Decision Tree Classifier – e.g., classification of DR realization types, achieved by F1-Score of 0.36 (Accuracy 39%). So, even though the data set was augmented by adding features from the agglutinative nature of Turkish, this attempt has contributed very little improvement – e.g., +0.3 in F1-Score in classification of DR realization types. So, with the clear view of the requirement of appropriate encoding mechanisms, parallel to the developments in the area, neural network based encoding mechanisms are researched in order to extract the potential discourse information in the text.

The next chapter explains Study 2, where we encoded the TDB dataset with USE and BERT Turkish PLM for our discourse tasks, where we abandon the idea of 15-way multi-class classification and attempt to classify DR realization types and their senses separately.

CHAPTER 5

STUDY II: ENCODING TDB DATASET WITH USE AND MONOLINGUAL BERT FOR DR IDENTIFICATION AND SENSE IDENTIFICATION EXPERIMENTS

In Study 2, experiments have been carried out to improve the DR realization type identification task with PLMs, and a separate sense identification experiment has been run. Although several experiments have been done, only the most prominent experiments will be reported here. The overview of the conducted experiments are summarized in Table 7 (in compliance with Figure 1) and detailed in the rest of the chapter. In general, the experiments of Study 2 can be considered as precursors of the experiments in Study 3. So, the final experiment series applied on the multilingual dataset in order to realize Cross-lingual Transfer Learning (Study 3) (see Chapter 6) are also summarized in the table to provide a complete picture of Study 2 and Study 3, which have yielded more satisfactory results than Study 1.

Table 7: Overview of the Experiments

Experiment #	Encoding PLM	Dataset	Purpose
Section 5.1	USE	TDB 1.2	Apply legacy classification algorithms over the encoded TDB data in order to form a baseline performance level in DR type and sense classification.
Section 5.2	Turkish-BERT	TDB 1.2 T-TED-MDB	Explore the effectiveness of monolingual BERT for both in-domain (TDB) and out-domain datasets (TED-MDB) in the classification of DR realization types and senses.
Chapter 6	Turkish-BERT Chinese-BERT English-BERT Multilingual BERT	TDB 1.2 TED-CDB PDTB 3.0	Explore the effectiveness of multilingual BERT encoding on multilingual data aggregation, as compared to monolingual models for the DR type classification task.

Since our attempts to improve the success level of DR Classes by linguistic features have not proven useful, we turned to works related to PLMs for finding the best solution for the thesis research ques-

tions, and solve our two major tasks, namely, DR realization type identification and sense identification, taken as separate targets. Starting with a lightweight version multilingual model (USE), we also experimented with the BERT Turkish PLM and the Multilingual BERT. In this chapter, we report and discuss the experiments results of USE and BERT Turkish PLM .

5.1 Encoding TDB 1.2 with USE for Testing Multiple Classification Algorithms

Besides improving the performance of our discourse parser that better incorporates the nature of the data into the approach, the main purpose of this Study is two folds: (i) to evaluate the contribution of encoding with a lightweight PLM (i.e., USE) to the DR realization type and sense identification problems, (ii) to find a new direction for our research and obtain noteworthy performance gains by assessing the potential of encoding.

The first attempt to use a PLM started with an experiment in which TDB 1.2 was encoded with USE, and the relations' arguments were mapped to a fixed-length vector representation with the maximum vector size of 512. That is, a maximum vector size of numbers is created for each input retrieving a 2D data matrix for the corpus of size (n x 512), where n stands for the number of the realized DR tokens and 512 is the maximum vector size. One more column is added for the DR realization category labels of DR realization types and senses listed in Table 1.

As the text encoder, the PLM *universal-sentence-encoder-multilingual*¹, trained on 16 languages (*Arabic, Chinese-simplified, Chinese-traditional, English, French, German, Italian, Japanese, Korean, Dutch, Polish, Portuguese, Spanish, Thai, Turkish, Russian*), is used. An assembly of six prominent classifiers, involving the Support Vector Classifier (SVC), Gaussian Naive Bayes (GaussianNB), K-nearest Neighbor (KNN), Feed Forward NN, Convolutional NN (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) are run on the same data matrix of encoded text produced by USE.

A six-way classification of TDB DR types and a five-way classification of their Level-1 senses were targeted. The accuracy levels of the algorithms are listed in Table 8. The table reveals that the NN based approaches outperform legacy machine learning algorithms in the six-way DR type classification and a five-way sense classification. Due to the small size of the dataset, the differences are very small but there are slight accuracy improvements, showing that the BiLSTM classification reached the best results with 61% for type and 58% for tense classification.

Table 8: Accuracies of the Classification Algorithms on the TDB 1.2

CL	Category #	SVC	GaussianNB	KNN	FFNN	CNN	BiLSTM
DR Types	6	0.561	0.487	0.441	0.563	0.469	0.611
DR Senses	5	0.503	0.473	0.231	0.521	0.453	0.578

TDB 1.2 is encoded with USE PLM

CL: Classification Level; SVC: Support Vector Classification; GaussianNB: Gaussian Naive Bayes; KNN: k-nearest Neighbors Algorithm; FFNN: Feed Forward Neural Network; CNN: Convolutional Neural Network; BiLSTM: Bidirectional long short-term memory

¹ <https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/>

Naturally, the accuracy of multi-class classification decreases by the increase in the class number. But this is not the case with this experiment, even though the number of categories in DR sense classification (5) is one less than that of DR type classification (6). The results imply that distinguishing among the DR senses is far more complicated than the DR types. In the best results yielded by BiLSTM, the accuracy of DR sense classification is 3.3 % less than that of DR type classification, which has more number of classes. This shows that DR sense classification on its own is still an important problem to be resolved in discourse research.

5.2 Experiments Based on Encoding TDB 1.2 with Turkish BERT PLM

5.2.1 Experimental Setup

All the experiments based on encoding with BERT PLMs are conducted according to a standard rule set as detailed in this section. In order to learn the labels over DRs, we used BERT, in the architecture shown in Figure 11, to obtain the intrinsic representation of each word as the label-specific word representation while preserving the properties of the text (syntactic, contextual, etc.).

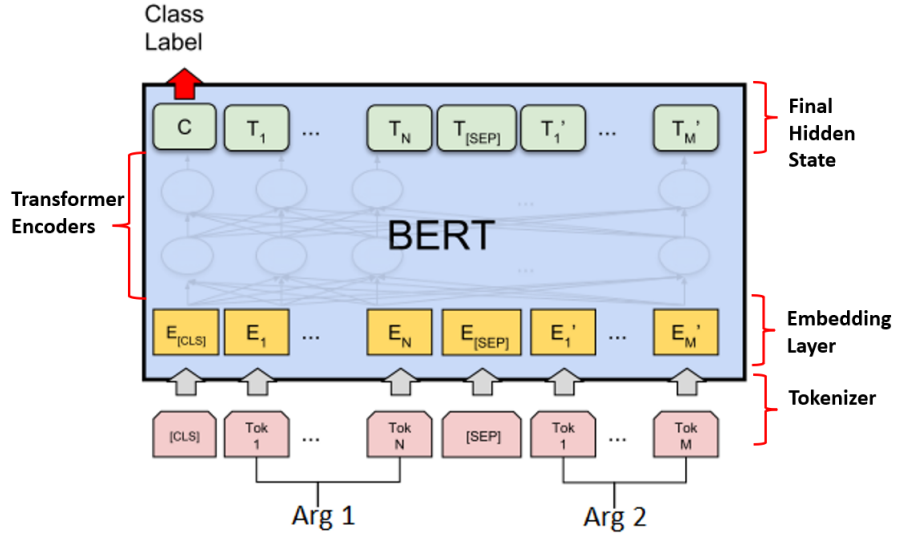


Figure 11: BERT Classification Architecture

Using the cross validation approach to address problems related to the small corpus size [75, 7], classification experiments with BERT are conducted through the infrastructure depicted in Figure 13, and through the following steps:

- For each DR, merge the arguments and (any available) discourse connective into a form of single line text as labelled with its relation labels (type, Level-1 sense, etc.).
- Convert sentences into segments by adding [CLS] (class) and [SEP] (separator) tags.
- Split sentences into tokens with the BERT tokenizer and convert the tokens into indexes of the tokenizer vocabulary.

- Pad or truncate the sentences into the maximum length long vectors.
- Create an attention mask.
- Return a dictionary of outputs.
- Convert each tokenized DR vector into a tensor and create embedding (see Example 5.2.1).

Example 5.2.1 DR ID: TDB 1.1.1850

Arg1: *Halk tarafından iyi görülmediler.* (They were not accepted as good by the public.)

Arg2: *Halbuki, onun yerine fıkıh hükümleriyle tanzim edilerek tatbik olunsalardı halk bunları hiç itirazsız kabul ederdi.* (However, if they were arranged and applied with religious law provisions instead, the people would have accepted them without any objection.)

Type: *AltLex*, **Sense:** *Expansion*, **DR Class:** *AltLex.Expansion*

Token IDs: tensor[2, 2987, 2418, 2370, 8178, 1985, 4150, 13314, 2864, 3097, 23661, 3680, 35238, 2308, 29859, 8190, 19956, 4981, 74909, 2987, 4900, 9416, 7198, 2599, 2948, 2831, 3, 0, 0, 0, ...]

- Train with *Bert_MultiClass* model with the main hyper-parameters of Batch size: 32, Learning Rate (Adam): 5e-5 and Number of Epochs: 9. The architecture always resulted with the normally expected loss decrease and accuracy increase as shown in Figure 12, standing for the realization of sound training and testing.

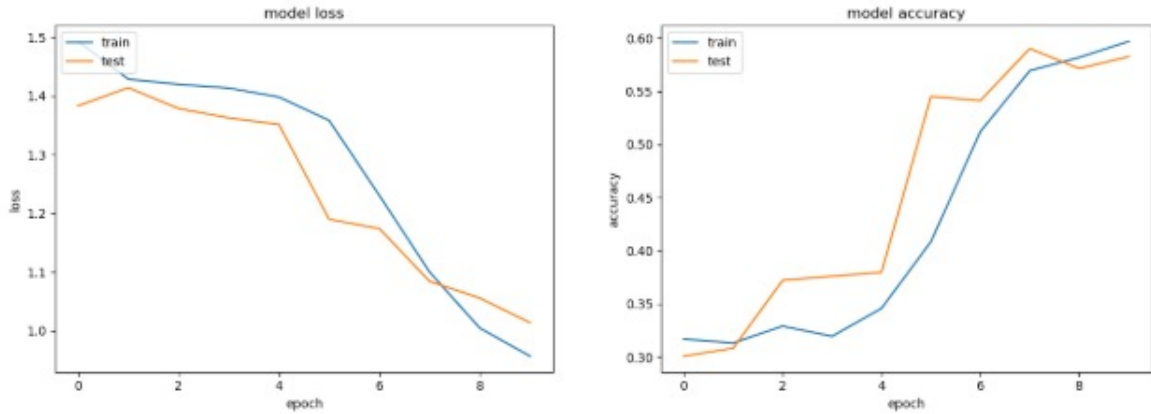


Figure 12: The plotting of loss and accuracy values during the experiments for fine-tuning of Turkish BERT Model over TDB 1.2

Searching for a better classification model in each trial, the *Bert_MultiClass* model was built by the inheritance from the *transformers.TFBertModel* class² as depicted in Figure 20 and detailed in Table 22 of Appendix B.

The TDB 1.2 dataset is fed into BERT, working with the transformer PLM for Turkish (*bert-base-turkish-128k-uncased*) (referred to as the monolingual model throughout the thesis), released in Hugging Face library. The library contains the PyTorch implementation of the state-of-the-art NLP models

² https://huggingface.co/docs/transformers/v4.15.0/en/model_doc/bert#transformers.TFBertModel

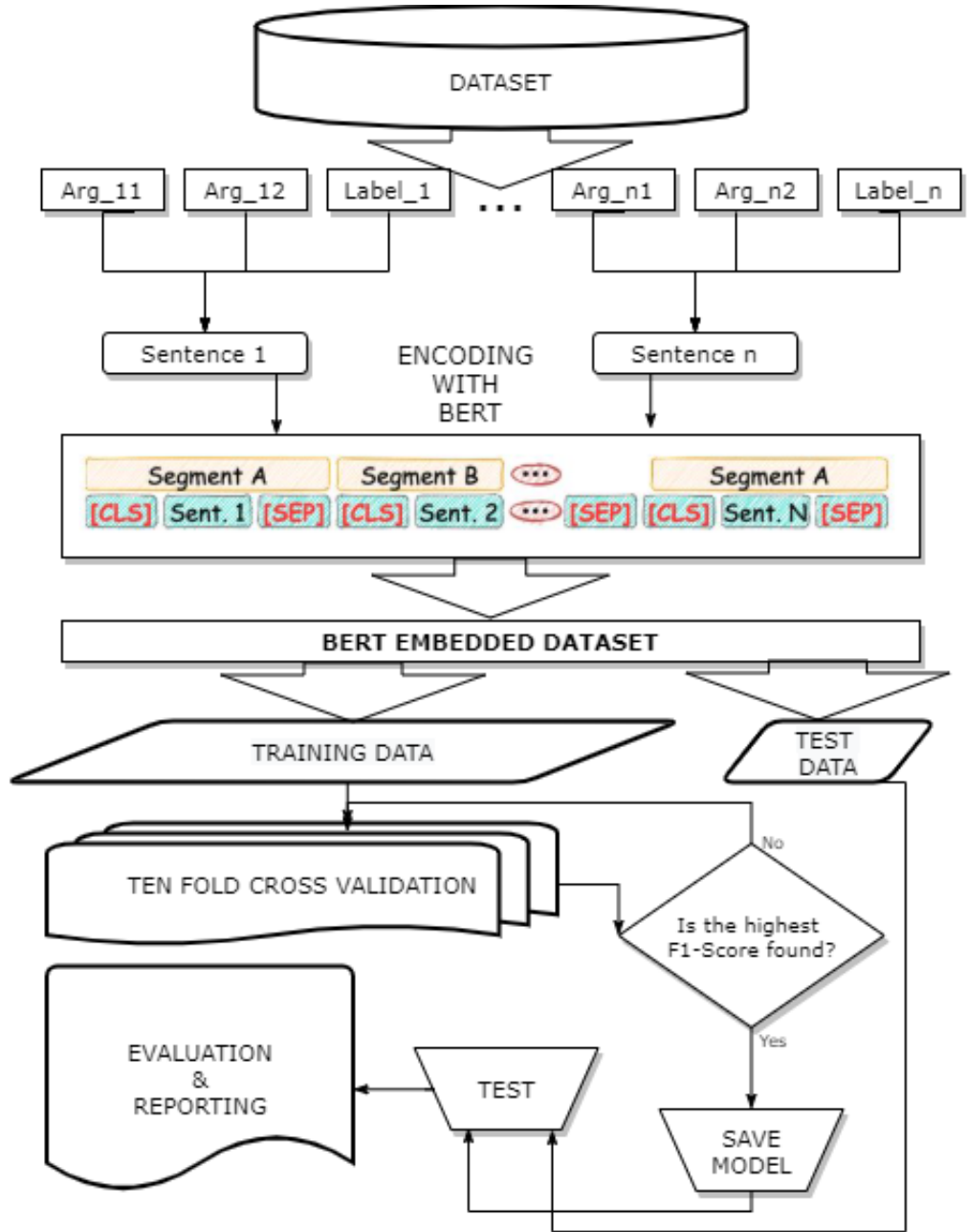


Figure 13: The Workflow of the Experiments Conducted by Encoding with BERT Models

including BERT and PLM weights. Taking the TDB 1.2 dataset as the input, classification of dataset against DR realization types and senses separately in the same workflow (see Figure 13).

A summary of the train, development and test splits of the TDB 1.2, used in the experiments with BERT, is presented in the tables below: The distribution of the labels in the DR realization type classification experiments is provided in Table 9 and the distribution of the Level-1 sense labels of the explicit and implicit type of relations is given in Table 10. As it will be explained below, the table of

Level-1 senses are important as they will be incorporated in a four-way classification involving explicit and implicit classifiers separately.

Table 9: The distribution of labels in the DR realization type classification experiments

TYPE	Train	Dev	Test
AltLex	113	13	20
EntRel	196	24	13
Explicit	1157	147	163
Hypophora	68	2	7
Implicit	1394	186	163
NoRel	167	15	21

Table 10: The distribution of labels in the DR sense classification experiments of Explicit and Implicit type of DR realizations

SENSE	Explicit			Implicit		
	Train	Dev	Test	Train	Dev	Test
Comparison	205	23	31	130	15	17
Contingency	206	29	33	264	35	34
Expansion	429	57	54	879	118	93
Temporal	317	38	45	121	18	19

The results of the phase, where the classification experiments run on the encoded TDB 1.2 by BERT, are provided in Table 11, revealing how BERT outperforms the algorithms described in the first phase with USE. Encoding and classification with BERT reached the accuracy level of 77% for type and 68% for sense classification, showing that the performance of the model is far better than USE (see Section 5.1).

Table 11: DR Classification Evaluation of TDB 1.2

CL	Category #	F1-Score	Recall	Precision	Accuracy
DR Types	6	0.596	0.565	0.654	0.765
DR Senses	5	0.559	0.553	0.566	0.675

CL: Classification Level of DR

TDB 1.2 is encoded with BERT Turkish PLM

Classification testing on monolingual BERT-coded data caused prominent improvements such that the accuracy level of DR type classification increased by 15.4% and DR sense by 9.7%. Also the F1-Scores provided in Table 11 are over 50%, confirming that the results of the experiments are solid improvements for the task, especially despite the scarcity of the training and testing data.

5.2.2 Discourse Relation Realization Type Classification

As defined in Section 2.1.2, DR realization type classification focuses on identifying how precisely discourse relations are realized in a given text span, given the PDTB’s five relation realization types

(implicit, explicit, AltLex, Hypophora, EntRel). If no such relation is found, the model is supposed to label the text as having a NoRel relation, mimicing the PDTB annotation style. The results of our experiments are provided in Table 12, at first sight revealing that the Turkish BERT model achieves almost 74% accuracy and an F1-Score of 0.77 over all relations. The relatively low F1-Score suggests that the task is more challenging than it looks; however, it must be highlighted that the model does not have access to any information regarding the argument boundaries or whether there is a connective in the text span or not. Hence, we find the results to be promising.

Table 12: DR realization type classification results over the TDB 1.2

DR Type	F1-Score	Recall	Precision
AltLex	0.63	0.58	0.69
EntRel	0.32	0.40	0.32
Explicit	0.90	0.89	0.92
Hypophora	0.74	0.79	0.70
Implicit	0.77	0.78	0.76
NoRel	<u>0.11</u>	0.11	0.12
Accuracy (%)	73.90		
F1-Score	0.77		

TDB 1.2 is encoded with BERT Turkish PLM

Of all relation realization types, the explicitly realized ones turn out to be the most easily identifiable relations with the F1-Score of 0.9. This result suggests that the model learns to recognize discourse connectives. On the other hand, the EntRels and NoRels are identified very poorly.

As it is shown well in the confusion matrix of DR realization type classification, in which TDB 1.2 is encoded with BERT Turkish PLM, provided in on the left of Figure 22 of Appendix C, despite its small sample size, Hypophora is very well classified. The confusion matrix reveals high numbers of true-positives for Explicit and Implicit DRs indicating that they are very well understood by the model. However, the Implicit DRs also yielded false positives and are mostly mistaken for EntRels, NoRels and AltLexes.

According to the confusion matrix, the model frequently mixes EntRels with implicits because bulk of the EntRels are predicted to be implicits and this finding was not surprising: [17] report that human annotators also struggle with telling these two relations apart. Implicit relations conveying an Expansion sense (especially the Level-2 sense of ‘level-of-detail’) look very similar to EntRels as they also tend to talk about a common entity.

As for NoRels, they are almost always classified as an implicit relation (18 out of 21 relations). Therefore, it is clear that the model did not learn the difference between these two relations. This is probably the case because the number of NoRels is pretty limited in the data and more importantly, these non-relations do not have any specific clue for the model to pick-up (e.g., a set of recurrent tokens such as connectives); so the model tends to classify them as the more frequent (inter-sentential) implicit relations that also do not have any characteristic explicit clue.

NoRels consist only 5.1% of TDB and signals a complete absence of a relation whatsoever in the text span, unlike the rest of labels which convey an existence of a relation in one way or another. Hence, NoRels directly contradict the remaining labels.

5.2.3 Semantic Similarity Analysis of Discourse Relation Realizations

The literature has shown that instead of using static vectors (e.g., word2vec), the use of contextualized word representations is useful in nearly every NLP task and the biggest step has been the innovation of sentence embedding as demonstrated by BERT, which makes a semantic similarity analysis feasible (this is detailed below).

A cosine similarity analysis through the investigation of classification test scores of relation realization types of TDB 1.2 has shown that within a discourse realization type category, the higher the level of heterogeneity among the text parts formed by the arguments of its discourse relations, the better the performance of deep learning for shallow discourse parsing.

In order to gain further insight into the model’s decisions, we first analyzed the results in the form of confusion matrix provided on the left of Figure 22 of Appendix C. Then, given that contextualized word representations have proven to be more useful than static vectors such as word2vec in almost every NLP task, we extracted the sentence embeddings of the textual elements, formed by concatenating Arg1, Arg2 and the discourse connective (if available) of each discourse relation in the TDB 1.2, by encoding with the same BERT Turkish PLM. Computing the cosine similarity of each textual element versus all other textual elements within each set of similarly realized relation types, each DR realization type received $(n*n)$ number of similarity scores, where n is the number of relations in a category.

The bar-chart showing the semantic similarity analysis of DRs in the TDB 1.2 (encoded with BERT Turkish PLM) is plotted in Figure 14, and it reveals that similarity scores are all in a small range of 0.2 – 0.24. So, the bar chart does not reveal sufficiently useful results. It could only be said that NoRel and EntRel type of relations bear more similarity as compared to other types.

In Figure 15, the similarity scores of each DR realization type are plotted in a box-plot and their category-wise averages are shown in a bar-chart. The box-plot shows that even small differences are significant because the upper and lower quartiles are very close to each other in all categories and most of the scores are squeezed in very small ranges. Thus, the analysis shows that the average similarity scores of NoRels, EntRels and Hypophora are higher than those of other relation realization types. This finding proves one of the reasons for the low classification performance of particularly NoRels and EntRels, as they seem to confirm the phenomenon that the lesser the diversity among the samples in a group, the lower the distinguishability is.

5.2.4 Discourse Relation Sense Classification

We attempted two sense classification tasks over the data. The first sense classification task has already been mentioned above; it is a five-way classification, which focuses on the classification of the senses conveyed by all discourse relation types targeting the four Level-1 senses, i.e. Expansion, Comparison, Contingency, and Temporal categories (which exist on Explicit, Implicit and AltLex DRs) plus the category "None", which is an assembly tag created to involve the DRs that receive to sense tag in the corpus (Hypophora, EntRel, NoRel). This five-way sense classification experiment is performed over all relation types at once and the confusion matrix in Figure 16 is produced, which reveals that except for Expansion, most of the senses are inaccurately classified at high rates. The Expansion class is also mistaken for all other four senses at considerable levels; this could be claimed as the

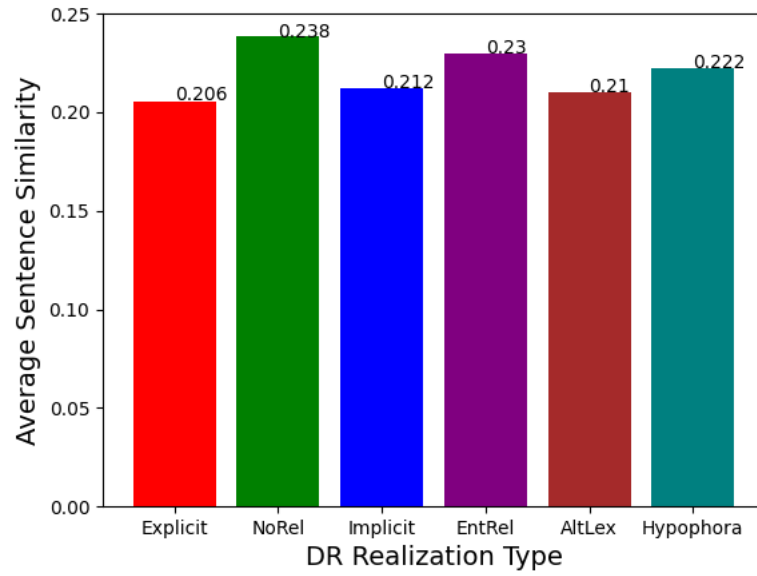


Figure 14: The bar-chart showing the semantic similarity analysis of DR types in the TDB 1.2 (encoded with BERT Turkish PLM)

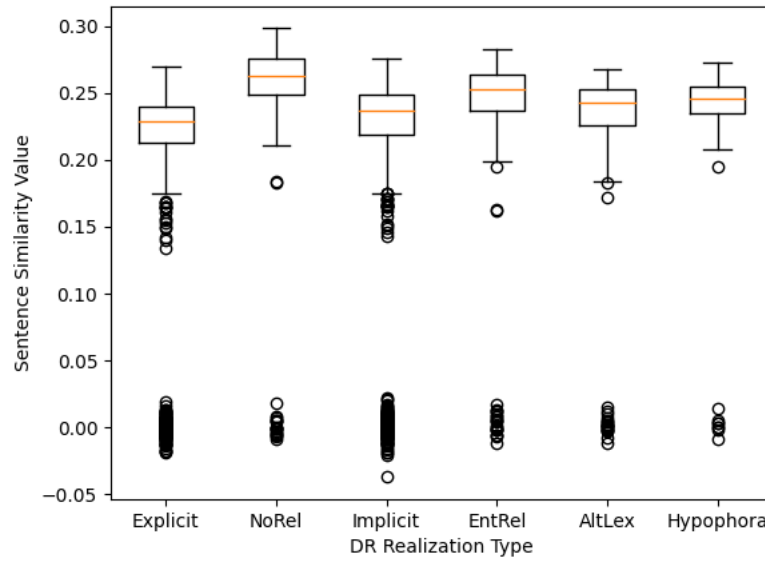


Figure 15: The box-plot showing the semantic similarity analysis of DR types in the TDB 1.2 (encoded with BERT Turkish PLM)

primary classification mistake in classifying the senses. The category None is at the lowest level of distinguishability and is often mistaken for Expansion.

Comparison	41	6	24	8	1
Contingency	3	56	28	12	5
Expansion	20	29	187	38	18
None	4	7	44	49	9
Temporal	2	2	24	5	58
	Comparison	Contingency	Expansion	None	Temporal

Figure 16: The confusion matrix of DR realization Level-1 sense classification, where the model is trained over TDB 1.2 (encoded with BERT Turkish PLM)

Not having obtained a satisfactory result for sense classification with the five-way classification technique using the BERT Turkish PLM, we ran a new experiment, with a method that closely mimics the nature of the annotations. Here, we trained separate sense classifiers for explicit and implicit discourse relations (we excluded the "None" category) and ran four-way sense classifications over the two major DR categories. Ideally, one could also train one for AltLex relations as they also convey a sense but that was not possible for the TDB 1.2, due to the lack of enough training data (there is a total of 152 AltLexes in the TDB 1.2 (see Table 1)).

The results of the four-way Level-1 sense classification experiment are provided in Table 13. As expected, sense-wise, explicitly conveyed discourse relations are much easier to classify than implicitly conveyed ones. The explicit sense classifier achieves almost two times better performance than its implicit counterpart (0.54 vs 0.82 F1-Score) and the classification is very stable across four major sense categories: The classifier achieves 0.81+ F1-Score for each major sense category, which suggests that it is robust to the uneven distribution of labels in the training data (e.g., Expansion relations are twice as frequent as Contingency relations (see Table 10). Hence, it should be safe to conclude that such explicit sense classifiers require only several hundreds of examples per label to achieve steady performance.

However, when compared to previous work in English, even the earliest studies conducted on the PDTB report much higher performances; for example, [25] achieves 93.09% accuracy with a much smaller model. Although the amount of training data in English is unquestionably larger, we also

Table 13: Explicit and implicit sense classification results over the TDB 1.2

SENSE	Explicit			Implicit		
	F1-Score	Recall	Precision	F1-Score	Recall	Precision
Comparison	0.85	0.82	0.89	<u>0.13</u>	0.12	0.17
Contingency	<u>0.81</u>	0.80	0.83	0.37	0.33	0.48
Expansion	<u>0.81</u>	0.76	0.87	0.69	0.72	0.66
Temporal	0.82	0.91	0.74	0.45	0.50	0.43
Accuracy (%)	82.20			53.37		
F1-Score	0.82			0.54		

TDB 1.2 is encoded with BERT Turkish PLM

suspect that the lower performance in Turkish may also have something to do with the nature of the Turkish connectives.

For example, unlike English, both words and morphemes may act as connectives in Turkish. Moreover, due to vowel harmonization, these morphemic connectives may be realized in various surface forms, greatly enlarging the list of possible connectives. We leave the more detailed study of identification of Turkish connectives as a future study.

The sense classification of implicit discourse relations is a notoriously challenging task and often regarded as the most difficult step in shallow discourse parsing. Accordingly, we achieve much lower performance in the disambiguation of implicit relations. (However, considering the size of the training data, the results are in line with our expectations; for example in the PDTB 2.0, the same setup achieves the F1-Score of 0.52 [76].

Unlike explicit relations, the performance varies significantly across different labels, where Expansion relations are clearly more successfully classified. This variation can be partly explained by the label distribution in the data: Expansion relations occur almost four times more frequently than the second most frequent label (Contingency).

However, the frequency of labels do not explain the poor performance on the Comparison relations since Temporal relations are classified much better despite being slightly less frequent in the training set. Therefore, in addition to the insufficient exposure to some labels, certain relations may be inherently more challenging to classify.

5.2.5 A Cross-Domain Experiment

The lack of annotated data in discourse parsing does not only make it challenging to train high-performance discourse parsers but also hinders the models from generalizing across different domains [77]. Therefore, it is crucial to evaluate the performance of the models on different domains in order to get a complete picture of their real performance. To this end, in this section, we report the performance of our classification models i.e. DR realization type and DR sense classifiers on the Turkish part of the TED-MDB corpus [18]. The TED-MDB is a multilingual corpus that follows the same annotation principles of the PDTB and the TDB. Yet, unlike those corpora, the TED-MDB is annotated on the

subtitles of six TED talks. The annotated talks differ from each other in terms of their subject matter, which makes it the perfect candidate for such a cross-domain evaluation. In total, there are 695 discourse relations (317 explicit and 210 implicit relations) in the Turkish part of the corpus.

The classification performance of the three models (i.e., DR realization types, and Level-1 senses for explicit and implicit categories) are provided in Table 14 and 15 and as expected, the performance of all drop on the T-TED-MDB[78]. Yet, compared to the values of Table 13, the accuracy drop is within acceptable margins for both explicit (82.2% vs. 73.2% (-9%)) and implicit (53.3% vs. 49.8% (-3.5%)) in the four-way sense classification test. On the other hand, genre change seems to have affected the DR realization type classifier considerably, where the performance drops almost 0.23 (0.77 vs. 0.54 (-0.23)) in F1-Score (see Table 12).

Table 14: Cross-domain DR realization type classification results over the T-TED-MDB

DR Realization Type	F1-Score	Recall	Precision
AltLex	0.37	0.42	0.35
EntRel	0.18	0.18	0.34
Explicit	0.77	0.80	0.74
Hypophora	0.11	0.20	0.10
Implicit	0.47	0.46	0.49
NoRel	<u>0.09</u>	0.06	0.35
Accuracy (%)	59.85		
F1-Score	0.54		

T-TED-MDB is encoded with BERT Turkish PLM

Table 15: Cross-domain four-way sense classification results over the T-TED-MDB

SENSE	Explicit			Implicit		
	F1-Score	Recall	Precision	F1-Score	Recall	Precision
Comparison	0.71	0.69	0.75	0.16	<u>0.14</u>	0.21
Contingency	0.75	0.73	0.77	0.21	0.36	0.17
Expansion	0.80	0.78	0.82	0.65	0.61	0.76
Temporal	<u>0.65</u>	0.83	0.55	0.41	0.40	0.56
Accuracy (%)	73.18			49.75		
F1-Score	0.76			0.52		

T-TED-MDB is encoded with BERT Turkish PLM

The performance drops over all DR realization types, but the AltLexes suffer the most significant performance loss, where the classification performance almost drops to half. Considering that, AltLexes are rather an open class; compared to explicit connectives, it seems that the models do not learn AltLexes well enough to generalize over unseen AltLex phrases. Overall, although there is much room for improvement, considering the size of the T-TED-MDB, the cross-domain performance of our models are in line with our expectations and can indeed be useful in mining relations on different text types.

In the next chapter, we explain the details of our final set of experiments, i.e. Study 3, where we attempt to aggregate TDB with a multilingual data set for Cross-lingual Transfer Learning.

CHAPTER 6

STUDY III: ENCODING MULTILINGUAL DATASET WITH THE MULTILINGUAL BERT FOR CROSS-LINGUAL TRANSFER LEARNING EXPERIMENTS

In the final set of experiments (Study 3), multilingual dataset variations are encoded with the multilingual BERT for Cross-lingual Transfer Learning, exploited to leverage DR realization type identification in Turkish. Through a series of Cross-lingual Transfer learning experiments, we explore the effect of multilingual data aggregation on DR realization type classification (Section 6.2) and discuss the extent to which the lack of training data that arises due to the cost of manual annotation can be alleviated in a low-resource scenario such as ours. The classifier model to be utilized is the same as described in Appendix B.

6.1 Reasoning for Cross-lingual Transfer Learning Experiments

The overarching problem in discourse parsing studies in general is the data bottleneck. Although the results provided in the previous sections are on the similar levels with what is achieved for English despite the significantly lower size of the TDB, clearly there is much room for improvement. In these experiments, we aimed to explore whether enriching our training data with a larger resource in another language can help us to improve our scores.

To this end, we focused on the DR realization type sub-task and performed Cross-lingual Transfer Learning experiments using the CDB (see Section 3.1.2) and the PDTB 3.0 (see Section 3.1.3) paired with TDB 1.2, and one final experiment where all three were evaluated at once.

We have focused on the following research questions about Cross-lingual Transfer Learning applied by the virtue of multilingual data additions and encoding with multilingual BERT PLM:

- As the first and foremost question: to what extent does the multilingual training data grow affect the performance of Turkish DR realization type identification?
- Is there a performance loss due to incorporated languages as compared to the levels with their monolingual encoding?
- Are there performance differences among experiment results conducted on language pairs and all languages together? If there is any, what could be the reason?

6.2 Cross-lingual Transfer Learning Experiments and the Results

Firstly, the standard approach (of the classification architecture of the study) is applied on each language with its specific monolingual BERT model in order to form a basis for comparison as detailed here. Then the same method is applied on the multilingual datasets, derived by pairing languages as detailed in Section 3.1.3, in order to realize a sound Cross-lingual Transfer Learning. So, a new multilingual model is produced by fine-tuning the BERT multilingual PLM over the combination of training sets of three languages. A separate test is applied for each language to do language specific assessments.

The Multilingual Dataset, introduced in Table 2, is summarized in Table 16, with respect to the following: (i) the DR token numbers for the training and testing phases of the experiments, (ii) the weights of datasets in three different experiments reported in this section.

The annotated tokens in the English and Chinese corpora are abundant in the experiments of language pairs, and in the experiment with all languages, the annotated tokens of the English corpus exceeds all other languages. This implies that the PDTB could contribute much to the performance of a low resourced language, but as we will show below, this was not the case at the end.

Table 16: DR token statistics of the three datasets in the joined multilingual dataset formations

Language	Training	Test	Total	TDB+CDB	TDB+PDTB	ALL
English (PDTB-3)	41,175	10,294	51,469	0	93%	76,3%
Chinese (CDB)	9,579	2,396	11,975	75%	0	17,7%
Turkish (TDB 1.2)	3,189	798	3,987	25%	7%	6%
Total	53,943	13,488	67,431			

As always, training is applied on 80% of data.

Joined multilingual datasets are encoded with multilingual BERT PLM.

In total, we have considered two different experiment settings: (i) The Cross-lingual Transfer Learning settings where a model is trained on the aggregation of all possible group variations among the three datasets, introduced in Section 3.2.1 (i.e. the PDTB 3.0 with the TDB 1.2, the CDB with the TDB 1.2 and all three together¹), encoded with multilingual BERT PLM², with the results given in Table 17. (ii) The language specific monolingual settings for separate experiments where a model is trained on its own dataset, encoded with its own monolingual BERT PLM³ and the newly multilingual PLM fine-tuned in the first phase, with the results given in Table 18.

So, the results of the Cross-lingual Transfer experiments are evaluated in two categories: Table 17 is designed (i) to compare the performance of the TDB 1.2 monolingual baseline with the Cross-lingual Transfer experiment results, i.e. with those of multilingual encoded models of both language pairs and the concatenation of three languages, where performance differences are reported by F1-Scores, and

¹ Only the PDTB 3.0 with the CDB pairing is skipped since the main research direction is towards enhancing Turkish (a low resourced language).

² <https://huggingface.co/bert-base-multilingual-cased> for multilingual setting.

³

<https://huggingface.co/dbmdz/bert-base-turkish-128k-uncased> for TDB,

<https://huggingface.co/bert-base-chinese> for CDB,

<https://huggingface.co/bert-base-cased> for PDTB.

(ii) to check the occurrence of any performance loss caused by multilingual encoding for the joined languages by comparing the results with their monolingual encoding. Table 18 lists the F1-Scores of DR realization type classification experiments of the TDB 1.2, CDB and PDTB 3.0, encoded with their respective monolingual BERT, and with the new PLM, produced by fine tuning the multilingual BERT with the training dataset formed by joining the training portions of three datasets.

Going from the bottom to the top in Table 17, there is a slight increase (+1.3%) in the accuracy of the main task such that the DR types of TDB 1.2 are classified more accurately with the PLM fine-tuned over three datasets. The result with the experiment of ‘TDB 1.2 + PDTB 3.0’ dataset is substantially lower as compared to the monolingual baseline and the experiment of ‘TDB 1.2 + CDB’. So, the experiments of language pairs show that the main contribution comes from the CDB contrary to what was expected from the PDTB 3.0, because we expected the common rule (the more the training samples, the merrier the model performance) to be at work.

The counter-expectations are not limited to the number of training samples. Despite the fact that the text genres of the PDTB 3.0 (news) and the TDB 1.2 (news, fiction, biography, etc.) are similar to each other, the PDTB 3.0 did not help much to the improvement of the target task. On the other hand, although the CDB is quite a different domain and involves DR annotations on the transcripts of oral speech (TED Talks), the results are very convincing for the potential improvement of Turkish with Chinese, i.e., they are promising in leveraging DR type identification through multilingual DR realization identification. (While investigating the causes of this confusing result, we carried out a fact-finding analysis described in Section 6.3.)

Table 17: **F1-Scores** of Cross-lingual Transfer experiments that classify DR realization types in TDB 1.2.

	Mono	Multi-I	Multi-II	Multi-III
DR Type	TDB 1.2	TDB 1.2 + CDB	TDB 1.2 + PDTB 3.0	ALL
AltLex	0.63	0.60 (-0.03)	0.22 (-0.41)	0.54 (-0.09)
EntRel	0.32	0.22 (-0.10)	0.07 (-0.25)	0.20 (-0.12)
Explicit	0.90	0.88 (-0.02)	0.82 (-0.08)	0.89 (-0.01)
Hypophora	0.74	0.57 (-0.17)	0.44 (-0.30)	0.49 (-0.25)
Implicit	0.77	0.72 (-0.05)	0.74 (-0.03)	0.74 (-0.03)
NoRel	0.11	0.20 (+0.09)	0.07 (-0.04)	0.19 (+0.08)
Accuracy (%)	73.90	74.93 (+1.03)	66.92 (-6.98)	75.20 (+1.3)
F1-Score	0.77	0.76 (-0.01)	0.67 (-0.10)	0.74 (-0.03)

Mono: TDB 1.2 is encoded with BERT Turkish PLM fine-tuned in monolingual setting (see Section 5.2.2) – the same results in Table 12.

Multi-I: TDB 1.2 is encoded with the multilingual BERT PLM fine-tuned in multilingual setting where the TDB 1.2 is paired with TED-CDB.

Multi-II: TDB 1.2 is encoded with the multilingual BERT PLM fine-tuned in multilingual setting where the TDB 1.2 is paired with PDTB 3.0.

Multi-III: TDB 1.2 is encoded with the multilingual BERT PLM fine-tuned in multilingual setting where three of the TDB 1.2, TED-CDB and PDTB 3.0 formed a training set together.

In Table 17, performance drop occurs most sharply for Hypophora relation in all settings with the maximum decrease in the ‘TDB 1.2 + PDTB 3.0’ setting and this setting has also caused much performance loss with EntRel and AltLex relations. On the other hand, the performance in implicit relations is very close to the monolingual baseline and the performance drop remains within a reasonable margin.

When the training data is composed of both the PDTB 3.0 and the CDB, the results increase perceptibly. The performance becomes slightly better than that of the monolingual baselines, with 1.03% increase in accuracy and almost the same F1-Score. A very valuable development holds for NoRel and it also appears in the three-lingual setting.

Table 18 summarizes the results of the DR realization type classification experiments conducted to evaluate the performance of each language, encoded with both its own BERT PLM and the multilingual BERT PLM fine-tuned with the combined dataset of three languages. From the view of performance comparison, there are quite few losses for the CDB and the PDTB 3.0 caused by multilingual encoding. Among the monolingual encoding experiments, the BERT PLM for English appeared to be the one with best performance, making the PDTB 3.0 reach 77% accuracy, whereas the multilingual PLM caused the highest level of imperfection for English with an accuracy loss of the highest magnitude (-2.67%). However, the evaluation of all datasets by F1-Scores does not reveal any accuracy increases, showing instead that they maintain their levels very well. This result suggests that the DR realization type classification may not be a language-specific task and the addition of another language to the training set could help to bring about a solid improvement in the relation detection and identification.

Table 18: **F1-Scores** of DR realization type classification experiments of: (i) TDB 1.2, CDB and PDTB 3.0, encoded with their respective monolingual BERT PLM, (ii) The combination of three languages, encoded with the multilingual BERT PLM.

DR Type	TDB 1.2		CDB		PDTB 3.0	
	Mono	Multi	Mono	Multi	Mono	Multi
AltLex	0.63	0.54 (-0.09)	0.33	0.39 (+0.06)	0.50	0.47 (-0.03)
EntRel	0.32	0.20 (-0.12)	0.27	0.25 (-0.02)	0.57	0.57 (0)
Explicit	0.90	0.89 (-0.01)	0.82	0.80 (-0.02)	0.84	0.81 (-0.03)
Hypophora	0.74	0.49 (-0.25)	0.19	0.44 (+0.25)	0.13	0.30 (+0.17)
Implicit	0.77	0.74 (-0.03)	0.77	0.75 (-0.02)	0.75	0.73 (-0.02)
NoRel	0.11	0.19 (+0.08)	0.14	0.02 (-0.12)	0.00	0.01 (+0.01)
Accuracy (%)	73.90	75.20 (+1.3)	71.27	70.86 (-0.41)	77.00	74.33 (-2.67)
F1-Score	0.77	0.74 (-0.03)	0.70	0.69 (-0.01)	0.75	0.74 (-0.1)

F1-Scores of monolingual and multilingual DR realization type classification experiments are given for three languages as such:

Mono: The dataset is encoded with its monolingual BERT PLM

Multi: The dataset is encoded with the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages for the Cross-lingual Transfer learning purpose.

A test with downsized samples:

Conforming to the scenario of this chapter, the results in Table 18 are obtained over the number of DR realization tokens reported in Table 16, which obviously reveals a very unbalanced structure in terms of sample size for each language. In order to assess the outnumbering effect, the experiment is repeated by downsizing the CDB and the PDTB 3.0 to the token numbers of the TDB 1.2, given in Table 1, i.e., we randomly chose the same number of tokens for each DR realization type category and rerun the experiment. The experiment of equalized datasets produced the results in Table 19 revealing losses in the F1-Scores for each language. This could imply that the increase in the number of samples in any language could help to improve the training performance of classification for relation detection and identification.

Table 19: F1-Scores of the same DR realization type classification experiments with **equal number of tokens in each DR realization type category of each language**.

DR Type	TDB 1.2		CDB		PDTB 3.0	
	Mono	Multi	Mono	Multi	Mono	Multi
AltLex	0.63	0.10 (-0.53)	0.33	0.00 (-0.33)	0.50	0.17 (-0.33)
EntRel	0.32	0.35 (+0.03)	0.27	0.18 (-0.09)	0.57	0.32 (-0.25)
Explicit	0.90	0.84 (-0.06)	0.82	0.74 (-0.08)	0.84	0.69 (-0.15)
Hypophora	0.74	0.32 (-0.42)	0.19	0.20 (+0.01)	0.13	0.18 (+0.05)
Implicit	0.77	0.74 (-0.03)	0.77	0.73 (-0.04)	0.75	0.65 (-0.05)
NoRel	0.11	0.04 (+0.08)	0.14	0.16 (+0.02)	0.00	0.08 (+0.08)
Accuracy (%)	73.90	71.00 (-2.9)	71.27	66.00 (-5.27)	77.00	60.00 (-7.00)
F1-Score	0.77	0.69 (-0.08)	0.70	0.63 (-0.07)	0.75	0.59 (-0.16)

The earlier experiment, given in Table 18, is repeated by downsizing the sample numbers of each DR realization type category of the CDB and the PDTB 3.0 to those of the TDB 1.2 (see Table 1).

The F1-Scores of monolingual and multilingual DR realization type classification experiments are given for three languages as such:

Mono: The whole dataset is encoded with its monolingual BERT PLM

Multi: The downsized dataset is encoded with the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages for the Cross-lingual Transfer learning purpose.

A Kappa statistic to compare the results of monolingual and multilingual models:

The confusion matrices of DR realization type classification experiments of the TDB 1.2, the CDB and the PDTB 3.0, encoded with their respective monolingual BERT, and the experiment with the combination of three languages, encoded with the multilingual BERT are plotted in Figures 22, 23 and 24 of Appendix C; in order to summarize the classification performance of the monolingual and the multilingual models (in their normal sample size), and a κ (Kappa) statistic is computed to compare the results. We find that the experiments with multilingual PLMs produce κ coefficient decreases by 1.8%, 0.9% and 3.5% (see Table 23) as compared to the κ coefficients reached by the respective monolingual PLMs of Turkish, Chinese and English.

But except English, the decreases are very small. This leads us to suggest that the TDB 1.2 and the CDB may be taken as the two datasets that can be used in a multilingual setting to leverage any classification performance in either language. Thus, contrary to the findings of the recent work of [79] on connective prediction, where the authors also report that the language specific models trained on the target language outperform the multilingual model trained on a concatenation of different languages, our results stand for the advantage of a multilingual training model.

Table 17 and our Kappa analysis suggest that the types by which discourse relations are realized in Turkish and English are diverse enough to prevent any knowledge transfer even between the resources that are annotated following the same framework, but the transfer from a different genre (even from a typologically different language) could create quite the opposite affect favoring the multilingual PLM development.

In order to shed more light into these discrepancies between Turkish and English, we have performed a manual error analysis of the predictions of the multilingual model that is trained on both resources.

According to our preliminary analysis, the following points stand out as the possible reasons behind the poor generalization across these two languages:

- Discrepancies between English and Turkish in expressing AltLexes: The largest performance drop occurs in AltLex relations, which are linguistic expressions that can act as discourse connectives in certain contexts. Hence, languages considerably vary when it comes to AltLexes, an observation also put forward by [80].
- EntRels manifest a large performance drop and they are frequently confused with (inter-sentential) implicits and NoRels. These are quite similar to the errors in monolingual experiments (see Section 5.2.2), showing that neither the monolingual model nor the multilingual model learns the EntRel and NoRel labels properly.
- Discrepancies between English and Turkish explicit relations: There is a small performance drop in the prediction of explicit relations, and our manual error analysis shows that one of the reasons of performance loss is due to the intra-sentential relations conveyed by converbial suffixes, annotated as a type of explicit connectives in Turkish. For instance, in Example 6.2.1, despite the presence of the suffixal connective *-sek* ‘if’, the relation is mislabeled as an implicit relation.

Example 6.2.1 *Üç kişi versek, güç çevreleriz.*
‘If three of us come together, we could hardly encircle it.’

Such examples show that English and Turkish annotate discourse connectives belonging to different syntactic classes (e.g., single words versus suffixes), and this could be one reason why the success of Cross-lingual Transfer (from English to Turkish) decreases in our experiments.

6.3 Semantic Similarity Analysis of Discourse Relation Realizations

Despite the poor contribution of English to Turkish, the Cross-lingual Transfer Learning experiment revealed a valuable effect of the dataset from a typologically different language and genre. Since our knowledge of Chinese and the kind of annotations created in the CDB is insufficient to carry out a linguistic analysis of why this effect has been obtained, a semantic similarity analysis could help understand the positive contribution of the Chinese dataset to the Turkish dataset.

Similar to the semantic similarity analysis of DR realizations, explained in Section 5.2.3, the same method is applied to Turkish-Chinese and Turkish-English language pairs on the basis of DR realization type category. Computing the cosine similarity of each textual element versus all other textual elements within each set of similarly realized relation types, each DR realization type received ($n*n$) number of similarity scores, where n is the number of relations in a category. The similarity scores are plotted as box-plots and their category wise averages are charted in the Figures 17 and 18.

The box-plot of the semantic similarity scores and the bar-chart of the semantic similarity averages of DR text elements per each realization type category of CDB and TDB 1.2 (encoded with the Multilingual BERT) are given in Figure 17. The figures reveal that almost all of the similarity scores, calculated between all DR realizations in Turkish and Chinese datasets, falls in a very small spectrum, squeezed

in a range decreasing from 0.04 to -0.05. Thus, upper and lower quartiles in the box-plot are very close to each other, meaning that Turkish and Chinese relations bear too small similarity scores in all categories. In addition, the type wise averages are all below zero with too small magnitudes. So, it is detected that the DR realizations in the CDB and the TDB carry almost totally different characteristics even though they are from the same DR realization type.

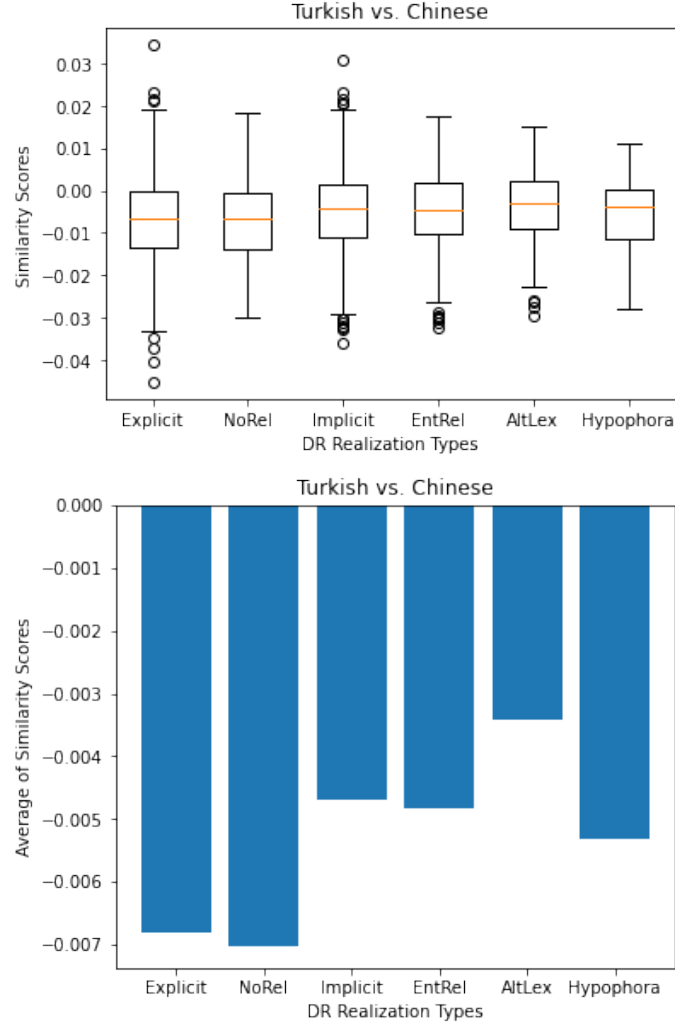


Figure 17: The box-plot of the semantic similarity scores and bar-chart of the semantic similarity averages of DR realization text elements per each realization type category between CDB and TDB 1.2 (encoded with the Multilingual BERT)

Quite different from Figure 17, the box-plot of the semantic similarity scores and the bar-chart of the semantic similarity averages of DR text elements per each realization type category of PDTB 3.0 and TDB 1.2 (encoded with the Multilingual BERT), given in Figure 18, reveal that almost all of the similarity scores, calculated between all DR realizations in Turkish and English datasets, falls in a fairly broader spectrum, in a range increasing from -0.023 to 0.08. Thus, upper and lower quartiles are very far to each other, meaning that Turkish and English relations bear higher similarity scores in all categories as compared to the Chinese-Turkish case. In other words, the discourse relations in

the PDTB 3.0 and the TDB carry similar characteristics, e.g., one DR realization type annotated in Turkish is very similar to the same type annotated in English. Despite the averages of similarity scores between Turkish and English are all positive and bear higher magnitude than it is with Chinese, the distinguishability of one category from the other is low and this finding is parallel to one of the reasons for the decrease in classification performance when we shift to the multilingual setting of TDB 1.2 + PDTB 3.0. The performance loss is apparently caused by the side effect of similarity among the samples⁴ in multiclass classification tasks.

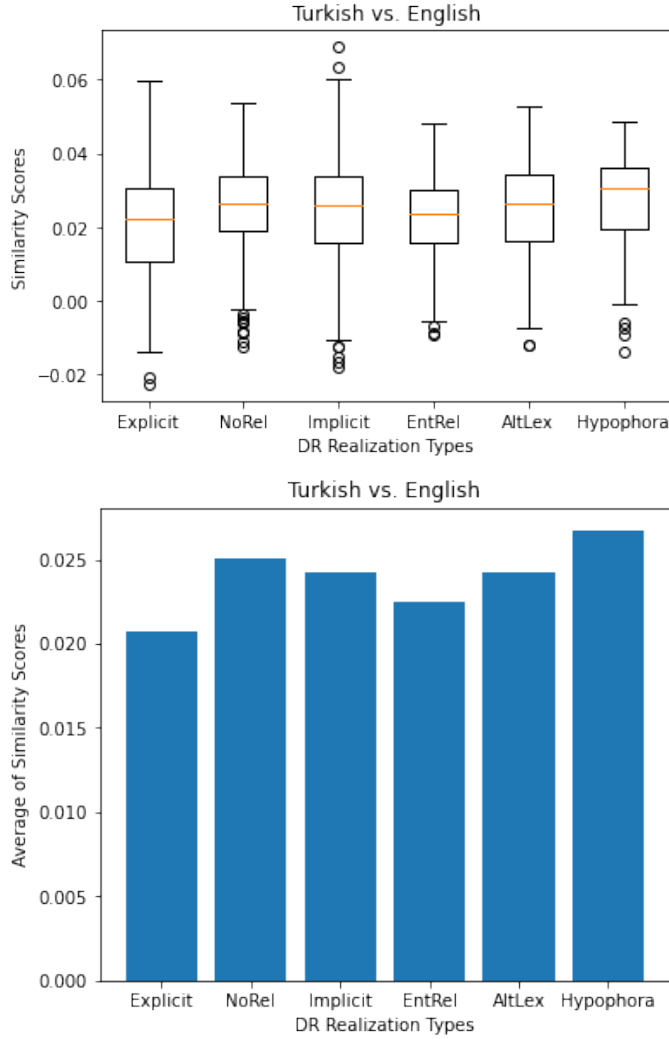


Figure 18: The box-plot of the semantic similarity scores and bar-chart of the semantic similarity averages of DR realization text elements per each realization type category between PDTB 3.0 and TDB 1.2 (encoded with the Multilingual BERT)

To summarize, in this chapter, we detailed Study 3, where we tried to leverage the DR realization type classification through multilingual data aggregation methods and in the main, we suggested that the

⁴ i.e., The phenomenon that the lesser the diversity among the data parts of the samples in categories, the lower the distinguishability holds among them.

Turkish dataset is positively affected by the Chinese dataset rather than the English dataset. The next chapter will conclude the thesis.

CHAPTER 7

CONCLUSION AND FUTURE WORK

This research focused on the automatic identification of discourse relations between two individual discourse units by using manually-annotated discourse corpora with various techniques and methods. Shallow discourse parsing (the detection of all manually-annotated categories of a discourse corpus) is an important step towards discourse understanding and would be a prominent contribution to NLU research in general. However, despite its importance, most of the existing work is still confined to English, leaving the field largely understudied in the non-English context. In this work, we aimed to help remedy this problem by performing various sub-tasks of the shallow discourse parsing pipeline on Turkish. Although our work falls short of developing a full end-to-end parser, it constitutes the most exhaustive study on Turkish so far.

Through three Studies, in this thesis, we focused on a rather overlooked task – the classification of discourse relation realization types, which focuses on understanding whether there is a discourse relation in a given text piece or not (NoRel), if there is a discourse relation, how it is realized (explicitly, implicitly, alternatively lexicalized, or as an entity relation). Besides, we performed various sense classification experiments, though only at a coarse level, i.e. at the Level-1 categories of the PDTB sense hierarchy – Expansion, Temporal, Contingency, Comparison. We attempted the well-known tasks of sense classification of explicit and implicit discourse relations separately.

All tasks are modelled as multi-class text classification problems and different from Study 1, in Study 2 and 3, we used PLMs to encode the binary arguments of DRs – constitutive units of DRs, which are available as annotated data in all the corpora we used. In other words, using PLMs, we extracted sentence embeddings by concatenating the arguments of discourse relations. The results presented in these two Studies suggest that despite the scarcity of the available training data, all tasks can be performed with a satisfactory accuracy, in monolingual and multilingual settings, and the results are largely parallel to the reported results in the literature.

In the final phase of the thesis (Study 3), we performed a cross-lingual investigation to see if it is possible to leverage information from the much bigger resources of the PDTB 3.0 and the CDB on the DR realization type classification (sense classification was left out of scope of Study 3). In order to gain an insight into the DR realization type classification decisions of both monolingual and multilingual PLMs in Study 3, we analyzed the results in terms of confusion matrices and cross checked the outputs with a κ coefficient analysis applied to each confusion matrix. This gives an objective score enabling us to compare the results of different experiment settings.

In addition, we conducted a cosine similarity analysis over the encodings of textual elements within DR realizations types in Study 2 (among the DR realization types in TDB) as well as in Study 3,

comparing the similarity of the of monolingual PLM encodings with multilingual PLM encodings among the DR realization types of each language pair with TDB. The results of this analysis implied that the higher the level of heterogeneity among the textual parts of discourse relations, the better the performance of deep learning will be in various sub-tasks of shallow discourse parsing.

In a nutshell, after having worked on the task of DR classification by grouping DR types and senses into Classes in Turkish DB and incorporating linguistic features in machine learning methods for improved results, our approach changed, so the task was turned into a new multi-class classification problem, where the classification of DR types and senses were targeted separately. With this approach, experiments that use text element encoding, PLMs and NN-based classifiers were conducted over DR realization types as well as DR senses. All in all, ranging from linguistic feature detection to context aware encoding in data processing; from machine learning methods to neural network based classification frameworks; from monolingual low resourced language settings to Cross-lingual Transfer Learning with a multilingual corpus, all modern methods have been experimented with. This approach has allowed us to deal with many difficult issues in the discourse parsing field and eventually helped us reveal outputs with a good level of general impact.

7.1 A Discussion of the Research Questions

We started the thesis with the following research questions:

- **What is the best performance level that could be reached by linguistic feature engineering and syntactic parsers in the classification of DR realization identification? Considering that feature engineering might work for a scenario where the number of annotated data is limited, how cost effective could be the deep learning models which require a lot of training?**
- **What is the impact of PLM based sentence-level text encoding mechanisms over DR realization identification (predicting the labels over DRs such as Explicit, Implicit, etc.) and their senses?**
- **What could be the architecture of a yet another Neural Network-based classifier that might possibly reach the best performance in predicting the labels of DR types (Explicit, Implicit, etc.) and their senses separately?**
- **Will a broad Cross-lingual Transfer Learning classification experiment that uses a multilingual dataset encoded by a multilingual PLM for text encoding circumvent the data scarcity problem?**
- **What kind of an effect could be produced by the model, fine-tuned within such an experiment, over the performance of the DR realization identification task?**

Our task was identification and classification of discourse relations, firstly modelled together with their type and sense into a category we named Classes in Study 1. For this task, we started with machine

learning models enriched with linguistic features and obtained the first best score (F1-Score 0.36 in C5 Decision Tree Classifier algorithm). In Study 2 and 3, we adopted a new approach and shifted to PLMs, and we experimented with sense classification of implicit and explicit DRs separately. This approach has been linguistically and cognitively more plausible, as it is parallel to the procedure of the annotations created by humans.

Each new attempt has increased the performance, as summarized in Table 20, and revealed new opportunities for further innovation in the problem area where the needs are becoming urgent.

Table 20: Overall DR Realization Type Classification Evaluation for TDB

Experiment Name	F1-Score	Accuracy
Feature Selection (Study 1)	0.36	39%
USE + BiLSTM (Study 2)	0.54	61.1%
Monolingual BERT (Study 2)	0.77	73.9%
Multilingual BERT (Study 3)	0.74	75.2%

Monolingual BERT: TDB 1.2 encoded with BERT Turkish PLM

Multilingual BERT: Multilingual dataset of three languages encoded with multilingual BERT PLM

Expressed in terms of F1-Scores of the classifications, our efforts throughout the three Studies of the thesis have resulted in: (i) rising from 0.36 to 0.77 for discourse relation realization types (see Table 20), (ii) achieving 0.82 in a four-way classification of the Level-1 senses of explicit relations, and 0.54 of implicit relations (see Table 21). The Level-2 senses increase the category number to such a high level that it becomes very difficult to end up with a sound classification performance with the number of samples available in the TDB. Thus, the study of Level-2 senses is left to future works.

Table 21: Overall DR Realization Sense Classification Evaluation for TDB

Experiment Name	Category #	DR Type	F1-Score	Accuracy
Monolingual BERT (Study 2)	4	Explicit	0.82	82.2%
Monolingual BERT (Study 2)	4	Implicit	0.54	52.37%

Monolingual BERT: TDB 1.2 encoded with BERT Turkish PLM

In the light of the findings of the Study 3 of the thesis, particularly regarding the experiments where we aggregated the TDB with the Chinese DB, it can be argued that the solutions offered by Cross-lingual Transfer Learning are mature enough to deal with the inherently complex semantic problem of DR type classification and can alleviate the data scarcity problem caused by the need to manually annotate discourse structure. This result has been backed up by the small κ coefficient differences between the classifications performed by monolingual encoded data and multilingual encoded data, and has been a good indication for the potential efficiency of multilingual encoding. Thus, investing in and relying on Cross-lingual Transfer could bring possible solutions to discourse parsing and ultimately to NLU and dialogue systems.

Although there might be reasons other than data scarcity leading to the results obtained in the Study 3 of the thesis (such as the potential complexity of the discourse structures included in the TDB which future work could reveal), this work demonstrated that, in order to create global benefits, the multilingual PLMs can serve very well in Cross-lingual Transfer Learning techniques.

NLU needs to sense the tacit linguistic information and the knowledge in text or speech. Language representations supported with lexical meanings, syntactic structures, semantic roles, even pragmatics can serve for that purpose. So, sufficient amounts of discourse-annotated data containing rich semantic/pragmatic labels have become the most urgent need for further solutions because neither syntactic features nor modern text encoding mechanisms are able to support machines to understand texts beyond the clause level.

Additionally, it is experienced that high-dimensional embeddings and the larger model size (i.e. hidden layers, maximum total input sequence length and number of trainable parameters) enhanced the classification performance, i.e. the best performance is reached by exploiting *BERT_MultiClass* model empirically developed within the research.

7.2 Highlights of the Contributions and Notes on Further Research

Classification for DRs, both in terms of their type and sense, still remains as an important problem to be resolved for languages that lag behind well-studied languages. DR realization types and the meaning the specific DRs convey are important annotation categories in discourse annotation, their automatic identification is one of the prominent steps that would contribute to not only discourse understanding but also NLU. Methods to disambiguate discourse connectives (which is in a sense, Explicit DR identification) in highly resourced languages have been proliferating, with success levels approaching 100%, e.g., in [81]. However, work on many non-English languages still have a long way to go even in such a tasks, possibly due to lack of adequate data. The thesis is expected to make an important contribution to shallow discourse parsing by the methods it employs, namely, the encoding of all DRs at once (rather than a broad explicit versus implicit type categorisation) through the latest PLMs and the classification of all DRs simultaneously with BERT-specific NN models for classification.

The other main contribution of the thesis has been the focus on Turkish discourse. The experimental results presented here will form baselines for further research on Turkish, and it is expected that future research will be able to improve considerably upon these baselines. For instance, it is clear that increasing the number of non-English samples in the multilingual dataset with the PDTB-style annotated DRs will improve the results of Cross-lingual Transfer Learning.

The results of the thesis can also support automatic DR annotation required to produce synthetic data for further studies on Turkish, and can be used by other less-resourced languages in DR realization type and sense classification.

There are several issues not covered in the thesis, such as the classification of discourse senses at a more fine-grained level (e.g., at the Level-2 of the PDTB sense hierarchy) and the detection of the constitutive units of discourse (the binary arguments of discourse relations).

The PDTB-3.0 brought almost 12.5K more intra-sentential relations and nearly 1K more inter-sentential relations than the PDTB-2; this shows that there is room for the exploration of more effective disambiguation methods to handle the larger range of DC ambiguities [2] and, the utilization of the intra-/inter-sentential implicit DR information in the training might also enhance the approach proposed by the thesis. These tasks can be tackled in the future, possibly benefiting from the pipeline we devel-

oped here and avoiding the techniques that have led to low scores. Thus, the thesis would ultimately contribute to our understanding of Turkish discourse structure on one hand and to NLU on the other.

Finally, in the PDTB style corpora, targeting an end-to-end discourse parsing would involve sub-tasks that mimic the human annotations such as: (i) the recognition of DCs by distinguishing them from non-DCs (i.e., connective disambiguation), (ii) the distinction of explicits from implicits, (iii) the recognition of the implicit and explicit discourse connective's senses (i.e., sense disambiguation), and (iv) the recognition of the textual elements of the constitutive units of discourse relations. In the current thesis, particularly the sub-tasks of (i) - (iii) have been carried out, but our system has the ability to handle (iv) as well.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.
- [2] B. Webber, R. Prasad, and A. Lee, “Ambiguity in explicit discourse connectives,” in *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, (Gothenburg, Sweden), pp. 134–141, Association for Computational Linguistics, May 2019.
- [3] K. Oflazer and M. Saraçlar, *Turkish Natural Language Processing*. Springer, 2018.
- [4] K. Oflazer and C. Bozşahin, “Turkish natural language processing initiative: An overview,” in *Middle East Technical University, Citeseer*, 1994.
- [5] E. K. Akkaya and B. Can, “Transfer learning for Turkish named entity recognition on noisy text,” *Natural Language Engineering*, vol. 27, no. 1, pp. 35–64, 2021.
- [6] G. Şeker and G. Eryiğit, “Extending a crf-based named entity recognition model for Turkish well formed text and user generated content,” *Semantic Web*, vol. 8, no. 5, pp. 625–642, 2017.
- [7] G. Eryiğit, J. Nivre, and K. Oflazer, “Dependency parsing of Turkish.,” *Computational Linguistics*, vol. 34, no. 3, pp. 357–389, 2008.
- [8] R. Çakıcı, M. Steedman, and C. Bozşahin, “Wide-coverage parsing, semantics, and morphology,” in *Turkish Natural Language Processing*, pp. 153–174, Springer, 2018.
- [9] C. F. Baker, C. J. Fillmore, and L. J. B., “The Berkeley FrameNet Project,” *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL ’98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference*, pp. 86–90, 1998.
- [10] M. Palmer, D. Gildea, and P. Kingsbury, “The Proposition Bank: An annotated corpus of semantic roles,” *Computational linguistics*, vol. 31, no. 1, pp. 71–106, 2005.
- [11] J. Bos, V. Basile, K. Evang, N. Venhuizen, and J. Bjerva, *The Groningen Meaning Bank*, pp. 463–496. Dordrecht: Springer Netherlands, 06 2017.
- [12] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber, “The penn discourse treebank 2.0.,” in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco*, Institute for Research in Cognitive Science, University of Pennsylvania, 2008.
- [13] R. Prasad, B. Webber, and A. Joshi, “Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation,” *Computational Linguistics*, 2014.

- [14] N. Asher, *Reference to Abstract Objects in Discourse*. Dordrecht/Netherlands: Kluwer, 1993.
- [15] A. Moschitti and R. Basili, “Complex linguistic features for text classification: A comprehensive study,” in *Proceedings of the 26th European Conference on Information Retrieval (ECIR)*, vol. 2997, pp. 181–196, 04 2004.
- [16] Z. Lin, H. NG, and M. Kan, “A PDTB-styled end-to-end discourse parser,” *Natural Language Engineering*, vol. 20, no. 2, pp. 151–184, 2014.
- [17] D. Zeyrek and M. Kurfali, “TDB 1.1: Extensions on Turkish discourse bank,” in *Proceedings of the 11th Linguistic Annotation Workshop, LAW@EACL 2017, Valencia, Spain, April 3, 2017*, pp. 76–81, Association for Computational Linguistics, 2017.
- [18] D. Zeyrek and M. Kurfali, “An assessment of explicit inter- and intra-sentential discourse connectives in Turkish discourse bank,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, European Language Resources Association (ELRA), 2018.
- [19] D. Zeyrek, A. Mendes, Y. Grishina, M. Kurfali, S. Gibbon, and M. Ogrodniczuk, “TED multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style,” *Language Resources and Evaluation*, vol. 54, 04 2019.
- [20] W. Long, B. Webber, and D. Xiong, “TED-CDB: A large-scale chinese discourse relation dataset on TED talks,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 2793–2803, Association for Computational Linguistics, 2020.
- [21] B. Webber, R. Prasad, A. Lee, and A. Joshi, “The penn discourse treebank 3.0 annotation manual,” 3 2019.
- [22] Z. Zhao and B. Webber, “Revisiting shallow discourse parsing in the PDTB-3: Handling intra-sentential implicits,” in *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, (Punta Cana, Dominican Republic and Online), pp. 107–121, Association for Computational Linguistics, Nov. 2021.
- [23] R. Prasad, B. Webber, and A. Lee, “Discourse annotation in the pdtb: The next generation,” in *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pp. 87–97, Association for Computational Linguistics, 2018.
- [24] D. Zeyrek, A. Mendes, and M. Kurfali, “Multilingual extension of PDTB-style annotation: The case of TED Multilingual Discourse Bank,” in *LREC*, 2018.
- [25] E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi, “Easily identifiable discourse relations,” *Technical Reports (CIS)*, p. 884, 2008.
- [26] E. Pitler and A. Nenkova, “Using syntax to disambiguate explicit discourse connectives in text,” in *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pp. 13–16, Association for Computational Linguistics, 2009.

- [27] B. Polepalli Ramesh, R. Prasad, T. Miller, B. Harrington, and H. Yu, “Automatic discourse connective detection in biomedical text,” *JAMIA (Journal of the American Medical Informatics Association)*, vol. 19, pp. 800–8, 06 2012.
- [28] S. Gopalan and S. L. Devi, “BioDCA Identifier: A system for automatic identification of discourse connective and arguments from biomedical text,” in *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pp. 89–98, The COLING 2016 Organizing Committee, 2016.
- [29] A. Al-Saif and K. Markert, “Modelling discourse relations for Arabic,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, (Edinburgh, Scotland, UK.), pp. 736–747, Association for Computational Linguistics, July 2011.
- [30] E. Pitler, A. Louis, and A. Nenkova, “Automatic sense prediction for implicit discourse relations in text,” in *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pp. 683–691, The Association for Computer Linguistics, 2009.
- [31] D. Marcu and A. Echihiabi, “An unsupervised approach to recognizing discourse relations,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp. 368–375, ACL, 2002.
- [32] Z. Lin, M. Kan, and H. Ng, “Recognizing implicit discourse relations in the Penn Discourse Treebank,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 343–351, ACL, 2009.
- [33] A. Rutherford and N. Xue, “Discovering implicit discourse relations through brown cluster pair representation and coreference patterns,” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pp. 645–654, The Association for Computer Linguistics, 2014.
- [34] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, “Class-based n -gram models of natural language,” *Computational Linguistics*, vol. 18, no. 4, pp. 467–480, 1992.
- [35] S. Mukherjee, A. Tiwari, M. Gupta, and A. Singh, “Shallow discourse parsing with syntactic and (a few) semantic features,” in *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, CoNLL 2015, Beijing, China, July 30-31, 2015*, pp. 61–65, ACL, 2015.
- [36] N. Xue, H. Ng, S. Pradhan, R. Prasad, C. Bryant, and A. Rutherford, “The conll-2015 shared task on shallow discourse parsing,” in *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, CoNLL 2015, Beijing, China, July 30-31, 2015*, pp. 1–16, ACL, 2015.
- [37] A. Zeldes, D. Das, E. G. Maziero, J. D. Antonio, and M. Iruskieta, “The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection,” in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, (Minneapolis, MN), pp. 97–104, Association for Computational Linguistics, jun 2019.

- [38] P. Muller, C. Braud, and M. Morey, “ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents,” in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, (Minneapolis, MN), pp. 115–124, Association for Computational Linguistics, June 2019.
- [39] L. Liang, Z. Zhao, and B. Webber, “Extending implicit discourse relation recognition to the PDTB-3,” *CoRR*, vol. abs/2010.06294, 2020.
- [40] J. Chen, Q. Zhang, P. Liu, X. Qiu, and X. Huang, “Implicit discourse relation detection via a deep architecture with gated relevance network,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, The Association for Computer Linguistics, 2016.
- [41] Y. Liu and S. Li, “Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1224–1233, The Association for Computational Linguistics, 2016.
- [42] L. Qin, Z. Zhang, and H. Zhao, “A stacking gated neural architecture for implicit discourse relation classification,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 2263–2270, The Association for Computational Linguistics, 2016.
- [43] W. Lei, X. Wang, M. Liu, I. Ilievski, X. He, and M. Kan, “Swim: A simple word interaction model for implicit discourse relation recognition,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 4026–4032, 2017.
- [44] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained models for natural language processing: A survey,” *CoRR*, vol. abs/2003.08271, 2020.
- [45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.
- [46] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, Curran Associates, Inc., 2013.
- [47] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *CoRR*, vol. abs/1409.3215, 2014.
- [48] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014.
- [49] M. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *CoRR*, vol. abs/1508.04025, 2015.
- [50] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, 2019.

- [51] A. Raganato, Y. Scherrer, and J. Tiedemann, “Fixed encoder self-attention patterns in transformer-based machine translation,” *CoRR*, vol. abs/2002.10260, 2020.
- [52] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.
- [53] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil, “Universal sentence encoder,” *CoRR*, vol. abs/1803.11175, 2018.
- [54] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” *CoRR*, vol. abs/2005.14165, 2020.
- [55] Y. Ji and J. Eisenstein, “One vector is not enough: Entity-augmented distributed semantics for discourse relations,” *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 329–344, 2015.
- [56] A. Nie, E. Bennett, and N. D. Goodman, “Dissent: Learning sentence representations from explicit discourse relations,” in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4497–4510, Association for Computational Linguistics, 2019.
- [57] Y. Kishimoto, Y. Murawaki, and S. Kurohashi, “Adapting BERT to implicit discourse relation classification with a focus on discourse connectives,” in *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pp. 1152–1158, European Language Resources Association, 2020.
- [58] W. Wang, V. W. Zheng, H. Yu, and C. Miao, “A survey of zero-shot learning: Settings, methods, and applications,” *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 13:1–13:37, 2019.
- [59] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.
- [60] M. Yazdani and J. Henderson, “A model of zero-shot learning of spoken language understanding,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 244–249, The Association for Computational Linguistics, 2015.
- [61] Y. N. Dauphin, G. Tür, D. H. Tür, and L. P. Heck, “Zero-shot learning and clustering for semantic utterance classification,” in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

- [62] P. Pasupat and P. Liang, “Zero-shot entity extraction from web pages,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 391–401, The Association for Computer Linguistics, 2014.
- [63] Y. Ma, E. Cambria, and S. Gao, “Label embedding for zero-shot fine-grained named entity typing,” in *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pp. 171–180, ACL, 2016.
- [64] R. Funaki and H. Nakayama, “Image-mediated learning for zero-shot cross-lingual document retrieval,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 585–590, The Association for Computational Linguistics, 2015.
- [65] O. Levy, M. Seo, E. Choi, and L. Zettlemoyer, “Zero-shot relation extraction via reading comprehension,” in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pp. 333–342, Association for Computational Linguistics, 2017.
- [66] S. Schuster, S. Gupta, R. Shah, and M. Lewis, “Cross-lingual transfer learning for multilingual task oriented dialog,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 3795–3805, Association for Computational Linguistics, 2019.
- [67] T. Caselli and A. Üstün, “There and back again: Cross-lingual transfer learning for event detection,” in *Proceedings of the Sixth Italian Conference on Computational Linguistics*, vol. 2481, CEUR Workshop Proceedings (CEUR-WS.org), 2019.
- [68] D. Zeyrek and B. Webber, “A discourse resource for Turkish: Annotating discourse connectives in the METU corpus,” in *Proceedings of the 6th Workshop on Asian Language Resources*, 2008.
- [69] D. Zeyrek and K. Başbüyük, “TCL - a lexicon of Turkish discourse connectives,” in *Proceedings of the First International Workshop on Designing Meaning Representations*, (Florence, Italy), pp. 73–81, Association for Computational Linguistics, Aug. 2019.
- [70] D. Zeyrek and M. E. Er, “A description of Turkish Discourse Bank 1.2 and an examination of common dependencies in Turkish Discourse,” 2022.
- [71] W. Spooren and L. Degand, “Coding coherence relations: Reliability and validity,” *Corpus linguistics and linguistic theory*, vol. 6, no. 2, pp. 241–266, 2010.
- [72] M. Kurfali and R. Östling, “Zero-shot transfer for implicit discourse relation classification,” in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pp. 226–231, Association for Computational Linguistics, 2019.
- [73] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017.

- [74] O. T. Yıldız, B. Avar, and G. Ercan, “An open, extendible, and fast Turkish morphological analyzer,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, (Varna, Bulgaria), pp. 1364–1372, INCOMA Ltd., Sept. 2019.
- [75] W. Shi and V. Demberg, “Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification,” in *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, (Gothenburg, Sweden), pp. 188–199, Association for Computational Linguistics, May 2019.
- [76] N. Kim, S. Feng, C. Gunasekara, and L. Lastras, “Implicit discourse relation classification: We need to talk about evaluation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 5404–5414, Association for Computational Linguistics, July 2020.
- [77] E. Stepanov and G. Riccardi, “Towards cross-domain PDTB-style discourse parsing,” in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, (Gothenburg, Sweden), pp. 30–37, Association for Computational Linguistics, Apr. 2014.
- [78] F. Kutlu, D. Zeyrek, and M. Kurfali, “Towards a shallow discourse parser for Turkish, (revised version under evaluation),” 2023.
- [79] T. Chapados Muermans and L. Kosseim, “A BERT-Based Approach for Multilingual Discourse Connective Detection,” in *Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings*, (Berlin, Heidelberg), p. 449–460, Springer-Verlag, 2022.
- [80] S. Özer, M. Kurfali, D. Zeyrek, A. Mendes, and G. V. Oleskeviciene, “Linking discourse-level information and the induction of bilingual discourse connective lexicons,” *Semantic Web*, vol. 13, no. 6, pp. 1081–1102, 2022.
- [81] R. Knaebel and M. Stede, “Contextualized embeddings for connective disambiguation in shallow discourse parsing,” in *Proceedings of the First Workshop on Computational Approaches to Discourse*, (Online), pp. 65–75, Association for Computational Linguistics, Nov. 2020.
- [82] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [83] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.
- [84] M. Grandini, E. Bagli, and G. Visani, “Metrics for multi-class classification: an overview,” 2020.
- [85] M. J. Warrens and Y. Wu, “New interpretations of cohen’s kappa,” *Journal of Mathematics*, vol. 10.1155/2014/203907, 9 2014.
- [86] T. Byrt, J. Bishop, and J. Carlin, “Bias, prevalence and kappa,” *J Clin Epidemiol*, vol. 46(5), pp. 423–429, 5 1993.
- [87] A. J. Tallón-Ballesteros and J. C. Riquelme, “Data mining methods applied to a digital forensics task for supervised machine learning,” in *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications* (A. K. Muda, Y.-H. Choo, A. Abraham, and S. N. Srihari, eds.), vol. 555 of *Studies in Computational Intelligence*, pp. 413–428, Springer, 2014.

APPENDIX A

PDTB 3.0 SENSE HIERARCHY

The leftmost column contains the Level-1 senses and the middle column, the Level-2 senses in Figure 19. For asymmetric relations, Level-3 senses are located in the rightmost column [2].

While the TDB 1.2 and PDTB 3.0 datasets both assign senses from all three levels, the present work exploits Level-1 senses only.

Level-1	Level-2	Level-3
TEMPORAL	SYNCHRONOUS	–
	ASYNCHRONOUS	PRECEDENCE
		SUCCESSION
CONTINGENCY	CAUSE	REASON
		RESULT
		NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF
		RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT
		RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND
		ARG2-AS-COND
	CONDITION+SPEECHACT	–
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND
		ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	–
	PURPOSE	ARG1-AS-GOAL
		ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER
		ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
	SIMILARITY	–
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT
		ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE
		ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL
		ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER
		ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST
		ARG2-AS-SUBST

Figure 19: PDTB 3.0 Sense Hierarchy [2]

APPENDIX B

THE CLASSIFICATION MODEL, METHOD OF FINE-TUNING AND TESTS

As shown in Figure 20, a neural network based classification model is devised and magnified up to 768 hidden layers with 184,345,344 parameters for the improvement of results by the virtue of the GPU memory size, which is fully utilized.

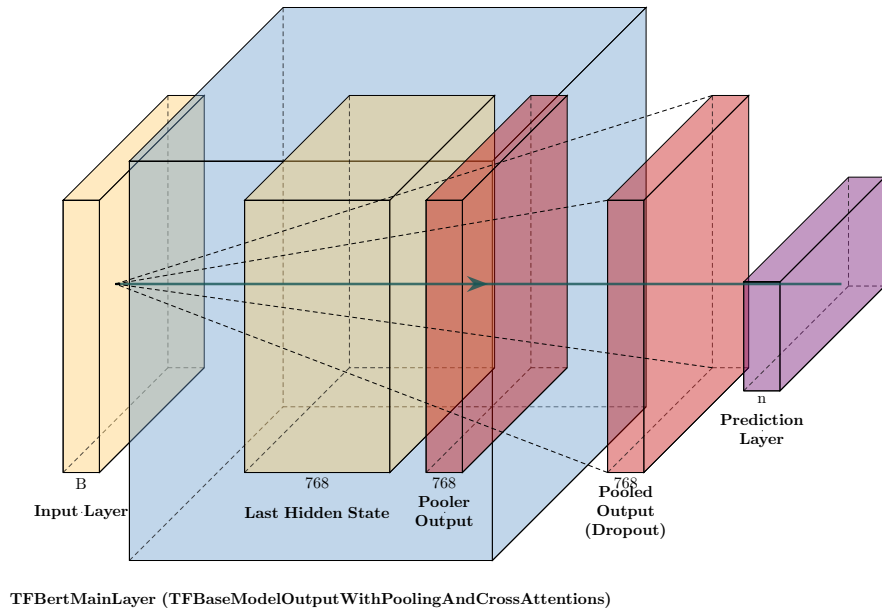


Figure 20: The Symbolic Representation of the BERT MultiClass TensorFlow Model

The first component of the model is the Input Layer that feeds the model with B (Batch Size) number of DRs (encoded by BERT) in each iteration. The main body of the model is TensorFlow BERT, the Main Layer is *TFBaseModelOutputWithPoolingAndCrossAttentions* class, released in transformers library¹ and forms a base class for the model's outputs that also contains a pooling of the last hidden states. The class has two components as the Last Hidden State and the Pooler Output. The Last Hidden State

¹ https://github.com/huggingface/transformers/blob/main/src/transformers/modeling_tf_outputs.py

of the TensorFlow BERT Main Layer is a TensorFlow tensor of the shape B, M (maximum sequence length (128)) and H (number of hidden layers (768))². Since the the number of words in the 99% of the DR realization arguments' text elements in the TDB 1.2 are less than 128, as shown in Figure 21, taking the maximum sequence length as 128 caused better results and optimal usage of memory.

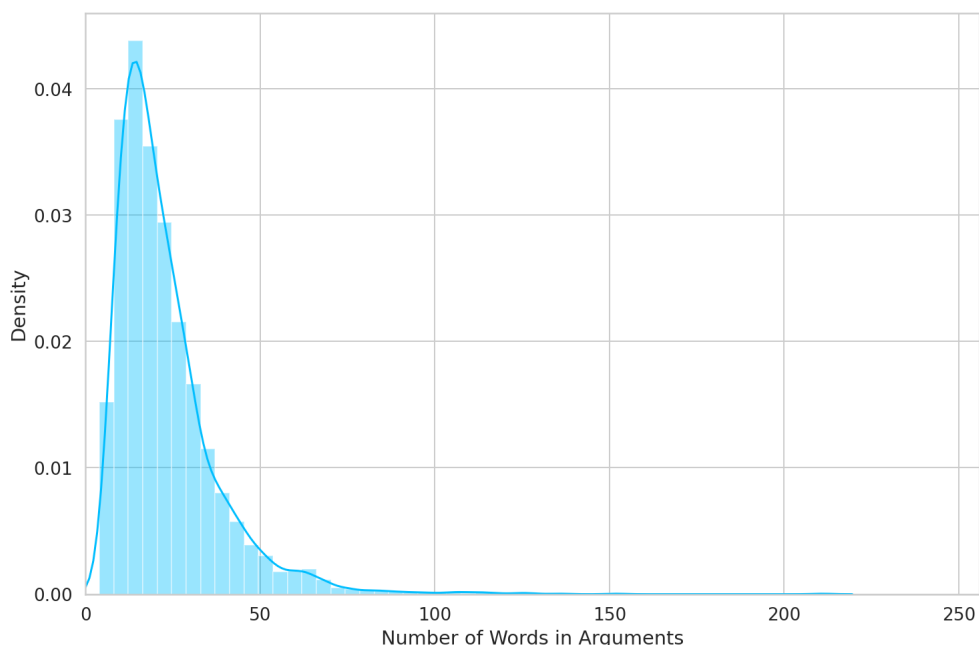


Figure 21: The bar chart of the number of words in the textual elements of a DR in TDB 1.2 listed in Table 1.

The BERT Main Layer is followed by the Pooler Output, which is a TensorFlow tensor of the shape B and H. The Pooler Output is the last layer where hidden-state of the first token of the sequence (classification token) further processed by a linear layer and a Tanh activation function. The Pooled Output is the drop out layer of shape B and H, which collects the results for the softmax function. Finally, the Prediction layer of shape n (number of classes) and B, which creates a discrete prediction for each DR by an argmax function.

In all the experiments, the fine-tuning of the specific PLM involves the use of a supervised learning approach, which is literally a training that results in a new PLM to be used for the encoding of the test set. The other aspects of the experiment setup are as follows:

- The classification task is conducted separately for all DR realization types and their Level -1 senses.
- The input is pairs of arguments (Arg1 and Arg2), and the output is a label, such as a DR realization type label or a sense label annotated in the data.
- For each DR realization, the arguments and the discourse connective (if available) are concatenated into a single line to form text element of the each input data item.

² It forms the sequence of hidden states at the output of the last layer of the model.

- The DR realization type labels and the sense labels form the "category" feature of the input for both training and test phases.
- 4/5 of the dataset is used for training and the rest is used for testing.
- Hyper-parameter tuning is done empirically by repeating the steps below:
 - Tokenize maximum input sequence length (128) number of words from each text element with the BERT tokenizer and convert all into the indexes of the tokenizer vocabulary.
 - Pad or truncate the texts into the maximum length long vectors.
 - Create an attention mask and return a dictionary of outputs and convert each tokenized DR vector into a tensor.
 - Starting with the parameters in the bench-marking architectures of the best practices and fine-tune the PLM by training it with the *Bert_MultiClass* classification model, depicted in Figure 20, by using AdamW optimizer [73] with the learning rate of $5e - 5^3$.
- The DRs in the test set are encoded with the fine-tuned model and classified by using the same *Bert_MultiClass* classification model.

As shown in Table 22, the model has been magnified up to 768 hidden layers with more than 184 million of parameters for the improvement of results, by the virtue of the GPU memory size, which is fully utilized.

Table 22: BERT_MultiClass TensorFlow Model

Layer (type)	Output Shape	Parameter #
input_ids (InputLayer)	(None, 128)	0
bert (TFBertMainLayer)	TFBaseModelOutput- WithPoolingAndCross- Attentions(last_hidden- _state=(None, 128, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)	184,345,344
pooled_output (Dropout)	(None, 768)	0
category (Dense)	(None, Number of Classes (n))	4614 (n+n.768)
All parameters	: 184,349,958	(bert layer + category layer)
Trainable parameters	: 184,349,958	
Non-trainable parameters	: 0	

³ Train the models by monitoring the Micro-F1 score on the validation set and stop the training if there is no increase in 50,000 consecutive steps.

APPENDIX C

KAPPA ANALYSIS OVER CONFUSION MATRICES

The classifier results could be presented in the form of confusion matrices in order to provide valuable quantitative values, including the total number of samples, precision, recall and F1-Score values, but we also need an objective method to compare the performance of our models. For this purpose, we use the Cohen’s Kappa Association Coefficient (κ) [82].

For more than five decades κ has been used as an associating measure for providing an agreement score between two observers on a nominal scale and its formula is built upon an N by k observation matrix in which the elements n_{ij} represent the number of observers who assigned the i -th case in the j -th class:

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}, \quad P_i = \frac{1}{n(n-1)} \left(\sum_{j=1}^k n_{ij}^2 - n \right) \quad (3a)$$

$$P_o = \frac{1}{N} \sum_{i=1}^N P_i, \quad P_e = \sum_{j=1}^k p_j^2, \quad (3b)$$

where N is number all data samples annotated, p_j is the proportion of all assignments to the j -th class, P_i is the extent of agreement among the n observers for the i -th sample, P_o is the observed overall agreement, and P_e is the expected mean proportion of agreement due to chance [83].

So, the Kappa statistic is defined as the degree of actually attained agreement in excess of chance ($P_o - P_e$), normalised by the maximum agreement attainable above chance ($1 - P_e$) [84]:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (4)$$

The κ statistic reduces the ratings of two observers to a single number [85] by taking into account a priori distribution that does not affect the distribution of the predictions among the target classes. The bias between observers and the distribution of data across the categories (prevalence) affect κ in very complex ways [86] and κ ’s ability to compensate for random hits makes it an interesting alternative for measuring the success levels of classifiers. Especially in working with unbalanced data such as ours, the κ coefficient can be helpful in comparing the performance of classification models. For inter-annotator agreement evaluation, the range of κ coefficients extend from -1 to +1 such that -1, 0 and +1 indicate strong disagreement, chance-level agreement and strong agreement, respectively.

We calculated the κ coefficients of our monolingual and multilingual classification models using the values in the confusion matrices as follows [87]:

$$\kappa = \frac{\sum_{i=1}^m CM_{ii} - \sum_{i=1}^m Ci_{corr} \cdot Ci_{pred}}{N^2 - \sum_{i=1}^m Ci_{corr} \cdot Ci_{pred}}. \quad (5)$$

where

- CM_{ii} represents the diagonal elements of the confusion matrix,
- Ci_{corr} is the number of correct samples in the i -th class,
- Ci_{pred} is the number of predicted samples picked for the i -th class.

The confusion matrices of DR realization type classification experiments of TDB 1.2, CDB and PDTB 3.0, encoded with their respective monolingual BERT, and the experiment with the combination of three languages, encoded with the multilingual BERT (see Table 18 in Section 6.2) are plotted in Figures 22, 23 and 24.

In order to measure the κ coefficients of the values in confusion matrices by the Formula 5, we calculate a Random Accuracy by taking the sum of all multiplications of the number of correct samples in the i -th class (Ci_{corr}) with the number of each predicted sample picked for the i -th class (Ci_{pred}). Then, in order to calculate a κ coefficient for each confusion matrix, we divided the difference between the sum of all True Positives (the diagonal elements of the confusion matrix) and Random Accuracy into the difference between the square of the data sample size in the test (N) and Random Accuracy.

The results are given in Table 23, showing κ coefficient decreases by 1.8%, 0.9% and 3.5% for the experiments with multilingual PLMs. Considering the TDB 1.2’s loss (18 per thousand) as a very small rate, the results could imply that the loss of accuracy is negligible and the multilingual encoder cannot be regarded as inefficient.

Table 23: κ Coefficients of the Confusion Matrices in the Figures 22, 23 and 24 Calculated by the Formula 5

	TDB 1.2 (TURKISH)		CDB (CHINESE)		PDTB 3.0 (ENGLISH)	
	Mono	Multi	Mono	Multi	Mono	Multi
κ	0.597	0.579	0.535	0.526	0.612	0.577
κ Difference		-0.018		-0.009		-0.035

κ Coefficients of monolingual and multilingual DR realization type classification experiments are given for three languages as such:

Mono: The dataset is encoded with its monolingual BERT PLM

Multi: The dataset is encoded with the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages for the Cross-lingual Transfer experiment purpose.

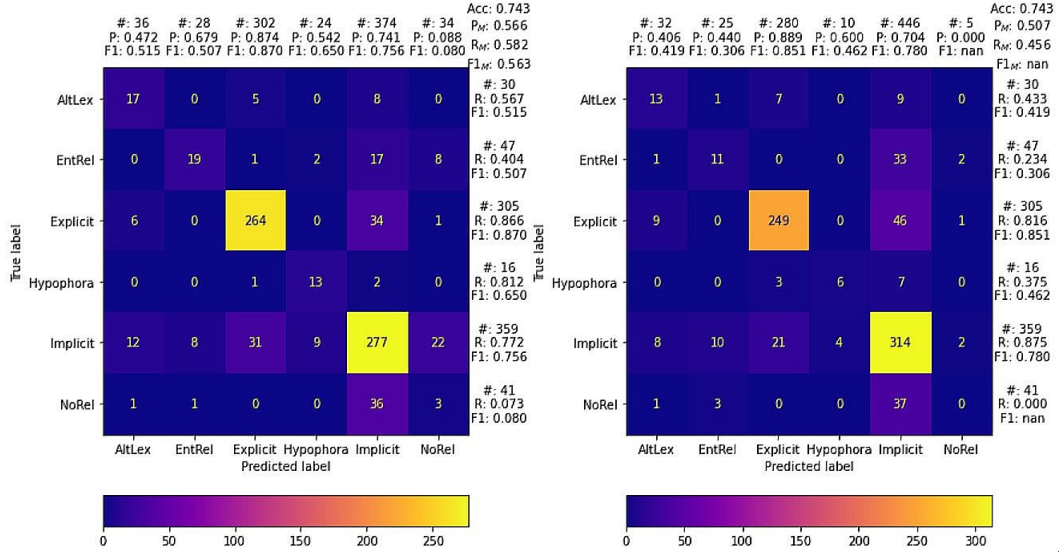


Figure 22: Confusion Matrices of Classification of DR Types in TDB 1.2 (Turkish) Encoded with BERT Turkish PLM (left) and the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages (right).

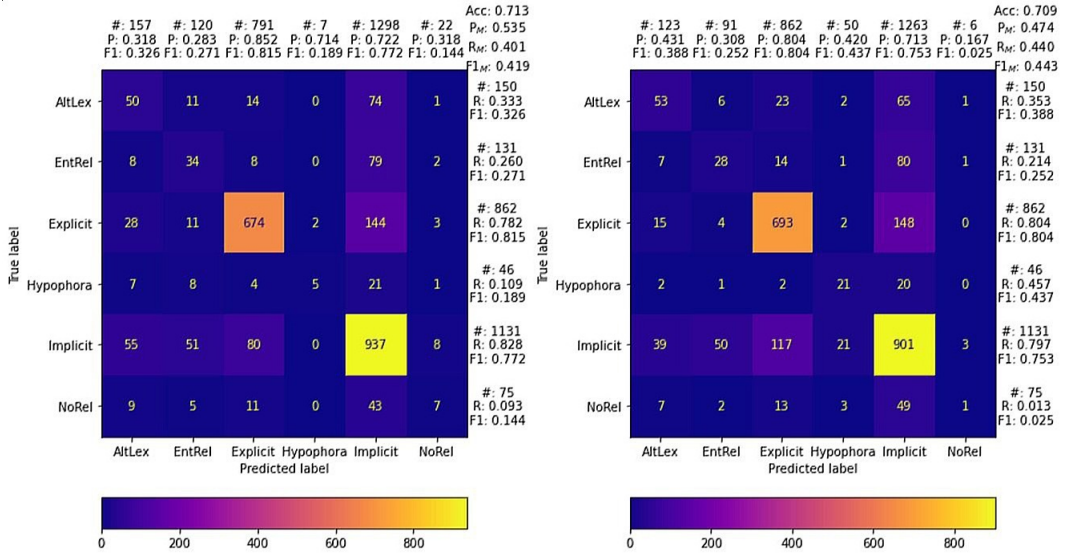


Figure 23: Confusion Matrices of Classification of DR Types in CDB (Chinese) Encoded with BERT Chinese PLM (left) and the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages (right).

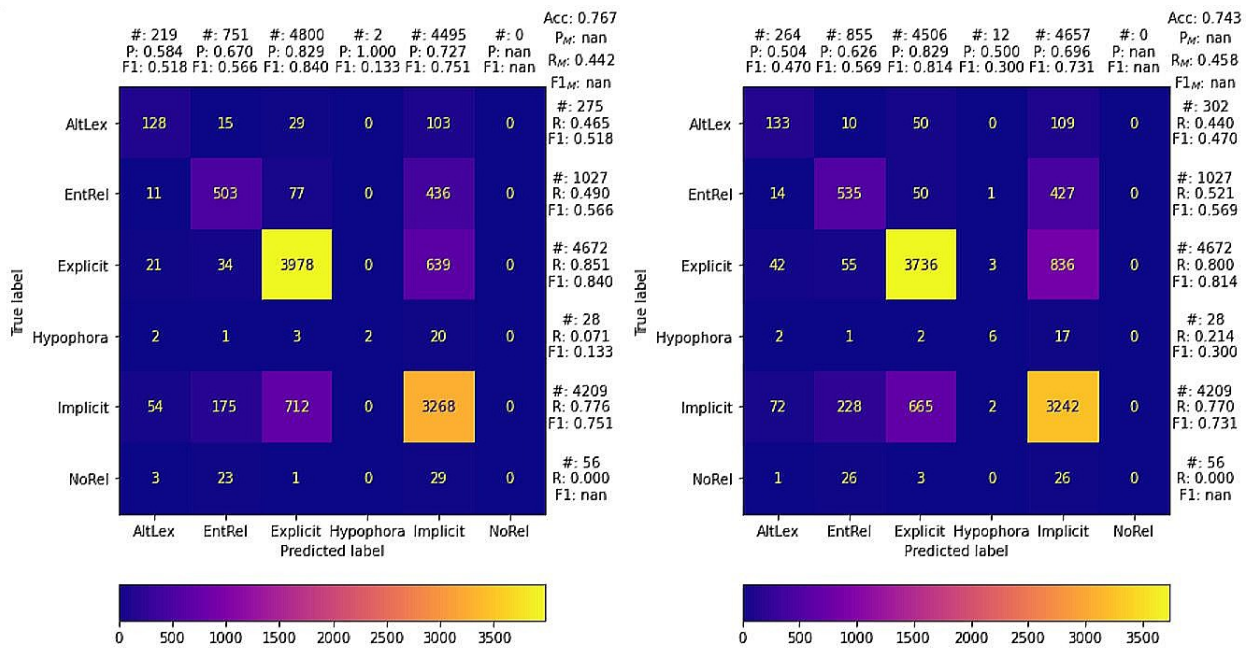


Figure 24: Confusion Matrices of Classification of DR Types in PDTB-3 (English) Encoded with BERT English PLM (left) and the multilingual BERT PLM which is newly fine-tuned by the contribution of these three languages (right).

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: KUTLU, Ferhat

Nationality: Turkish (TC)

Date and Place of Birth: October 17, 1970, Çivril/DENİZLİ

Marital Status: Married

Phone: +90 505 3170301

Fax: +90 0312 2103745

MASTER OF SCIENCE

Degree	Institution	Year of Graduation
M.S.	Bilkent University, Computer Eng. Dept.	2001
B.S.	Kara Harp Okulu, System Eng. Dept.	1998
High School	Kuleli Askeri Lisesi	1988

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
5	KoçSistem, İstanbul	R&D Project Manager
2	ENTES Electronics, İstanbul	R&D Dept. Tech. Manager
2	METU, Graduate School of Informatics	Research Assistant
11	Turkish Land Forces HQs	R&D Project Manager
6	Turkish Land Forces	Unit Manager

PUBLICATIONS

Journal Publications

F. Kutlu, D. Zeyrek and M. Kurfalı, Towards a Shallow Discourse Parser for Turkish, Natural Language Engineering (revised version under evaluation).

International Conference Publications

F. Kutlu and H. A. Güvenir, Categorization in a Hierarchically Structured Text Database, in Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2001), A. Acan, I. Aybay, and M. Salamah (Eds.), Gazimagusa / TURKISH REPUBLIC OF NORTHERN CYPRUS) (June 2001), 218-227.

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences

☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences

☐

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics

☐

Enformatik Enstitüsü / Graduate School of Informatics

☒

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences

☐

YAZARIN / AUTHOR

Soyadı / Surname : KUTLU

Adı / Name : Ferhat

Bölümü / Department : Bilişsel Bilimler EABD Başkanlığı

TEZİN ADI / TITLE OF THE THESIS (İngilizce / English):

Identification of Discourse Relations in Turkish Discourse Bank

TEZİN TÜRÜ / DEGREE: Yüksek Lisans / Master

☐

Doktora / PhD

☒

1. Tezin tamamı dünya çapında erişime açılacaktır. / Release the entire work immediately for access worldwide. ☒

2. Tez iki yıl süreyle erişime kapalı olacaktır. / Secure the entire work for patent and/or proprietary purposes for a period of two year. * ☐

3. Tez altı ay süreyle erişime kapalı olacaktır. / Secure the entire work for period of six months. * ☐

* Enstitü Yönetim Kurulu Kararının basılı kopyası teze birlikte kütüphaneye teslim edilecektir.
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature

Tarih / Date 25 Ocak 2023