Doctoral Thesis

Metaphor Processing in Tweets

Omnia Zayed

Supervisors

Dr. Paul Buitelaar

Dr. John P. McCrae

Examiners

Dr. Ekaterina Shutova

Dr. Edward Curry

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

Insight SFI Research Centre for Data Analytics
Data Science Institute
College of Science and Engineering,
National University of Ireland Galway

March 15, 2021

DECLARATION

I, Omnia Zayed, declare that this thesis, titled "*Metaphor Processing in Tweets*", is composed by myself, and that the work contained herein is my own except where explicitly stated otherwise in the text. I also confirm that appropriate credit has been given within this thesis where work of others has been consulted and that all main sources of help have been acknowledged. Finally, I confirm that this work has not been submitted for any other degree or professional qualification.

Galway, March 15, 2021

Omnia Zayed

To those who love metaphors, to you who are reading this...

ACKNOWLEDGEMENTS

A PhD is a life-changing journey that comes with mental challenges especially when pursued abroad away from home, family and friends. I enjoyed most of it and I enjoyed the quite long ride studying this fascinating figure of speech, metaphor. I won't deny that destiny played a role in choosing to study metaphors in English text. I was amazed by figures of speech, specifically idioms. Then, I crossed path with Prof Katja Markret during my time in Germany. She advised me to study the very interesting paper by Shutova et al. 2013 about statistical metaphor processing. And I have to admit, I fell in love with this intriguing phenomenon and the challenges of its computational modelling. I am really thankful to Prof Markret and for this unplanned chance.
Before saying goodbye to this chapter of my life, I would like to express my deep appreciation to many people who supported me throughout this adventure both professionally and emotionally.

This work would not have been possible without the support and guidance of my amazing supervisors, Dr Paul Buitelaar and Dr John McCrae, who motivated, pushed, taught, believed and helped me throughout this tough and long journey. Throughout the years, they were always keen to support me with their invaluable experience, ideas and, above all, time and patience.

Special thanks to my supportive Graduate Research Committee (GRC) members, Prof John Breslin, Dr Edward Curry and Dr Adegboyega Ojo, who witnessed the progress of this work over the years and supported me with their insightful discussions and advice.
I would also like to extend my appreciation to Dr Ekaterina Shutova, who acted as the external examiner of this thesis. I was highly fascinated and challenged by her research in the area of metaphor processing. I would like to thank her for the insightful feedback and discussion that she provided during the viva.

I am lucky to have many friends who supported me and stood by my side. Mennatullah Siam, my dearest and closest friend, whose support can not be described in words. I would especially like to thank Salma Abdulaziz, whom I met during my early days in Galway, and we became good friends over the years. Her cheerful and positive personality supported me in a way that she might not even imagine. I also want to thank Shereen Esmael for being by my side in the hard times before the good ones. Finally, I am thankful to Reham, Marwa, Ghada and Mai, as well as Ethar, Shaimaa and Rasha, for their moral support and encouragement.

I would like to thank my wonderful colleagues at the Data Science Institute (formerly Insight, Galway) generally and my team at the Unit of Natural Language Processing specifically, whom some of them became dear friends. They supported me over the years in many ways that probably they did not even realise, maybe through a passing comment or advice, a word of support, or even a friendly smile.
Special thanks to Cecile Robin, Sapna Negi, James O'Neill, Ian Wood, Fionn Delahunty, Sina Ahmadi, Mariam Masoud, Daniel Torregrosa, Joana Barros, Nivranshu Pasricha, Tobias Daudert, Marc Mellotte and Bianca Pereira. I wish also to thank Hazem Safwat, Ihab Salawdeh, Manel

Zarrouk and Wassim Derguech and his wife Mejda Allani.
I am also very grateful to Mihael Arcan, Housam Ziad and Sameh Mohamed for their continuous support, advice and guidance since the beginning of this journey.

I would like to extend my sincere thanks to my lovely landlady, Mary Gilson and her son Tomás. Although they needed the apartment, they accepted the extension of my rental agreement for a couple of months once they heard that I am finalising my thesis.

Finally, A great deal of gratitude and appreciation goes to my family, my mum (the first linguist in my life), my dad and my brother. They have been standing by my side, even when a continent was between us. I would have never gotten where I am today without their support, trust, guidance and, above all, love. My twin sister, Doaa, deserves a separate word of appreciation for her indefinite care and love. Her support has been beyond measure and I would probably need pages and ages to describe.

The work presented in this thesis was supported by Science Foundation Ireland under grant numbers SFI/12/RC/2289 and SFI/12/RC/2289_2 (Insight).

Omnia Zayed
Galway, Ireland
March 15, 2021

PREFACE

We are very happy to introduce the PhD thesis by Omnia Zayed on metaphor processing in tweets, which makes significant contributions to the definition and processing of metaphor in social media language data and beyond. First of all, the thesis work produced several new metaphor data sets that take a principled relational-level view of metaphor and in addition several existing word level datasets were extended in the context of this thesis with a relational-level view annotation. A second focus of the work presented here has been on the development of novel methods for metaphor identification, specifically in the context of tweets but with applications beyond this. Significant improvements were obtained by using an innovative neural architecture for metaphor identification based on contextual modulation that borrows from visual reasoning by the use of affine transformations. Finally, a further focus has been on improving methods for metaphor interpretation, which in the context of this thesis was reformulated as definition generation. The approach taken here uses a sequence-to-sequence language model with a dual encoder-decoder architecture and contextualised sentence embeddings to represent metaphorical expressions and corresponding target definitions. The thesis work has been published at several high impact venues and will certainly influence other ongoing and future work in the field of metaphor processing, social media analysis and natural language processing in general.

Galway, March 15, 2021

Dr. Paul Buitelaar, Dr. John McCrae

ABSTRACT

Metaphor plays an important role in defining the interplay between cognition and language. Despite its fuzziness, this ubiquitous figurative device is an essential element of human communication that allows us (as humans) to better understand and, thus, communicate unfamiliar experiences and concepts in terms of familiar ones. Metaphor comprehension and understanding is a complex cognitive task that includes grasping the interaction between the underlying concepts. This is very challenging for humans, let alone computers. The last few decades have witnessed a growing interest in automating this cognitive process by introducing a wealth of ideas to model the computational recognition and comprehension of metaphors in text. Many approaches and techniques have been introduced to explore the automatic processing of different types of metaphors and the preparation of metaphor-related resources.

In spite of the attention that metaphor processing has gained recently, the majority of existing approaches do not process metaphors in informal settings such as social media. Twitter offers a novel way of communication that enables users all over the world to share their thoughts and experiences. The social media content circulated on this platform through the short informal tweets poses a challenge for automatic language processing due to the unstructured nature and brevity of the text as well as the vagueness of topics. Such unique characteristics of tweets, coupled with the importance of studying metaphoric usage on social media motivated me to study metaphor processing in such a context. Metaphor processing in tweets can be beneficial in many social media analysis applications, including political discourse analysis and health communication analysis.

In this thesis, I investigate the automatic processing of metaphors in tweets focusing on two main tasks, namely metaphor identification and interpretation. My aim is to improve metaphor identification to study the usage of metaphoric language in healthcare communication and political discourse in social media. Furthermore, I aim to improve metaphor interpretation to aid language learners and to enrich lexical resources. I, therefore, study various NLP and deep learning techniques to automatically identify and interpret metaphors in tweets. To the best of my knowledge, there has been no attempt to process metaphors in tweets in part due to the lack of tweet datasets annotated for linguistic metaphor. Thus, the focus of the work presented here is not only introducing models to process metaphors in tweets but also developing the necessary datasets.

Overall, the work is divided into three main research themes; the first focuses on the development of metaphor annotation schemes and the preparation of datasets for both tasks. The second is concerned with the automatic identification of linguistic metaphors in tweets under a relational paradigm which explores three main ideas, namely distributional semantics, meta-embedding learning and contextual modulation. Finally, the last theme focuses on metaphor interpretation along the more complex "definition generation" approach, which provides full explanation of a given metaphoric expression. Experiments are conducted on the introduced datasets of tweets as well as benchmark metaphor datasets to show the effectiveness of the proposed approaches. Furthermore, the proposed datasets and the best models from this thesis will be made publicly available to facilitate research on metaphor processing in general and in tweets specifically.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

"How do people use words to make meanings? This is a question that has fascinated and baffled thinkers who have concerned themselves with the nature of human language, from Plato to the present day."

(Hanks, 2013)

1.1 MOTIVATION

Metaphor is an essential element of human communication, especially in informal settings such as social media. Despite its fuzziness, metaphor is a fundamental feature of language that orchestrates the relation between cognition and language. This ubiquitous figurative device allows us (as humans) to better understand and, thus, communicate unfamiliar experiences and concepts in terms of familiar ones. Over the last couple of decades, there has been an increasing attention towards metaphor processing and its applications, either as part of natural language processing (NLP) tasks such as machine translation (Koglin and Cunha, 2019), text simplification (Wolska and Clausen, 2017a; Clausen and Nastase, 2019) and sentiment analysis (Rentoumi et al., 2012) or in more general discourse analysis use cases such as in analysing political discourse (Charteris-Black, 2011), financial reporting (Ho and Cheng, 2016) and health communication (Semino et al., 2018).

Twitter offers a novel way of communication that enables users all over the world to share their thoughts, experiences and ideas. The social media content circulated on this platform through the short informal massages, namely tweets, poses a challenge for automatic language processing due to the unstructured nature and brevity of the text as well as the vagueness of topics. Such unique characteristics of tweets and the availability of Twitter data motivated me to study metaphor processing in such a context. Metaphor processing in tweets can be very useful in many social media analysis applications including, as hinted earlier, political discourse analysis and health communication analysis. Also, it will open a wide range of possibilities for researchers from various fields such as linguists, sociologists, psychologists and political scientists to scientifically study and explore metaphoric usage on social media which will allow deeper understanding of the language on many computational and linguistic levels.

This thesis focuses on the automatic processing of metaphors in tweets with the aim of improving many applications such as analysing healthcare communication and political discourse in social media as well as language learning and lexical resources development. The introduced work focuses on two main metaphor processing tasks, namely, metaphor identification and interpretation. Despite the variety of approaches for processing metaphor, there is still a need for better models that mimic human cognition. The key question is how to design a system that can

generalise well beyond the level of metaphoric analysis and the text genre. In this thesis, I study various NLP and deep learning techniques to automatically identify and interpret metaphors in tweets. Furthermore, I highlight the encountered methodological challenges and limitations.

To the best of my knowledge, there has been no attempt to identify and interpret metaphors in tweets in part due to the lack of tweet datasets annotated for linguistic metaphor. This is one of the challenges that this thesis addresses for both metaphor identification and interpretation. One of the contributions of this work is to create and publish annotated datasets to facilitate the research on metaphor identification and interpretation in tweets. Among the other contributions of this work is to design, build and evaluate computational models that identify and interpret metaphors in tweets. Throughout this thesis, I, first, explore previously introduced methods and analyse their limitations. Then, I focus on developing deep learning approaches to identify metaphors and interpret their intended meaning in a given context. Furthermore, I develop the necessary annotated datasets of tweets for both the metaphor identification and interpretation tasks. The objectives and contributions of the work presented in the context of this thesis will be discussed in detail in the next sections.

1.2 RESEARCH OBJECTIVES AND CONTRIBUTIONS

This thesis is guided by three main research themes focusing on two main tasks of metaphor processing, which are metaphor identification and metaphor interpretation. As mentioned earlier, the main scope of this work is to study metaphor processing in tweets. The following subsections discuss the research themes and the associated research questions in detail.

1.2.1 Resource Preparation

One of the main challenges that face the research of metaphor processing is the corpora availability and preparation. The majority of previous approaches pertaining to metaphor identification have focused on formal well-structured text selected from a specific corpus to create datasets to model and evaluate their approaches. A common issue of all the available datasets is that they are specifically designed for a certain task definition focusing on a certain level of metaphor analysis which makes their annotation schemes difficult to generalise. Additionally, the majority of available datasets lack coverage of metaphors and text genres as they rely on predefined examples of metaphors from a specific domain during the creation process. On the other hand, automatic metaphor interpretation is much less explored in part due to the lack of publicly available datasets. To the best of my knowledge, there exist only two datasets that were prepared in the context of metaphor interpretation for different task formulations. These available datasets have important limitations in terms of size, representativeness and quality. This lack of reliable annotated datasets hindered the progress in this area.

Manually annotating a dataset for either metaphor identification or interpretation is a very demanding task which requires effort and time from a human annotator to first identify the metaphoric expression then to figure out its meaning and finally to provide a literal explanation (if possible) for it. Besides, it is a highly subjective task that depends on the context where the expression occurs and the followed definition of metaphor as well as the cultural background

of the annotator. Several aspects should be considered while preparing datasets for metaphor processing, starting from the data selection until developing a reliable annotation scheme, in order to ensure creating a high-quality dataset of tweets annotated for linguistic metaphors. In the light of this, this research theme seeks to answer the following research questions:

RQ1.a Can metaphor be defined in such a way that results in a high inter-annotator agreement?

RQ1.b How to adapt existing benchmark datasets to better suit relation-level metaphor identification using minimal annotation effort while maintaining annotation accuracy and consistency?

RQ1.c How can lexical resources be employed to prepare a dataset of reliable definitions (interpretations) of metaphoric expressions?

The main goal of the work done under this research theme is to create and publish annotated datasets to facilitate the research on metaphor identification and interpretation in tweets. I summarise the main contributions under this research theme as follows:

- Proposing an annotation scheme that results in an expert annotated dataset of tweets for metaphor identification on the relation level with the aim of achieving large coverage of metaphoric usages and text genres while maintaining high annotation accuracy.
- Adapting existing word-level benchmark datasets for relation-level metaphor identification by employing a semi-automatic approach to avoid the need for extensive manual annotation and to facilitate future research in relation-level metaphor processing.
- Introducing an annotation scheme for metaphor interpretation by casting the task as definition generation and employing this scheme to create a dataset of metaphor interpretation focusing on verb-noun expressions.

1.2.2 Metaphor Identification in Tweets

This research theme focuses on metaphor identification which is the most studied among the metaphor processing tasks. Identifying metaphors in text is an essential step which supports other metaphor processing tasks as well as more specific NLP applications. It is necessary to recognise the metaphor in text in order to interpret it or discern its underlying source-target relation or even translate it or analyse its sentiment. This task can be addressed at different levels of granularity, namely at the sentence level (a sentence is metaphorical or not) or at the word level (a given word is used metaphorically or literally) or at the relational level (the grammatical/semantic relationship between a pair of words, such as a verb and its noun object or a noun and its adjectival modifier, is metaphoric or literal). In the context of this thesis, I will focus on identifying linguistic metaphors in tweets by adopting the relational paradigm. My main aim is to model the interaction between the metaphor components in order to capture the metaphoricity in a way that mimics the human comprehension of metaphors.

To this end, I first explore the use of distributional semantics in the identification of metaphors on the relational level. Then, I study the features employed in the literature to identify metaphors in text focusing on defining the limitations of the state-of-the-art approaches either on the word or relation levels. Finally, I address these limitations by introducing a novel architecture

for identifying relation-level metaphoric expressions of certain grammatical relations based on contextual modulation. The work presented under this research theme is formulated as three research questions as follows:

RQ2.a To what extent can a minimally supervised approach based on distributional representation accurately identify metaphors in short texts?

RQ2.b Can employing an ensemble of linguistic and advanced contextual features to learn meta-embeddings in a neural architecture improve metaphor identification in tweets?

RQ2.c Can contextual modulation improve the performance for relation-level metaphor identification?

The answers to these questions can be expressed in terms of the following contributions:

- Employing distributional semantics to introduce a semi-supervised approach to identify metaphors in text with the aim of aiding in the creation and annotation of a dataset of tweets annotated for linguistic metaphors on the relation level.
- Studying the effectiveness of an ensemble of features to identify linguistic metaphors of certain grammatical relations by utilising meta-embedding learning methods.
- Proposing a novel approach for context-based textual classification based on contextual modulation through affine transformation and apply it on relation-level metaphor identification.

1.2.3 Metaphor Interpretation in Tweets

Metaphor comprehension and understanding is a complex cognitive task that includes interpreting metaphors by grasping the interaction between the meaning of the metaphor components. It is a very challenging task for humans and, as discussed earlier, understanding the intended meaning of a metaphor is highly subjective and depends on various linguistic, cultural and psychological aspects. Various task formulations have been adopted by previous approaches that addressed the automatic interpretation of metaphors. In this thesis, I define metaphor interpretation as a definition generation task with the aim to aid language learners and non-native speakers to understand metaphors as well as enrich the process of developing lexical resources. I investigate the feasibility of such formulation through employing advanced neural models trained on benchmark datasets of definitions to automatically interpret and explain the intended meaning of a metaphor. Therefore, this research theme focuses on exploring the following research question:

RQ3 Can an advanced neural architecture be implemented to generate reliable definitions (interpretations) of metaphoric expressions that aid people in understanding them?

The following contributions are made under this research theme:

- Approaching the metaphor interpretation task as definition generation and investigating definition modelling of metaphoric expressions.
- Employing an attention-based sequence-to-sequence neural model that utilises a dual encoder architecture and contextualised sentence embeddings to interpret metaphors as they occur in text.

1.3 THESIS STRUCTURE

Chapter 2 sets the stage for the rest of the thesis by introducing the necessary background knowledge and technical concepts. It begins by presenting the linguistic definition of metaphor and highlighting its main components and various types followed by an overview of the prominent views and theories developed to understand its comprehension. The chapter also provides some other technical background related to dataset annotation, word embeddings and artificial neural networks as well as the evaluation strategies for metaphor identification.

Chapter 3 reviews the literature related to metaphor processing focusing on the related work corresponding to each research theme. Although this review is done from a chronological perspective, it also highlights the various adopted paradigms to process metaphors on the sentence, relation, and word levels in order to investigate how these paradigms affected the choice of approaches, developed architectures and selected features. Finally, the chapter summarises the strengths and limitations of the proposed approaches.

Chapter 4 investigates the first research theme related to resource preparation in the context of this thesis. It focuses on demonstrating the work done to prepare the necessary resources for the identification and interpretation of linguistic metaphors on the relation level in English text. This chapter covers **RQ1.a**, **RQ1.b** and **RQ1.c**.

Chapter 5 demonstrates the work done under the second research theme that focuses on metaphor identification in tweets. It explains the proposed approaches to identify metaphors on the relation level either using distributional semantics, meta-embeddings learning and contextual modulation in detail. It also discusses the conducted experiments either on benchmark datasets or on the metaphor dataset of tweets developed as part of this thesis. The work presented in this chapter seeks answers to **RQ2.a**, **RQ2.b** and **RQ2.c**.

Chapter 6 seeks an answer to **RQ3** under the third, and last, research theme in this thesis. This chapter focuses on metaphor interpretation in tweets by casting the task as definition generation. It explores the feasibility of such task formulation by studying definition modelling. It then proposes a neural approach based on a sequence-to-sequence architecture that employs a dual encoder. The chapter provides the conducted experiments on the proposed metaphor interpretation dataset as part of this thesis.

Chapter 7 provides the main conclusions of this thesis and summarises the main contributions. Furthermore, it discusses the potential future directions.

1.4 PUBLICATIONS

Some of the work done in the context of this thesis has been published in relevant venues as follows:

- **Omnia Zayed**, John P. McCrae, and Paul Buitelaar. 2018. Phrase-level metaphor identification using distributed representations of word meaning. In Proceedings of the First Workshop on Figurative Language Processing, pages 81–90, New Orleans, LA, USA.
- **Omnia Zayed**, John P. McCrae, and Paul Buitelaar. 2019. Crowd-sourcing a high-quality dataset for metaphor identification in tweets. In Proceedings of the 2nd Conference on Language, Data and Knowledge, LDK '19, pages 10:1–10:17, Leipzig, Germany.

- **Omnia Zayed**, John P. McCrae, and Paul Buitelaar. 2020a. Adaptation of word-level benchmark datasets for relation-level metaphor identification. In Proceedings of the Second Workshop on Figurative Language Processing, pages 154–164, Online.
- **Omnia Zayed**, John P. McCrae, and Paul Buitelaar. 2020b. Contextual modulation for relation-level metaphor identification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Findings of of the Association for Computational Linguistics: EMNLP 2020, pages pages 388-406, Online. Association for Computational Linguistics.
- **Omnia Zayed**, John P. McCrae, and Paul Buitelaar. 2020c. Figure Me Out: A gold standard dataset for metaphor interpretation. In Proceedings of the 12th International Conference on Language Resources and Evaluation, LREC '20, pages 5810–5819, Marseille, France. European Language Resources Association.

2 BACKGROUND

"The essence of metaphor is understanding and experiencing one kind of thing in terms of another."

(Lakoff and Johnson, 1980)

This chapter discusses the background of the work presented in this thesis. The chapter starts by laying the foundation of what is meant by *"metaphor"* in this thesis with the aim to understand how we (as humans) understand, recognise and interpret metaphors in discourse. It begins with outlining the linguistic definition of a metaphor in the literature. Then, in a journey that goes back to Aristotle, the first part of this chapter navigates through the different views and theories of metaphor followed by an investigation of its different types. The differences between metaphor and other figures of speech will also be highlighted. Finally, the first part of this chapter concludes with outlining the definition, theory/view and type of metaphor adopted throughout this work. The second part of this chapter focuses on explaining some technical concepts and background used in the context of this thesis which are related to dataset annotation, word embeddings and artificial neural networks as well as the evaluation strategies for metaphor identification.

2.1 METAPHOR BACKGROUND

2.1.1 Metaphor Definition

Metaphor is a ubiquitous figurative device that represents the interaction between cognition and language. Despite its fuzziness, it is an essential element of human communication that defines the relation between how we understand things and how we express them (Cameron and Low, 1999). Metaphoric expressions allow us to express ideas, emotions and experiences that might be difficult to express using literal language. This section focuses on the definition of *"metaphor"* from a linguistic point of view.

The word metaphor comes from the ancient Greek word *metaphorá* which means "to transfer" or "to carry accross". Therefore, the simplest definition of metaphor is to transfer the name and attributes of something to another. Generally, a metaphor has two main components: the *tenor* and the *vehicle*. The *tenor* represents the topic of the metaphor while the *vehicle* is the term used metaphorically. The relation between them is defined by the *ground* which also gives the metaphor its meaning. Perceiving these components is essential to fully comprehend the metaphor. At least one of the two metaphor components must be explicitly present in the sentence; whereas the other can be represented by its properties (End, 1986). This led to a distinction into different types

of metaphors as will be explained in Section 2.1.3. But to complete the picture, Black (1962) listed some examples which he called "clear cases" of metaphor in order to allow the readers to agree on what could be the definition of metaphor or as he put it *"to analyze the notion of metaphor"*. I quote some of his examples as follows:

(2.1) *"The chairman plowed through the discussion."*

(2.2) *"A smoke screen of witnesses."*

(2.3) *"An argumentative melody."*

(2.4) *"Blotting-paper voices." (Henry James)*

(2.5) *"The poor are the negroes of Europe." (Chamfort)*

(2.6) *"Light is but the shadow of God." (Sir Thomas Browne)*

(2.7) *"Oh dear white children, casual as birds, Playing amid the ruined languages." (Auden)*

In the above examples, although the whole sentence is given as an instance of "clear case" metaphor, it can be agreed that the words *"plowed"*, *"smoke screen"*, *"argumentative"*, *"blotting-paper"*, *"negroes"*, *"shadow of god"* and *"ruined"* are the crucial words with the metaphoric sense that grabs the attention (Black, 1962). These words convey a metaphoric sense in these examples, and yet they could be used in their literal sense in other examples.

A conclusion from the previous discussion is that, a given word (lexical unit[1]) with a literal sense can also have a metaphorical one and both senses can be employed based on the context and the author's intentions. Moreover, a given sentence can contain some words used metaphorically and others used literally. This will take us to how metaphor, as a phenomenon, is perceived and analysed by philosophers and linguists and how that shaped the recent notion of metaphor.

2.1.2 Theories of Metaphor

Scholars formulated various theories and views of metaphor in order to define it and explain its comprehension. These theories and views can be traced back to Aristotle who gave his classic definition of metaphor as *"the transference of a name from genus to species, from species to genus, from species to species, or by analogy."* (Poetics XXI, 1457b). In other words, Aristotle simply defines a metaphor as the transfer of a name and its associated properties to another. This definition is then confirmed by Quintilian, in Institutio Oratoria[2] VIII, VI, 1, who defined a trope (focusing on metaphor) as *"the artistic alteration of a word or phrase from its proper meaning to another."* (Coulson, 2009). Based on these propositions, a number of linguists, philosophers and psychologists have developed various theories of metaphor which either contradict or support the Aristotelian view of metaphor. These views and theories are studied in extensive detail in (End, 1986; Gibbs, 1992; Shutova, 2011; Holyoak and Stamenković, 2018; Rai and Chakraverty, 2020). The next subsections will focus on the prominent ones that governed metaphor comprehension in the past few decades, which are: the comparison view, the interaction theory, the anomaly theory and the conceptual metaphor theory.

1 A lexical unit in the context of metaphor annotation is defined by Steen et al. (2010) and it can be a single word or more than one word such as multi-word expressions, compound nouns and phrasal verbs.

2 https://penelope.uchicago.edu/Thayer/E/Roman/Texts/Quintilian/Institutio_Oratoria/home.html

2.1.2.1 *The Substitution and Comparison Views*

As originally proposed by Aristotle, metaphor can be seen through substitution, comparison and similarity. A literal word expressing a certain topic can be replaced by one that literally does not express that topic. This borrowed word shares some properties with the topic (End, 1986) or in other words "resonates" with it (Black, 1962; Hanks, 2006). The common properties between the borrowed word (metaphor) and the topic is defined through the *ground* which orchestrates the linguistic relationship between the *tenor* and the *vehicle*. According to the comparison theory, a metaphor can be seen as an implied comparison or an analogy between two conceptual domains which are the source domain (*vehicle*) and the target domain (*tenor*)[3]. This view of metaphor was developed further by Gentner who introduced the notion of structure-mapping in order to capture the meaning of analogy and metaphor through the meaning of its parts (Gentner, 1983; Gentner and Clement, 1988; Gentner et al., 2001). Gentner highlighted the transfer of knowledge, in terms of attributes and relations, from one domain (source domain) into another (target domain).

2.1.2.2 *The Interaction View*

The interaction view of metaphor is considered as one of the prominent views on metaphor. The origins of this view go back to Richards (1936) and were then further developed by Black (1962). Black opposes the Aristotelian view of a metaphor as a comparison or condensed analogy. He proposed that a metaphor is the result of an interaction between two thoughts (subjects) and emphasised that some metaphors create new similarities, in addition to previously existing ones, between the two thoughts and these similarities provide the *ground* for the metaphor (Black, 1993; Holyoak and Stamenković, 2018). Therefore, this view focuses on the cognitive aspect of a metaphor by allowing for the possibility of introducing novel and new creative meanings of a given metaphor through the interaction of its components.

2.1.2.3 *The Anomaly View*

As opposed to the comparison view, the anomaly view of metaphor postulates that a metaphor represents a violation of semantic rules. Many scholars including Levin (1977) and Wilks (1978) adopted the semantic deviance view of metaphor and perceived a metaphor as a violation of selectional restrictions or preferences. This view proposes that a metaphor violates certain linguistic norms and rules (Wilks, 1978) and if it were interpreted literally it would be anomalous on the semantic, grammatical and conceptual levels (Gibbs, 1992). The examples *"My car drinks gasoline."*, *"... a Scottish Assembly should be given no executive powers ..."*, and *"... the line taken by the Shadow Cabinet ..."*, from (Wilks, 1978), illustrate the metaphoric usages of the given verbs as violations of certain semantic preferences.

2.1.2.4 *The Conceptual-Structure View*

Lakoff and Johnson (1980) introduced a conceptual (cognitive) view of metaphor based on the idea that our experiences, thoughts and actions are structured metaphorically. They introduced the conceptual metaphor theory[4] (CMT), and, unlike prior views which identify metaphors at the lexical level, they defined metaphor on the conceptual level through cross-domain mapping

3 The source and target domains terms are commonly used in psychological work on analogy to describe the *tenor* and the *vehicle*, respectively.

4 Also referred to as the cognitive metaphor theory.

or source-target mapping. The term *conceptual metaphor* is coined to represent an underlying conceptual mapping between the source and target domains. For example, a concept such as *"fragile object"* (source domain/vehicle) can be borrowed to express another such as *"emotions"* (target domain/tenor). This conceptual metaphor *"Emotions are Fragile Objects"* can be expressed in our everyday language in terms of linguistic metaphors such as *"shattered my emotions"*,*"break his soul"*,*"crushed her happiness"*, *"fragile emotions"* and *"brittle feelings"*. The various types of linguistic metaphors will be explained in the next section.

2.1.3 Types of Metaphor

As can be concluded from the previous discussion, a metaphor consists of a particular *tenor* and a *vehicle*. And the relation between them is provided by the *ground* which holds preexisting (or new) similarities, common properties and attributes. Either the *tenor* or the *vehicle* should be explicitly present in the sentence which leads to a distinction into different types of metaphors that can be summarised as follows:

2.1.3.1 *Conceptual and Linguistic Metaphor*

Lakoff and Johnson (1980) analysed the relation between the *tenor* and the *vehicle* through corpus studies. They proposed that linguistic metaphors, used in our everyday language, can be grouped together under what they called *conceptual metaphor*. Based on this study Lakoff et al. (1991) introduced the Master Metaphor List (MML) which is a collection of conceptual metaphors, representing cross-domain mappings, with the corresponding linguistic metaphors used in language. Here, I revisit a famous example in the literature (Lakoff and Johnson, 1980) of a *conceptual metaphor* and the corresponding *linguistic metaphors* that can be grouped under it as follows:

(2.8) *Argument is War*

a. Your claims are *indefensible*.

b. I *demolished* his argument.

c. I've never *won* an argument with him.

d. He *attacked* every weak point of my argument.

e. His criticisms were right on *target*.

f. He *shot down* all of my arguments.

Linguistic metaphor can exhibit different forms. The most common types are: lexical, multi-word and extended metaphors. I give an overview of them as follows:

1. **Lexical Metaphor**: The metaphoric sense is at the level of a single word (i.e. single *vehicle* term). Lexical metaphors can be sub-divided further based on their syntactic structure as:

   (a) *Nominal metaphors*: Also referred to as direct, explicit or analogical metaphors. This type focuses on noun metaphors where both the *tenor* and the *vehicle* are explicitly stated. Thus, the mapping between them is obvious and explicit. In this type, the *tenor* is usually compared with the *vehicle* through a copular verb such as *to be*. Examples 2.5 and 2.6 are clear cases of this type of metaphor as well as the following examples:

(2.9) *Juliet* is the *sun*.

(2.10) My *lawyer* is a *shark*.

A complex formulation of the nominal form is where the *vehicle* (metaphoric word) is in relation to another unstated concept from the target domain; this is called *proportional metaphor* such as the following example:

(2.11) Religion is the *opium of the people*.

This example is considered a four term analogical metaphor between since it compares four different entities which are religion, opium, people and addicts (i.e. religion to people is like opium to addicts) (Holyoak and Stamenković, 2018).

(b) *Predicate metaphors*: This type focuses on verb metaphors where the target domain is explicit whereas the source domain is implicit and is represented by its properties. Therefore, the mapping (relation) between the *tenor* and the *vehicle* is also considered implicit. Instances of this type include Examples 2.1, 2.8a, 2.8b, 2.8c and 2.8d as well as the following:

(2.12) I have to *swallow* my anger.

(2.13) Patients *see* hope in new treatment.

(2.14) This letter *stirred* my emotions.

(2.15) ... to *defend* this argument.

(2.16) ... *cruise to* victory.

This type is studied by either analysing the verb on the lexical level or by analysing the verb given a specific syntactic construction or grammar relation. For example, analysing the metaphoricity of the verb in a subject-verb-object (SVO) syntactic structure such as Example 2.14 in which the metaphoricity of the verb *"stirred"* comes from the subject *"letter"* and the object *"emotions"*. A complex formulation of this type is when the *tenor* is of a complex relation e.g. a nominal modifier such as in the following example:

(2.17) Doctors *see rays of hope* in the new treatment.

In this example, the metaphoricity of the verb *"see"* depends on the relation of *"rays"* to *"of hope"* which is fundamental in this case compared to, for example, *"patients see rays of light"*.

(c) *Attributive metaphor*: This type focuses on the metaphoric usage of adjectives in an adjective-noun syntactic relation. The adjective *"argumentative"* in Example 2.3 is used metaphorically with the noun *"melody"*. Other examples include:

(2.18) The child has *fragile* emotions.

(2.19) A candidate with a *colorful* personality.

(2.20) It is time to accept the *bitter* truth.

(d) *Adverbial metaphor*: This type focuses on the metaphoricity of an adverb in an adverbial phrase. Examples include:

(2.21) The two arguments are explained *separately*.

(2.22) Unemployment was a *heavily* politicized issue.

(e) *Prepositional metaphor*: Metaphor can be found in locative prepositional phrases as well. Examples include the metaphoric sense of the prepositions *"up"* and *"on"* in the following phrases:

(2.23) ... get my spirit *up*.
(2.24) ... *on* Monday morning.

The first four types are formulated in the literature as *Type-I–IV* metaphors (Krishnakumaran and Zhu, 2007; Neuman et al., 2013; Rai and Chakraverty, 2020) formulated. It is worth mentioning that *Types I–III* are the most studied by current computational processing approaches. This is supported by corpus-based studies as there has been an interest to investigate the relation between particular lexico-grammatical features such as syntactic structure (word classes) and metaphoricity. Goatly (1997) and Cameron (2003) studied the distribution of metaphors in a particular domain and across different text genre. According to their quantitative observations, metaphors exhibit unequal distribution across various word classes. In both studies, verbs (*predicate metaphor*) account for a larger percentage of the studied metaphors followed by nouns (*nominal metaphor*) and then adjectives and adverbs. Furthermore, Jamrozik et al. (2013) investigated the metaphoricity of relational words by analysing the usage of verbs, relational nouns, and entity nouns in a corpus search. The study shows that relational words exhibit higher metaphorical potential than entity words.

2. **Multi-word Metaphorical Expressions**: These expressions consist of multi-word vehicle terms such as compound nouns and phrasal verbs such as Example 2.8f or the following example:

(2.25) ... *wash off* all your sadness

3. **Extended Metaphor**: Usually used in literary writings and spans over longer discourse fragments such as multiple sentences, a paragraph or a poem. Shakespeare's famous quote, in his play *As You Like It*, is an example of this type of metaphor, I quote it as follows:

(2.26) "All the world's a stage,
And all the men and women merely players;
They have their exits and their entrances,
And one man in his time plays many parts,
His acts being seven ages."

2.1.3.2 *Novel, Conventional and Dead Metaphor*

We employ and produce metaphor unconsciously and automatically in our everyday language. Some metaphoric expressions became so conventionalised in our everyday language due to their repeated use to the extent that they might be no longer recognised as metaphoric. The idea that metaphors are initially created as novel decorative devices after which they become conventional in our daily language and finally dead is explained by Nunberg (1987) who depicted the journey of a metaphor as:

> *"Metaphors begin their lives as novel poetic creations with marked rhetorical effects, whose comprehension requires a special imaginative leap. As time goes by, they become a part of general usage, their comprehension becomes more automatic, and their rhetorical effect is dulled."*

Goatly (1997) categorised metaphors as *dead, inactive* and *active*. He described metaphoricity as a continuum that ranges from *active* (i.e. creative) metaphors at one end, to the most *dead* ones at the other. The difference between *dead* and *inactive* metaphors is highlighted as the former are perceived by language users as homonyms (e.g. *"pupil"* referring to a student), whereas the latter are regarded as polysemes (e.g. *"crane"* referring to lifting machine). The comprehension of novel metaphors (poetic or newly produced expressions) might be different than the conventionalised ones (entrenched in our everyday language) since the comprehension of the latter became effortless and unconscious. Deignan (2005) highlighted the difference between novel and conventional metaphor by describing the innovative and unconventional usage of novel metaphor. The effect of novel metaphors will be observed by most readers since the *vehicle* is not regularly mapped onto the target domain unlike conventional metaphors which regularly used.

From a CMT point of view, conventional metaphors such as the verb *"grasp"* as in *"grasp the idea"*, the verb *"contain"* as in *"contain your emotions"*, and the adjective *"sweet"* as in *"sweet dreams"* are still alive and active in our conceptual system. Although they no longer require comprehension effort or imagination they still convey the linguistic and conceptual definition of metaphor. On the other hand, there exist some words that lost their literal sense and only the metaphoric sense being recognised such as *"impress"* (original literal sense is: to press; or apply with pressure), *"overwhelm"* (original literal sense is: to swamp or submerge completely), *"pedigree"* (original literal sense: foot of a crane). These examples are of historical metaphors that died out at both the conceptual level as well as the language usage level.

2.1.4 Metaphor and Other Figures of Speech

This thesis distinguishes between metaphor and other figures of speech or tropes that in some sense express mappings such as similes, metonymies and idioms. The next subsections highlight the main similarities and differences.

2.1.4.1 *Metaphor and Simile*

A simile is a figure of speech that is used to compare two concepts by using comparative words such as *"as, like, than, resembles, etc"* to link the *tenor* and the *vehicle*. According to the comparison view, metaphor can be seen as a *condensed simile* since both of them present an underlying analogy or similarity between two domains. However, a simile is considered more explicit than a metaphor. Thus, removing the comparative words (similarity tools) from a simile will result in a *nominal metaphor* (*Type-I*). For example, consider reshaping Example 2.10 into a simile as follows:

(2.27) My lawyer is like a shark.

2.1.4.2 *Metaphor and Metonymy*

Metonymy is a figure of speech in which the lexicalisation of a closely associated attribute of a concept is used to refer to the concept itself (Crystal, 2008). Both metaphor and metonymy express a mapping but, unlike metaphor, the latter involves only one conceptual domain. The conceptual mapping in metonymy is done within the same domain (Gibbs, 1999). Many studies discussed the relation between metaphor and metonymy in detail highlighting the fact that metonymy is based on contiguity (association) while metaphor is based on similarity (Gibbs, 1999; Nerlich, 2009; Shutova, 2011). Examples of metonymy include:

(2.28) The *White House* announces ...

(2.29) The *pen* is mighter than the sword.

2.1.4.3 *Metaphor and Idiom*

An idiom is a phrase or an expression consisting of a group of words that conveys a figurative meaning different from their literal one. This meaning cannot be guessed from the meanings of the individual words, thus an idiom is considered an inseparable lexical unit. Examples include:

(2.30) Kick the bucket.

(2.31) Get off your high horse.

(2.32) Swept off my feet.

(2.33) Spilled the beans.

(2.34) Digging your own grave.

On the other hand, as discussed earlier, a metaphor is an implied analogy where a concept (represented by a word sense) is borrowed to represent another concept by exploiting single or common properties between both concepts (Lakoff and Johnson, 1980). Unlike idioms, the meaning of a metaphor can be determined by understanding its individual lexical units even if the listener did not encounter it before (Crystal, 2008).

Commonly used metaphors which became conventionalised in the language found their way into lexical resources (dictionaries) under the idioms category. Although I argue against this generalisation from a linguistic point of view, it is understandable to assign conventionalised metaphors (fixed expression) to an already existing label rather than creating a new one. Specially that many idioms have metaphorical roots (Gibbs, 1999) and some idiomatic expressions can be understood through conventional metaphoric mappings between two domains such as Example 2.33 that represents *"Secret is Food"* or Example 2.34 that represents *"Problems/Failure is Death"*.

2.2 TECHNICAL CONCEPTS AND BACKGROUND

2.2.1 Dataset Preparation

2.2.1.1 *Inter-Annotator Agreement*

Inter-annotator agreement (IAA) is used to assess the reliability of dataset annotation. Previous works have employed various forms of Kappa (Cohen, 1960) in order to assess the quality of the annotated metaphor datasets, as will be discussed in Section 3.2.2. These measures include: Cronbach's alpha (Cronbach, 1951), Fleiss' Kappa (Fleiss, 1971), Siegel and Castellan's Kappa (Siegel and Castellan, 1988), and Krippendorff's alpha (Krippendorff, 2004).

In this thesis, Fleiss' kappa is used to measure the IAA which is calculated as follows:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2.1}$$

where $\bar{P}$ is the mean proportion of agreement between k annotators and $\bar{P}_e$ is the mean proportion of agreement by chance. Typically, the scores are interpreted using the Landis and Koch (1977) scale where scores of over 0.6 are considered substantial. Table 2.1 revisits the IAA scores interpretation according to this scale.

Table 2.1: Inter-annotator agreement scores and their corresponding interpretation according to Landis and Koch (1977) scale.

Kappa Score	Strength of Agreement
<0.0	poor
0.0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.81	Substantial
0.81–1.0	Almost Perfect

2.2.1.2 *Annotators and Crowdsourcing*

Annotating datasets could be done by expert annotators, usually an in-house team of annotators with a relevant background, or through crowdsourcing platforms. These platforms such as Amazon Mechanical Turk[5] (MTurk) or Crowdflower[6] (rebranded as Figure-Eight then Apen) offer services to annotated datasets by recruiting a large number of human annotators (non-expert laypersons). The choice of either utilising such services or an in-house team of experts for an annotation experiment depends on several factors such as the time, budget and the task difficulty. In this thesis, the goal is to employ the ease-of-use of such platforms and at the same time hire expert annotators. MTurk allows such option unless the annotators register as a "Worker" (i.e. annotator) on the platform. Then, in the annotation experiment there is a custom qualification option that allows restricting the participation to specific "Workers" based on their *Worker ID*. As will be discussed in Chapter 4, a team of expert annotators, with similar background, were hired to work on the annotation experiments in the context of this thesis. The team were asked to register on MTurk to complete the annotation task and to receive their payments. Training were given to the annotators to familiarise them with the platform and to measure their understanding of the task. Further details will be given for each experiment in Section 4.2 and Section 4.4.

2.2.2 Linguistics and Word Representation

2.2.2.1 *Grammar Relations and Dependencies*

This thesis is concerned with linguistic metaphors of the predicate and attributive types. Therefore, it focuses on identifying expressions of verb-noun and adjective-noun grammatical relations where the verb or the adjective can be used metaphorically. These grammatical relations can be obtained using a dependency parser, such as the Stanford dependency parser (Chen and Manning, 2014). The verb-noun relation could be a *dobj* or a *nsubj* dependency and the adjective-noun relation is a *amod* dependency. The expressions "grammatical relations" or "dependency rellations" are used interchangeably in this thesis.

5 https://www.mturk.com
6 https://appen.com/

2.2.2.2 *Context (In-)Dependent Embeddings*

In this thesis, distributed word representations will be utilised, namely embeddings, which are used to encode semantic information about a given text into a dense vector using neural network architectures. These embeddings could be context-independent or context-dependent. Various pre-trained embeddings will be employed form different models such as Word2Vec, GloVe, ELMo and BERT. This section gives a brief overview of each of them. Context-independent word embeddings are distributed representations generated from large text corpora. Word2Vec (Mikolov et al., 2013a,b) is one of the well-known pre-trained embedding vectors that are created using context-predicting neural methods, namely the Continuous Bag-of-Words (CBOW) model and the Skip-Gram model. Global Vectors for Word Representation (GloVe) is a widely used word embeddings algorithm introduced by Pennington et al. (2014). One limitation of context-independent approaches to prepare word embeddings is that the cannot model polysemous words properly. The same vector representation will be assigned to a word with multiple senses. In order to overcome this issue and to utilise context, language models have been introduced to create word representations. The representations out of the pre-trained language models are called context-dependent embeddings and they encapsulate contextual, syntactic and semantic information of the language. Modelling deep contextualised word representations from language models was introduced by Peters et al. (2018). The representations from these models, referred to as Embeddings from Language Models (ELMo), are obtained from deep bidirectional language models that are based on LSTM networks to learn context-dependent aspects of word meanings along with syntactic aspects. This results in better representations of a word depending on its context. Devlin et al. (2018) introduced Bidirectional Encoder Representations from Transformers (BERT) by taking the idea of ELMo further by using the recent attention transformer architecture (Vaswani et al., 2017) to encode context. ELMo and BERT pre-trained embeddings can be employed directly in an NLP task as additional features or they can be fine-tuned to better suit the target task.

2.2.3 Recurrent Neural Networks

Recurrent neural networks (RNNs) are often used with sequential data such as text (a sequence of words). Traditional RNNs, also referred to as vanilla RNNs, read an input sequence and output hidden states at each time step. The hidden states encode information about the input from its beginning to its end (i.e. forward). Reading a given sequence backwards could help in encapsulating extra information. The hidden states from the forward and backward processes could be combined to form what is called bi-directional RNNs. One limitation of RNNs is when dealing with long sequence. Long Short Term Memory Networks (LSTMs) are proposed by Hochreiter and Schmidhuber (1997) to solve this problem, which was denoted as the short-term memory problem of RNNs, using a mechanism called gating to control the flow of information.

2.2.4 Evaluation Strategies for Metaphor Identification

This thesis, following the majority of previous works in this area, considers the task of metaphor identification as a classification task. Therefore, its evaluation employs the tradition metrics for

text classification which are accuracy, precision, recall and F1-score. The following equations describes each metric as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.5}$$

$$\tag{2.6}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively.

2.3 SUMMARY AND METAPHOR IN THIS THESIS

This chapter provided an overview of the technical concepts and background related to this thesis. The first part of this chapter presented the linguistic definition of metaphor highlighting its main components followed by an overview of the prominent views and theories developed to understand its comprehension. The difference between the comparison view and the interaction view was highlighted. While the former focuses on a preexisting linguistic relationship between the metaphor components, the latter focuses on the interaction between them allowing for new relations as well. Since not all views are based on the similarities between the underlying mapped domains, the anomaly view exploited the dissimilarities to view metaphor as a violation of semantic norms. Stemming from the interaction between language and cognition, the CMT of Lakoff and Johnson (1980) came to look at metaphor from a cognitive view. It emphasised that semantically-related linguistic metaphors can be grouped under certain conceptual headings by exploiting the mapping between two conceptual domains. In this work, I follow the CMT view of metaphor. However, I focus only on analysing linguistic metaphors not the conceptual ones governing them. I also agree with the interaction view in that metaphor can convey new and imaginary relations between its underlying conceptual domains.

Furthermore, this chapter discussed the difference between novel, conventional and dead metaphors. Due to their frequent usage, conventional metaphors do not call to attention the language user and they can be produced, used and understood effortlessly. On the other hand, novel ones call attention to themselves through the unusual mapping between the employed conceptual domains. I consider both novel and conventionalised metaphors to be equally important to human cognition. Therefore, I focus on both of them in this thesis. Dead metaphors are out of the scope of my study since they have lost their literal basic meaning and are only used currently in their metaphorical sense. Moreover, I also distinguish between metaphor and other figures of speech such as simile, metonymy and idiom.

This chapter, also, highlighted the different types of linguistic metaphors based on either their syntactic structure or the implicit or explicit presence of the metaphor components. This thesis focuses on lexical metaphors as well as multi-word metaphoric expressions considering predicate

and attributive metaphors. This thesis therefore focuses on metaphoric expressions of *vehicle-tenor* pair where the *vehicle* can be a verb or an adjective. Examples of such linguistic metaphors include *"bright mind"*, *"clear truth"*, *"blind love"*, *"hazy relationship"*, *"gloomy argument"*, *"pure love"*, *"absorb disapointment"*, *"attack cancer"*, *"beat the illness"*, *"contain your anger"*, *"stir excitement"*, *"twist the facts"*, and *"own my ambition"*.

Finally, the chapter highlighted the definition of some concepts and technical terms that are used in the context of this thesis. The presented technical background is related to dataset annotation, word embeddings and artificial neural networks as well as evaluation strategies for metaphor processing.

3 LITERATURE REVIEW

"... the distinction between past, present and future is only a stubbornly persistent illusion."

(Einstein, 1955)

The last few decades have witnessed a growing interest in metaphor processing in text by introducing a wealth of ideas to model its computational recognition and comprehension. Many approaches and techniques have been introduced to explore the automatic processing of different types of metaphors as well as the preparation of metaphor-related resources and the design of annotation schemes. These efforts have been increased and enriched by the introduction of a series of metaphor processing-related venues such as the Workshop on Metaphor in NLP (Shutova et al., 2013a; Klebanov et al., 2014b; Shutova et al., 2015; Klebanov et al., 2016b), the Workshop on Figurative Language Processing (Klebanov et al., 2018b, 2020) and the Shared Task on Metaphor Detection (Leong et al., 2018, 2020). A lot of effort has been done to summarise and review the literature covering various aspects of metaphor processing (Zhou et al., 2007; Shutova, 2015; Veale et al., 2016; Rai and Chakraverty, 2020).

This chapter will focus mainly on reviewing the previous research pertaining to processing linguistic metaphors in text. The chapter starts with explaining the different tasks of processing metaphors followed by a detailed explanation of the main tasks of interest in this thesis, namely metaphor identification and interpretation. This chapter gives a detailed overview of the state-of-the-art approaches and techniques which support 1) the automatic identification and interpretation of linguistic metaphors, 2) the development of metaphor annotation schemes and the preparation of resources and datasets for both tasks. A chronological perspective is taken here to review the literature focusing on the various adopted paradigms to process metaphors on the sentence, relation, and word levels in order to investigate how these paradigms affected the choice of approaches, developed architectures and selected features.

3.1 AUTOMATIC METAPHOR PROCESSING

Due to their nebulous nature, metaphors are quite challenging to comprehend and process by humans, let alone computational models. This intrigued many researchers to develop various automatic techniques to process metaphors in text. The computational processing of metaphors is concerned with designing and implementing computational models that can automatically recognise, understand and model metaphors in text with an acceptable level of precision by language users. Metaphor processing comprises several tasks including identification, interpretation and cross-domain mappings. These tasks can be defined as follows:

Metaphor identification: It is concerned with recognising the metaphoric word or expression in the input text. This task is the most studied among metaphor processing tasks in part due to the availability of datasets as will be discussed in Subsection 3.2.2.

Metaphor interpretation: This task focuses on discerning the meaning of the metaphor by inferring the *ground* and explaining the relation between the *tenor* and the *vehicle*.

Cross-domain mappings: This task, also referred to as source-target mappings, focuses on identifying the relation between the source and target domain concepts in a way that mimics the human formulation of metaphors. This mapping is produced by studying a set of multiple metaphorical expressions that describe one concept in terms of another. This task is important to support the creation of knowledge-bases of metaphoric language; however it is beyond the scope of this work.

These variation in tasks concerned with processing metaphor steered the creation of designated corpora for each task. Moreover, the levels of metaphor analysis informed the design of both the annotated corpora and the computational models of metaphor processing. The next sections will explain metaphor identification and interpretation in detail.

3.2 METAPHOR IDENTIFICATION

Metaphor identification is an essential step which supports other metaphor processing tasks as well as more specific NLP applications. We need first to recognise the metaphor in text in order to interpret it or discern its underlying source-target relation or even translate it or analyse its sentiment. Choosing the level of processing depends on the end application that one has in mind when designing and developing a computational model to identify metaphors. Moreover, the level of analysis determines choosing the annotated dataset for evaluation and comparison. This section highlights the difference between each level of processing then discusses how these levels affect the annotation of datasets for metaphor. After that, an overview of the various approaches pertaining to metaphor identification will be given.

3.2.1 Levels of Metaphor Identification

Identifying metaphors in text can be done on either the sentence, grammatical relation or word levels. These levels of analysis (paradigms) are already established in the literature and adopted by previous research in this area as will be explained in Subsections 3.2.2 and 3.2.3. These levels of analysis can be summarised as follows:

Sentence level: Approaches adopting this level classify the whole sentence that contains the metaphoric word/expression as either metaphoric or literal without the explicit annotation of the metaphoric expression.

Relation level: This level of metaphor identification focuses on certain grammatical relations by looking at pairs of words where both the source and target domain words are classified as a metaphoric expression. It is also referred to as phrase-level metaphor identification due to the way a sentence is divided into sub-phrases with various syntactic structures (we use these two terms interchangeably in the context of this thesis). The most commonly studied grammatical relations are verb-noun and adjective-noun relations where the metaphoricity

of the verb or the adjective (source domain/vehicle) is discerned given its association with the noun (target domain/tenor).

Word level: This level is also referred to as token-level metaphor identification and it means looking at each word in a sentence and deciding whether it is used metaphorically or not given the context. Approaches that adopt this paradigm treat the task as either sequence labelling or single-word classification. In both methods, only the source domain words (*vehicle*) are labelled as either metaphoric or literal. Many approaches are designed to identify metaphors of different syntactic types on the word level but the most frequently studied ones are verbs.

The following paragraphs will explain the differences between each level of processing, highlighting the main strengths and limitations of each level.

3.2.1.1 *The Broader View of Sentence-Level Analysis*

Identifying metaphor on the sentence level can be considered the broader type of metaphoric analysis. It classifies the whole sentence as either metaphoric or not, provided that it contains a metaphor. It does not differentiate between the type of metaphor or if the sentence contains more than one metaphoric expression. This type of analysis does not provide any information on where the metaphor is in the text, hence it could be used in applications that focus on the figurative nature of the sentence as a whole. An example of such applications is analysing sentiment in figurative tweets (Ghosh et al., 2015a). In this application, the main focus was not the metaphor itself but the figurative nature of the whole sentence which was initially collected based on lexical patterns that indicate metaphoricity such as the words *"figuratively"* and *"literally"*. One limitation of this paradigm is that it is hard to deal with text that comes without predefined sentence boundaries such as spoken discourse or user-generated social media text. For example, in the aforementioned application, the whole tweet was classified as metaphoric or not regardless of its length. Subsections 3.2.2 and 3.2.3 present an overview of the previous works that adopted this paradigm for dataset annotation as well as the modelling of computational approaches to metaphoric sentences classification, respectively.

3.2.1.2 *Word-Level vs. Relation-Level Metaphor Analysis*

The other two well-established paradigms in the literature are the relational and word-level metaphor annotation and identification. Although the main focus of both is discerning the metaphoricity of the *vehicle* (source domain words), relation-level approaches take the *tenor* (target domain words) associated with the *vehicle* under study into account while processing the metaphor which, in turn, gives the model a narrower focus in a way that mimics the human comprehension of metaphors. Thus, processing metaphors on the word level could be seen as a more general approach where the *tenor* of the metaphor is not explicitly highlighted as well as the relation between the source and the target domains. On the other hand, relation-level metaphor identification explicitly analyses the *tenor* and the source-target relation.

Figure 3.1 illustrates the difference between the relation and word levels of metaphor identification. As will be extensively explained later in this thesis, the word-level annotation could be done in a similar way to either sequence-labelling or single-word classification. Approaches that adopt the former way annotate each word in a given sentence as either metaphoric or not, the most prominent work that adopted this method is Steen et al. (2010). The latter approach focuses

on annotating a specific word-class (e.g. verbs) in a given sentence for metaphoricity. Birke and Sarkar (2006) were the first to introduce this approach in order to create an annotated dataset to identify the non-literal usages of particular verbs. The annotation scheme in both approaches formulates the definition of a metaphor based on either the human annotator's intuition of what a metaphor is (Birke and Sarkar, 2006) or by consulting a dictionary to check the basic meaning of a specific word (Steen et al., 2010). In the given example in Figure 3.1, the basic meaning of *"on"*, which should be the more concrete, body-related, more precise (often historically older) meaning, is related to physical (concrete) surfaces or objects. Therefore, this word will be annotated as metaphoric by giving it the label "1". When adopting the word-level annotation, only the targeted word (the *vehicle*) is annotated without any explicit highlighting of the argument (the *tenor*) that influenced its metaphoricity. As depicted, highlighting the syntactic relation is one step of the relation-level analysis, in which the *tenor* is highlighted. One way to do this is using dependency parsing. Then, the targeted candidate is classified based on the proposed approach that considers the contextual interaction between the *tenor* and *vehicle*. In this example, the whole expression *"on the weekend"* is labelled as metaphoric, where *"on"* is the *vehicle* that exhibits a metaphoric sense in this sentence due to the abstract *tenor "weekend"*.

Example: *The diligent reporter said that the new rules sparked a heated debate in the media on the weekend.*

Word-level Metaphor Identification:

implicit relations and no identification of the tenor (target domain words)

*The diligent reporter said that the new rules **sparked** a **heated** debate **in** the*

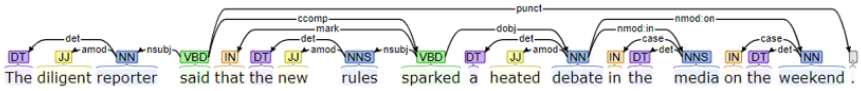
0 0 0 0 0 0 0 0 1 0 1 0 1 0

*media **on** the weekend.*

0 1 0 0

Relation-level (phrase-level) Metaphor Identification:

explicit relations and explicit identification of the tenor (target domain words)



grammar relation	expression	
amod	**diligent** reporter	0
nsubj	reporter **said**	0
nsubj	rules **sparked**	1
amod	**heated** debate	1
dobj	**sparked** debate	1
case	**in** media	1
case	**on** weekend	1

Figure 3.1: An illustration of the difference between word-level and relation-level metaphor identification. Stanford CoreNLP (Manning et al., 2014) is used to generate the dependencies. Binary labels 0/1 means literal/metaphoric.

Stowe and Palmer (2018) highlighted the importance of integrating syntax and semantics to process metaphors in text. Through a corpus-based analysis focusing on verb metaphors, the authors showed that the type of syntactic construction (dependency/grammar relation) a verb occurs in influences its metaphoricity. Relation-level metaphor processing requires an extra step to identify the grammatical relations (i.e. dependencies) that highlight both the *tenor* and the *vehicle*. Thus, it might be seen that processing metaphors on the word level is more straightforward and raises the question: why do we need relation-level metaphor identification?

The interaction between the metaphor components is less explicit in word-level analysis either when treating the task as sequence labelling or single-word classification, therefore there will be reasons that some downstream tasks would prefer to have such explicitly marked relations. The main distinction between the relation-level and the word-level paradigms is that the former makes the context more explicit than the latter through providing information about not only where the metaphor is in the sentence but also how its components come together through hinting at the relation between the *tenor* and the *vehicle* (i.e. the *ground*). It is a deeper level of analysis that captures information that is not captured on the word level, and this information is helpful to support other tasks e.g. metaphor interpretation and cross-domain mappings.

Although having various paradigms to analyse metaphor allowed researchers to look at its processing from different perspectives, one limitation of such diversity is that it will be difficult to fairly compare the proposed models across the different paradigms (Shutova, 2015). Direct mapping from word-level to relation-level annotation is not straightforward and requires extra annotation effort. Consider the following examples that contain verb metaphors:

(3.1) `The speech stirred the emotions.`

(3.2) `"History will judge you at this moment."`

(3.3) `Citizens see hope in the new regulations.`

Identifying metaphoric verbs on the word level will result in recognising the verbs *"stirred"*, *"judge"* and *"see"* as metaphoric in Examples 3.1, 3.2 and 3.3, respectively. In Example 3.1, both the subject and the object are responsible for the metaphoricity of the verb; while in Example 3.2, the subject gave the verb its metaphoricity and in Example 3.3 the object did. This is done implicitly in word-level annotation/identification. On the other hand, if we consider relation-level processing, the *tenor* associated with the verb has to be explicitly highlighted. Thus, annotating the above examples on the relation level focusing on verb-direct object relations (i.e. dobj) will result in identifying the expressions *"stirred the emotions"* and *"see hope"* as metaphoric in Examples 3.1 and 3.3, respectively and ignoring Example 3.2 since *"history will judge"* is a subject-verb (i.e. *nsubj*) relation. Therefore, adapting existing datasets annotated on the word level is required to arrive at explicit analysis of the *tenor* and the relation between the source and the target domains.

3.2.1.3 *Where does this thesis stand?*

I believe that understanding the interaction between the *tenor* and the *vehicle* is essential for metaphor identification, comprehension and interpretation. Therefore, I focus on processing metaphors on the relation level in order to model their comprehension in a way that mimics the human formulation of metaphors. The motivation behind choosing the relational paradigm can be summarised by a quote from End (1986) as follows:

> *"Metaphors are meaningful when the relationship between the topic and the vehicle is discovered and understood. It is necessary to determine the ground in order to understand a metaphor. Determination of the ground goes beyond understanding the meaning of two words, topic and vehicle. It involves an understanding of a more abstract relationship between them. The processes involved in the formation of the relationship between topic and vehicle are the focus of most psycholinguistic research on metaphor."*

This thesis adopts the principled relational-level paradigm by taking the *tenor* into consideration while annotating or identifying the metaphoric expressions in text. The proposed work in this

thesis will focus on identifying metaphoric expressions of adjective-noun and verb-noun grammar relations. On the computational level, this thesis hypothesizes that looking at metaphor explicitly at the level of a specific grammatical relation, by having an explicit representation of this relation, should allow the model to capture the implicit analogy between the metaphor components and draw its attention to the contextual interaction between the *tenor* and the *vehicle*. As highlighted in Section 2.3, I follow the conceptual metaphor theory (CMT) by analysing linguistic metaphors of specific grammatical relations which will allow the realisation of particular source-target domain relationships.

3.2.2 Metaphor Annotation and Corpora

In this section, the most relevant research in terms of the dataset preparation and the annotation of linguistic metaphors for metaphor identification will be discussed. As mentioned earlier, there are several factors that affect the creation of datasets annotated for metaphor and their annotation scheme. Among these factors are the level of metaphor analysis and the type of metaphor, in addition to the targeted application. These factors can push the dataset creation towards a specific domain or text type.

Past research in this area has either focused on formal well-structured text such as news or only targeted selected examples of metaphors. The majority of researchers formulate their own annotation guidelines based on the adopted theory of metaphor and its definition. One of the main challenges of this work is to introduce an annotation scheme that results in an expert annotated dataset for metaphor identification that have large coverage of metaphoric usages and text genres while maintaining high accuracy (more details in Section 4.2). The following subsections present an overview of the existing state-of-the-art datasets annotated to support metaphor identification on different levels of processing. Table 3.2 provides a detailed summary of the employed datasets in the literature to identify linguistic metaphors in English text highlighting the level of annotation for each dataset among other properties.

3.2.2.1 *Word-Level Datasets*

Word-level datasets[1] are prepared by focusing on annotating text for metaphoricity either by annotating a single word of a certain word-class in the sentence or by annotating every word in the sentence in a way similar to sequence labelling.

TroFi Example Base (Birke and Sarkar, 2006, 2007) is one of the earliest datasets annotated to identify metaphor on the word level. It is also referred to as the TroFi dataset. The dataset consists of 3,737 manually annotated English sentences extracted from the 1987-1989 Wall Street Journal corpus (WSJ) covering the literal and metaphoric senses of 50 selected verbs. The metaphoricity of the selected verbs on the word level is identified by manual annotation. The inter-annotator agreement (IAA) was calculated on a random sample of 200 annotated sentences scoring 0.77 in terms of Cohen's kappa (Cohen, 1960) among two annotators. This dataset has been frequently used to evaluate the performance of different approaches that identify verb metaphors either on the word level or on the relation level by focusing on subject-verb-object grammar relations.

Gedigian et al. (2006) also employed the WSJ corpus to prepare their dataset by utilising the one labelled with PropBank (Kingsbury and Palmer, 2002) frames, word senses and semantic

1 sometimes referred to as token-level datasets in the literature.

roles. The authors selected particular verbs related to the MOTION and CURE frames from FrameNet (Fillmore et al., 2003). Then, they selected the example sentences of which the PropBank senses are corresponding to the FrameNet selected verbs. The examples which comprise around 4,000 sentences are then annotated for metaphoricity. It is worth mentioning that 92% of the instances were metaphorical resulting in an unbalanced dataset. There is no information available regarding the IAA or the employed annotation scheme. Moreover, the dataset is not publicly available which limited its usage only to the approach introduced by Gedigian et al..

Around the time of the first two datasets, the Pragglejaz Group (2007) introduced a metaphor identification procedure (MIP) for human annotators. Its main aim was to annotate each word in a given text for metaphoricity. Steen et al. (2010) extended and employed this procedure as MIPVU to create the VU Amsterdam Metaphor Corpus (VUAMC). The VUAMC has become one of the most popular existing metaphor datasets nowadays. It is the largest corpus annotated for metaphors and has been used extensively to train, evaluate and compare models that identify metaphors on the word level. The corpus consists of 117 randomly selected texts from the BNC Baby version, a subset of the British National Corpus (BNC) (Burnard, 2007), which comprises various text genres, namely academic, conversation, fiction and news. Their collaborative annotation scheme annotates metaphors on the word level, regardless of the word's syntactic type, considering a word as a metaphor as long as its most basic meaning, derived from corpus-based dictionaries, contradicts its current contextual meaning. The basic meaning is typically the most physical or concrete meaning which does not have to be the first sense listed under a word entry. The MIPVU employs two other dictionaries in addition to the corpus-based dictionary. The IAA was measured in terms of Fleiss' kappa (Fleiss, 1971) among four annotators which averaged 0.84. One of the issues with this procedure is that the sense of every word in the text is considered as a potential metaphor, even idioms or fixed collocations, which are considered inseparable lexical units. Moreover, the annotators have to go through a series of complex decisions starting from chunking the given text into lexical units, then discerning their basic meaning, and finally the metaphoric classification. The uniformity of the basic meaning interpretation may vary from one annotator to the other. The corpus is published[2] in an XML format; Figure 3.2 shows an example of the corpus where the metaphoric words are tagged as *function="mrw"*.

The availability of the VUAMC encouraged many researchers to employ it while developing their computational approaches for metaphor identification on the word level. Furthermore, the NAACL 2018 Metaphor Shared Task (Leong et al., 2018) employed the VUAMC in order to introduce a metaphor detection shared task. Many researchers employed the corpus as part of the shared task to develop, train and test systems to identify metaphors on the word level; the top performing systems will be discussed in Subsection 3.2.3. The shared task consisted of two tracks, which are 1) *All Part-Of-Speech (POS)* to identify nouns, verbs, adverbs and adjectives that are labelled as metaphorical; 2) *Verbs* track which is concerned only with identifying metaphorical verbs. All forms of the verbs *"be, do,* and *have"* are excluded for both tracks. The corpus is then divided into training and test sets according to the focus of each track. A script is provided to parse the original VUAMC.xml file which contains the corpus, since the corpus is not directly downloadable due to licensing restrictions. Table 3.1 shows the statistics of the dataset as highlighted in (Leong et al., 2018).

2 The VUAMC was available online at: `http://ota.ahds.ac.uk/headers/2541.xml` but the website was unresponsive at the time of this publication.

```
<s n="441">
  <w lemma="such" type="DT0">Such </w>
  <w lemma="language" type="NN1">language </w>
  <w lemma="focus" type="VVD-VVN">
    <seg function="mrw" subtype="PP" type="met" vici:morph="n">focused</seg>
  </w>
  <w lemma="attention" type="NN1">attention </w>
  <w lemma="on" type="PRP">
    <seg function="mrw" type="met" vici:morph="n">on</seg>
  </w>
  <w lemma="the" type="AT0">the </w>
  <w lemma="individual" type="NN2">individuals </w>
  <w lemma="or" type="CJC">or </w>
  <w lemma="group" type="NN2">groups </w>
  <w lemma="who" type="PNQ">who </w>
  <w lemma="be" type="VBD">were </w>
  <c type="PUQ">‘</c>
  <w lemma="break" type="VVG">
    <seg function="mrw" type="met" vici:morph="n">breaking</seg>
  </w>
  <w lemma="the" type="AT0">the </w>
  <w lemma="law" type="NN1">law</w>
  <c type="PUQ">’</c>
  <c type="PUN">, </c>
  <c type="PUQ">‘</c>
  <w lemma="commit" type="VVG">
    <seg function="mrw" type="met" vici:morph="n">committing</seg>
  </w>
  <w lemma="criminal" type="AJ0-NN1">criminal </w>
  <w lemma="act" type="NN2">acts</w>
  ...
```

Figure 3.2: An example form the VU Amsterdam metaphor corpus (VUAMC) showing the data annotation format and the metaphoric words labelled with the metaphor-related word tag *(function="mrw")*.

Table 3.1: Statistics of the training and test data in the "Verbs" track in the NAACL metaphor shared task. %M is the percentage of metaphors as reported in (Leong et al., 2018).

Data	Training			Test		
	#texts	#tokens	%M	#texts	#tokens	%M
Academic	12	4,903	31 %	4	1,259	51%
Conversation	18	4,181	15%	6	2,001	15%
Fiction	11	4,647	25%	3	1,385	20%
News	49	3,509	42 %	14	1,228	46%

The main limitation of the VUAMC, and any dataset that stems from it, is that it only suits the identification of metaphors on the word level. Thus, it is not possible to apply the VUAMC in its current state to relation-level metaphor identification and there are no larger datasets designated to support relation-level metaphor identification.

Shutova and Teufel (2010) adopted the MIP annotation scheme, with some modifications, to annotate linguistic metaphors on the word level focusing on verbs in a subset of the BNC. The corpus comprises 761 sentences and 13,642 words. The authors exclude specific verb classes including: auxiliary verbs, modal verbs, aspectual verbs, and light verbs from the annotation arguing that these verbs exhibit weak metaphoric potential. The IAA was evaluated by means of Siegel and Castellan's Kappa (Siegel and Castellan, 1988) which averaged 0.64 among three native annotators. The authors reported that the conventionality of some metaphors is a source of disagreement. The authors extended the dataset annotation to include other part-of-speech tags. Although the dataset is not publicly available, it can be obtained from the authors.

Another work that exploits known corpora to prepare a metaphor annotated dataset is Hovy et al. (2013). The authors created their dataset by extracting sentences from the Brown corpus (Francis and Kucera, 1979) to identify metaphors of any syntactic structure on the word level. They used a list of 329 predefined metaphors as seed to extract sentences that contain the specified expressions. The dataset is manually annotated using crowd-sourcing through the Amazon Mechanical Turk (MTurk) platform. The annotators were asked whether a highlighted word in a sentence was used metaphorically or not based on its original meaning. The IAA among seven annotators was 0.57. The annotated instances were filtered out yielding a final corpus consisting of 3,872 sentences, out of which 1,749 contains metaphors. Although the dataset is quite sizable and balanced, it is not publicly available.

In a series of works, Klebanov and Flor (2013); Klebanov et al. (2014a, 2018a) took a different perspective to prepare metaphor annotated corpus, instead of utilising known corpora or news text, by focusing on non-native written English text. The dataset, referred to as the Essays dataset, comprises essays written for a large-scale college-level assessment of analytical writing. The essays were annotated for argumentation-relevant metaphors relying on the intuition of the human annotators to define metaphor. Klebanov and Flor (2013) proposed an annotation scheme that focuses on annotating metaphors relevant to the writer's arguments and based on the annotator's intuition of what a metaphor is. As a result, 116 essays were annotated by two annotators, with a background in linguistics, obtaining an IAA of 0.575 in terms of κ. The main topic of the essays is to discuss the following statement: *"High-speed electronic communications media, such as electronic mail and television, tend to prevent meaningful and thoughtful communication"*. The annotators were also asked to interpret the labelled metaphoric words and identify their argumentative contribution. Klebanov et al. (2014a) extended this dataset by introducing 174 essays under the previous topic about communication as well as a second topic to discuss the following statement:*"In the age of television, reading books is not as important as it once was. People can learn as much by watching television as they can by reading books."*. The dataset was divided in two sets according to each topic in which Set A comprises 85 essays and Set B comprises 79 essays. Two annotators were asked to annotate the essays obtaining an IAA of 0.58 and 0.56 for Set A and Set B, respectively, in terms of κ. In 2018, Klebanov et al. (2018a) sampled a dataset from the publicly available ETS Corpus of Non-Native Written English [3]. The dataset was annotated following the annotation scheme of Klebanov and Flor (2013). This dataset, namely the ETS Corpus of Non-Native Written English, was employed as part of the ACL 2020 Metaphor Detection Shared Task (Leong et al., 2020) to develop computational models that identify metaphors on the word level; the best performing systems will be discussed in Subsection 3.2.3. The dataset consists of 240 essays annotated on the word level. Figure 3.3 shows a snippet of the dataset where annotated metaphors are labelled with the prefix *M_*. This dataset is not available for distribution, but access can be requested for research purposes under a Data License Agreement.

In order to detect metaphors in conversational text on social media, Jang et al. (2014) collected around 300 sentences from three web discussion forums including a Massive Open Online Course, a breast cancer support group, and a forum for street gang members. Two annotators whose background is linguistics were asked to annotated each word in the given sentences for metaphoricity by adopting an annotation scheme similar to MIP. The IAA was measured in terms of Cohen's kappa and scored 0.49 at word level. As a continuation of this effort, Jang et al. (2015a) acquired a corpus of 1,562,459 posts from an online breast cancer support group. A set of seven

3 https://catalog.ldc.upenn.edu/LDC2014T06

```
As people M_climb M_the M_ladder of success their ideas tend to change from M_dynamic
and innovative to M_static and conservative .
I believe that succesful poeple M_focus and doing what they already know how to do
rather than M_exploring or trying out new things and taking risks .
M_Reaching a M_level of success whether in bussiness or in life M_requires time and
hard work , and upon M_reaching success risk would be to huge of a M_price .
...
```

Figure 3.3: A snippet from the ETS Corpus of Non-Native Written English (Klebanov et al., 2018a). Annotated metaphors are marked with a M_ prefix.

predefined words which are: *"boat, candle, light, ride, road, spice,* and *train"*, were used to retrieve all the posts in the corpus that contain them. These words can appear either metaphorically or literally in the corpus. An IAA of 0.81 was recorded in terms of Fleiss' kappa among five annotators on the MTurk platform who were provided with a Wikipedia definition of metaphor. The final dataset consists of 2,670 sentences and is not publicly available.

In an effort to create a multilingual dataset annotated for metaphor, Mohler et al. (2016) introduced the Language Computer Corporation (LCC) metaphor datasets. The English dataset was extracted from the ClueWeb09 corpus[4]. The freely available part of the dataset contains around 7,500 metaphoric pairs of any syntactic structure annotated by adopting an annotation procedure similar to the MIP. There is no clear information regarding the number of annotators or the final IAA of this subset. The dataset was annotated on multiple levels which are metaphoricity, cross-domain mappings and affect. The metaphoricity ratings were on a four-point scale to distinguish weak, conventional and clear metaphors from literal senses. Although the dataset was annotated on the word level targeting verbs, nouns, adjectives and adverbs as well as multi-word expressions (MWEs), the source and target domain words are highlighted. This allowed the authors to ask the annotator to link each source/target domain word to a corresponding list of source and target domains (covering 114 source domains and 32 target domains). Furthermore, the annotators were asked to rate the affect (emotional impact) of each annotated metaphor on a scale from -3 to 3. The IAA is done on a subset of the whole English dataset by two annotators reporting a value of 0.928 but there is no information regarding the type of metric. It is worth noting that the authors stated that this is just a "snap-shot" of the annotation quality and not the final IAA score of the released datasets.

A potential work to annotate metaphors in lexical resources is Mohammad et al. (2016) where different senses of verbs in WordNet (Fellbaum, 1998) were annotated for metaphoricity. Verbs were selected if they have more than three senses and less than ten senses yielding a total of 440 verbs. Then the example sentences from WordNet for each verb were extracted and annotated by 10 annotators using crowd-sourcing through the CrowdFlower platform[5]. The verbs that were tagged by at least 70% of the annotators as metaphorical or literal were selected to create the final dataset. The dataset consists of 1,639 annotated sentences out of which 410 were metaphorical and 1,229 literal. The dataset is quite sizable (yet unbalanced) for supervised machine learning approaches but might not suit neural approaches that require larger datasets for training. Furthermore, the average sentence length is relatively short (11 words) with

4 https://lemurproject.org/clueweb09/
5 by the time of writing this thesis the CrowdFlower platform was re-branded as Figure Eight

some sentences containing three words providing no reliable context to discern metaphoricity. However, this dataset, referred to as the MOH dataset, became a benchmark to evaluate metaphor identification approaches and has been widely used to model and evaluate systems identifying metaphoric verbs on the word level.

3.2.2.2 *Relation-Level Datasets*

These datasets were created to suit the relational paradigm of metaphor processing which focuses on discerning the metaphoricity of expressions of certain grammatical relations. Obtaining these relations could be done either automatically by employing a dependency parser or manually by highlighting targeted expressions in a specific corpus. These expressions are then manually annotated for metaphoricity given the surrounding context. As will be discussed in the coming paragraphs, almost all the proposed datasets under this category focused on *Type-III* metaphors by annotating expressions of adjective-noun grammar relations.

The first attempt to create a metaphor dataset on the relation level was by Turney et al. (2011). The authors created a dataset of 100 sentences from the Corpus of Contemporary American English (COCA) (Davies, 2009) to identify the metaphoricity of adjective-noun grammar relations (*Type-III* metaphors). The dataset focuses on five selected adjectives which are: *"dark, deep, hard, sweet,* and *warm"*, forming twenty adjective-noun pairs which were manually annotated by five annotators whose background was in psychology. The IAA was reported in terms of Cronbach's alpha (Cronbach, 1951) and scored 0.95. This dataset has a small size and is limited to the aforementioned adjectives.

Neuman et al. (2013) introduced two datasets annotated on the relation level to identify *Types-I-III* metaphors. The datasets are obtained from two large corpora, namely the Reuters corpus (Lewis et al., 2004) and the New York Times (NYT) corpus (Sandhaus, 2008). The authors focused on the government and governance domains by analysing five target nouns (concepts), which are: *"Governance, Government, God, Father,* and *Mother"*. The final annotated dataset consist of 1,378 expressions from the Reuters corpus and 1,003 expressions from the NYT corpus. The IAA was measured in terms of Cronbach's alpha among four annotators. The scores obtained for *Type-I, II,* and *III* were 0.78, 0.80, and 0.82, respectively, for the Reuters dataset and 0.80, 0.76, and 0.72, respectively, for the NYT dataset. Although Neuman et al. tried to increase the coverage of the annotated metaphor types, the dataset is still limited to the aforementioned domain concepts.

Tsvetkov et al. (2014) also was interested in analysing *Type-III* metaphors, and therefore created a relation-level metaphor dataset that focuses on adjective-noun grammatical relations. The dataset comprises around 2,000 adjective-noun pairs which were selected manually from collections of metaphors on the Web. It is divided into 1,768 pairs as a train set and 200 pairs as a test set. Only the test set contains the full sentences which were obtained from the English TenTen Web corpus (Jakubíček et al., 2013) by utilising SketchEngine[6] (Kilgarriff et al., 2014). The annotation scheme depended on the intuition of the human annotators to define the metaphoric expressions. An IAA of 0.76, in terms of Fleiss' kappa, was obtained among five annotators on the test set. The main limitation of this dataset is the absence of the full sentences in the training set which forces the models employing it to either ignore the context that surrounds the adjective-noun pairs or to use the small test set in a cross-validation experimental setting which makes the model prone to overfitting. This dataset is widely used by approaches targeting the

6 http://www.sketchengine.eu

identification of adjective-noun metaphoric expressions and thus became a benchmark, referred to as the TSV dataset, for evaluation and comparisons.

Another attempt that focuses on creating resources for *Type-III* metaphors is by Gutiérrez et al. (2017) who annotated a dataset of adjective-noun pairs focusing on 23 adjectives in particular. The dataset comprises 8,592 adjective-noun pairs out of which 4,601 are metaphorical. The adopted annotation scheme was the modified MIP scheme by Shutova and Teufel (2010) to identify metaphors on the relation level. The dataset was annotated by two native speakers obtaining an IAA of 0.80 in terms of Cohen's kappa. The authors filtered out all expressions that require wider context to establish their metaphoricity such as *"bright side* and *weak point"*. Therefore, the context (full sentences) around each adjective-noun expression was not provided.

3.2.2.3 *Adapted Word-Level Datasets*

Annotated datasets that support word-level metaphor identification are not suitable to support relation-level processing due to the annotation difference. Furthermore, the majority of relation-level metaphor datasets focused on adjective-noun metaphoric expressions. To overcome the limited availability of relation-level datasets, there has been a growing effort to enrich and extend benchmark datasets annotated on the word level to suit relation-level metaphor identification. Although it is non-trivial and requires extra annotation effort, Tsvetkov et al. (2014) and Shutova et al. (2016) introduced adapted versions of the TroFi and MOH datasets, respectively, to train and evaluate models that identify verb-noun metaphors on the relation level.

The first attempt to adapt existing word-level datasets to suit relation-level processing was made by Tsvetkov et al. (2014). In this work, the TroFi dataset, which was originally designed to classify particular literal and metaphoric verbs on the word level, was adapted in order to extract metaphoric expressions on the relation level. Tsvetkov et al. parsed the original dataset using the Turbo dependency parser (Martins et al., 2010) to extract subject-verb-object (SVO) grammar relations. The final adapted relation-level dataset consists of 1,609 sentences out of which 953 metaphorical and 656 literal instances. Since this is an adaptation of an already existing dataset, no further manual annotation is carried out assuming the correctness of the original annotation. Therefore, the IAA of the adapted dataset was not evaluated.

Another effort to enrich the available datasets for relation-level metaphor identification is done by Shutova et al. (2016). The authors adapted the benchmark MOH dataset, which was initially created to extract metaphoric verbs on the word level, to suit the identification of verb-noun expressions for metaphoricity. Verb-direct object and verb-subject dependencies were extracted using a dependency parser and filtered yielding a dataset of 647 verb–noun pairs, out of which 316 instances are metaphorical and 331 instances are literal. This subset, which was referred to as MOH-X in several papers, is available upon request from the authors.

3.2.2.4 *Sentence-Level Annotated Datasets*

Annotating datasets on the sentence level is done as a binary classification of whether the whole sentence is metaphoric or not, regardless of where the metaphor is. Since this coarse-grained annotation is required only for specific applications, few attempts were made to prepare datasets on this level of annotation.

Among the few attempts to annotate metaphor on the sentence level is the one proposed by Mohler et al. (2013a). The authors created a dataset focusing on linguistic metaphors in the

governance domain. The dataset consists of 500 documents which were extracted from political speeches, websites, and online newspapers. Manual annotation of the dataset was done by three annotators who annotated the sentences if some element in the text seemed to have been used figuratively. The annotation scheme is based on the annotator's intuition of what a metaphor is. The dataset consists of around 21,000 sentences out of which 1,269 are metaphoric. There is no information about the IAA. Since the dataset is created solely to evaluate the system proposed by the authors, which will be explained in Subsection 3.2.3, it is not publicly available.

Twitter datasets of a figurative nature were prepared in the context of the SemEval 2015 Task 11 on Sentiment Analysis of Figurative Language in Twitter (Ghosh et al., 2015a). The dataset is originally designed and annotated to support the classification and sentiment analysis of tweets containing irony, sarcasm, and metaphors. The available training, and test sets were collected by querying Twitter Search API using lexical patterns that indicate each phenomenon. For example, hashtags such as *"#sarcasm, #irony"* and words such as *"figuratively"* and *"literally"* are used as lexical markers[7] to collect the metaphoric tweets of ironic/sarcastic nature. The training and test dataset contains 2,000 and 800 tweets, respectively, which are categorised as metaphoric tweets. The dataset was annotated for sentiment by seven annotators using the CrowdFlower platform. Similar to the sentence-level dataset introduced by Mohler et al., this dataset was created with a specific application in mind. Therefore, it is not uniquely developed for the metaphor identification task.

3.2.2.5 *Others*

There has been a growing interest to study other aspects of metaphoricity. As discussed earlier, the LCC metaphor datasets were prepared to capture the metaphoricity strength as well as the affect of the annotated metaphors. Other researchers were also interested to study the novelty of metaphoric expressions. The following paragraphs briefly discuss these attempts.

Parde and Nielsen (2018) exploited the VUAMC to create a dataset of relation-level metaphors annotated for novelty. In this work, 18,000 metaphoric word pairs of different syntactic structures have been extracted from the corpus. Five annotators were then asked to score the highlighted metaphoric expression in a given context for novelty on a scale from 1 to 3. The annotation experiment was set up on the MTurk platform achieving an IAA of 0.435 in terms of κ.

Another work that addresses metaphor annotation for novelty is Do Dinh et al. (2018b) focusing on word-level metaphors however. Similar to Parde and Nielsen (2018), the authors exploited the VUAMC to annotate 15,180 metaphors for novelty using crowd-sourcing. Different annotation experiments were set up on the MTurk platform to decide: 1) the novelty and conventionality of a highlighted word, 2) the scale of novelty of a given metaphor, 3) scale of "unusualness" of a highlighted word given its context, and 4) the most novel and the most conventionalised metaphor from given samples. The annotators obtained an IAA of 0.39, 0.32 and 0.16 in terms of Krippendorff's alpha (Krippendorff, 2004) for the first three tasks, respectively. The authors attributed the simplicity of the annotation task description as one of the reasons behind the low IAA scores.

7 Shutova et al. (2013b) studied the reliability of such technique and discussed that the dependence on lexical markers as a signal of metaphors is not sufficient.

Table 3.2: Summary of the annotated datasets for linguistic metaphor identification employed in the literature. *The dataset is not directly available online but can be obtained by contacting the authors. (% M = percentage of metaphors; #Anno = number of annotators; IAA = Inter-annotator agreement.)

Level of analysis	Dataset	Benchmark Name	Syntactic structure	Data Source	Domain	Size	% M	Crowd-sourcing	# Anno	IAA	Available
word level	(Birke and Sarkar, 2006)	TroFi Example Base	verb	50 selected verbs (WSJ)	open	3,727 sentences	57.5%	in-house (on 200 sentences)	2	0.77	yes[8]
	(Gedigian et al., 2006)	-	verbs	selected examples (WSJ)	restricted	4,000 sentences	92%	in-house	-	-	no
	(Steen et al., 2010)	VUAMC	all POS	known-corpus (BNC)	open	∼16,000 sentences (∼200,000 words)	12.5%	in-house	4	0.84	yes[9]
	(Shutova and Teufel, 2010)	-	verb	known-corpus (BNC)	open	761 sentences	21.5%	in-house	3	0.64	yes*
	(Hovy et al., 2013)	-	any	known-corpus (The Brown Corpus)	open	3,872 sentences	45.17%	yes	7	0.57	no
	(Mohler et al., 2013a)	-	any	selected examples	restricted	21,000 sentences	6.04%	in-house	3	-	no
	(Klebanov and Flor, 2013) (Klebanov et al., 2014a)	The Essays dataset	any	non-native English text	restricted	175 essays	-	in-house	2	0.58	no
	(Jang et al., 2015a)	-	noun	selected examples (Social Media)	restricted	2,670 instances	68.01%	yes	5	0.81	no
	(Mohler et al., 2016)	LCC	any	known-corpus (ClueWeb09)	open	7,500 metaphoric pairs	-	no	-	-	partially[10]
	(Mohammad et al., 2016)	MOH	verb	selected examples (WordNet)	open	1,639 sentences	25%	yes	10	-	yes[11]
	(Klebanov et al., 2018a)	ETS corpus	any	non-native English text	restricted	240 essays	-	in-house	2	0.62	yes[12]
relation level (phrase level)	(Turney et al., 2011)	-	adj-noun	5 selected adjectives (COCA)	open	100 sentences	-	no	5	0.95	no
	(Neuman et al., 2013)	-	noun-isA-noun; subj-verb-obj; adj-noun	5 selected nouns (Reuters/NYT)	restricted	1,378/1,003 expressions	37.73/ 41.17%	in-house	4	0.80/ 0.76	no
	(Tsvetkov et al., 2014)	TSV	adj–noun	selected examples (Web)	open	∼2,000 adj-noun pairs	50%	in-house (on 200 sentences)	5	0.76	yes[13]
	(Gutiérrez et al., 2017)	-	adj–noun	selected examples	open	8,592 adj-noun pairs	53.54%	in-house	2	0.80	yes[14]
	(Tsvetkov et al., 2014)	adapted TroFi	subj-verb-obj	50 selected verbs (News)	open	1,609 sentences	59.23%	-	-	-	yes*
	(Shutova et al., 2016)	adapted MOH (MOH-X)	verb-direct obj; subj-verb	selected examples (WordNet)	open	647 sentences	48.8%	-	-	-	yes*
sentence-level	(Ghosh et al., 2015a)	The SemEval SAFL	-	selected examples (tweets)	open	2,800 tweets	-	-	-	-	yes*

[8] http://natlang.cs.sfu.ca/software/trofi.html
[9] http://ota.ahds.ac.uk/headers/2541.xml
[10] http://www.languagecomputer.com/metaphor-data.html
[11] http://saifmohammad.com/WebPages/metaphor.html
[12] https://github.com/EducationalTestingService/metaphor
[13] https://github.com/ytsvetko/metaphor
[14] http://pages.ucsd.edu/$\sim$e4gutier/m4p/AN-phrase-annotations.csv

3.2.3 Metaphor Identification Approaches

Over the last few decades, the focus of computational metaphor identification has shifted from rule-based (Fass, 1991) and knowledge-based approaches (Krishnakumaran and Zhu, 2007; Wilks et al., 2013) to statistical and machine learning approaches including supervised (Gedigian et al., 2006; Turney et al., 2011; Dunn, 2013a,b; Tsvetkov et al., 2013; Hovy et al., 2013; Mohler et al., 2013a; Klebanov et al., 2014a; Bracewell et al., 2014; Jang et al., 2015a; Gargett and Barnden, 2015; Rai et al., 2016; Bulat et al., 2017; Köper and Schulte im Walde, 2017), nearly unsupervised (Birke and Sarkar, 2006; Shutova et al., 2010) and unsupervised methods (Shutova and Sun, 2013; Heintz et al., 2013; Strzalkowski et al., 2013). These approaches employed a variety of lexical and semantic features to design their models. With the advances in neural networks, the focus started to shift towards employing more sophisticated models to identify metaphors. This section reviews previous research of linguistic metaphor identification on sentence, relation and word levels in English text.

3.2.3.1 *Rule-Based and Knowledge-Based Approaches*

One of the earliest attempts to identify metaphors in text was Fass (1991) who introduced the *met** method to distinguish metonymy from literalness, metaphor and anomaly. Fass adopted the anomaly view of metaphor; therefore, the proposed method starts by detecting violation of selectional preferences through utilising hand-coded rules to identify the non-literalness of a given phrase. Then, it distinguishes between metonymy and metaphor via another set of rules that look at the relation between pairs of word senses. Finally, the difference between metaphor and analogy is distinguished. The method was also able to provide an interpretation of an identified metaphoric expression through the mapping of concepts. In addition to the hand-coded rules, the method utilises a lexicon that comprises the sense-frames of around 500 words. The authors did not evaluate the system performance. The main limitation of this method is that it relies on a knowledge base and hand-coded patterns which limits the extensibility of this approach.

A data-driven approach to identify metaphors in lexical resources was taken in Peters and Peters (2000). The authors focused on detecting lexicalised systematic polysemy in WordNet. Semantic relations were identified between word senses which allowed the identification of metonymic and metaphorical relations. The WordNet hierarchy was searched for high-level concepts (nodes) that share the same word form (as hyponyms) among their descendants. It was found that some of these concepts reflect metaphoric and metonymic relations. There is no performance evaluation reported. Coverage might be one of the limitations of this approach, however it was an important step towards identifying metaphors in text despite the lack of annotated data.

Exploiting hyponym relations from WordNet to identify selectional preference violation was the basis for the approach by Krishnakumaran and Zhu (2007). The authors were interested to widen the scope of metaphor identification by identifying various types of metaphors focusing on *Types-I-III* metaphors (i.e. nominal, predicate and attributive metaphors). Their idea was to exploit the absence of the hyponym relation between the *tenor* and the *vehicle* to detect semantic preference violations. The proposed approach detects these hyponymy relations using WordNet to classify nominal metaphors; in addition, word bigram frequencies obtained from the Web 1T corpus (Brants and Franz, 2006) were employed to identify predicate and attributive metaphors. Although the authors utilised a dependency parser to extract the grammar relations of

subject-object, verb-noun and adjective-noun, the classification is done on the sentence level. The system was evaluated on a set of manually annotated examples from the Master Metaphor List (MML) (Lakoff et al., 1991) where conventionalised metaphors such as *"bright idea"* or *"unearthed new evidence"* are treated as literal (negative) examples reporting an accuracy of 0.58. Not dealing with literal examples and ignoring conventionalised metaphors questions the reliability of the evaluation and the generalisability of the approach.

Following Fass's idea, Wilks et al. (2013) attempted to automate the idea of acquiring selectional preferences from lexical resources to identify metaphors as a violation of these preferences. The authors also, similar to Krishnakumaran and Zhu, employed WordNet to implement their main approach in addition to a baseline that utilises VerbNet (Schuler, 2005). The main approach is based on the hypothesis that for a given word if a lower less frequent word sense in WordNet satisfies the selectional preference of its context in a given targeted expression more than its main (first) sense then it is likely to be a metaphor. The authors focused on identifying the metaphoricity of nouns and verbs on the relation level. Therefore, the Stanford dependency parser (De Marneffe et al., 2006) was employed to identify targeted grammatical relations of subject-verb, and verb-object where either the verb or the noun can be used metaphorically. Moreover, the TRIPS semantic parser and lexicon (Allen et al., 2008) were used to provide the semantic roles and selectional restrictions for a given verb by processing its definition in WordNet and identifying its nominal arguments then abstracting these arguments through their higher-level hypernyms in WordNet. In order to evaluate this approach, the authors manually annotated a set of 122 sentences from the governance domain to create a balanced dataset. The authors compared this approach to the VerbNet-based baseline to acquire the selectional preferences through hand-coded rules. The main limitation of this baseline, which the authors tried to remedy by utilising WordNet, is the limited coverage of VerbNet which forced the authors to assume that there is no selectional preference violation for the verbs that do not exist in the resource. An F-score of 0.49 is reported for identifying metaphors using the VerbNet-based baseline and of 0.67 for identifying metaphors using WordNet. This approach is based on several assumptions that lead to some limitations. The first is that it relies on the WordNet sense ranking assuming that the first sense is the frequent (literal) sense. Moreover, the authors assumed that there is only one literal sense for a given word. Finally, this approach also resulted in detecting metonymies along with metaphors, however there is no distinction made between the two figures of speech.

3.2.3.2 *Statistical and Machine Learning-Based Approaches*

WORD-LEVEL PROCESSING

Birke and Sarkar (2006, 2007) introduced *Trope Finder* (TroFi), which is a nearly unsupervised system to identify metaphorical sense of verbs through sentence clustering. The authors viewed the task as word sense disambiguation and adapted the statistical similarity-based approach by Karov and Edelman (1998) which clusters sentences based on their similarities to a predefined set of seed sentences annotated for word sense. The authors employed the same approach to classify literal and non-literal usages of verbs, however there is no distinction made between different types of non-literalness. The authors prepared the TroFi dataset as part of this work which targets the non-literalness of 50 particular verbs, as discussed in Subsection 3.2.2. The system obtained an F-score of 0.538 on 25 verbs.

Gedigian et al. (2006) was interested in the statistical modelling of metaphor through developing a supervised machine learning model to identify the metaphoricity of verbs. The authors

employed a subset of the WSJ corpus annotated by PropBank in addition to WordNet to develop their approach. The nominal arguments (subject and object) and their semantic roles associated with each targeted verb were extracted from the PropBank annotations. These arguments are then used as features to train a maximum entropy (maxEnt) classifier (Berger et al., 1996). The nominal arguments are represented as pronouns, named entities and WordNet synsets. The authors also prepared a dataset for training and testing their model which focused only on verbs with frames related to MOTION and CURE from FrameNet, as discussed in Subsection 3.2.2. The model achieved a performance of 95.12% in terms of accuracy on the test set with a slight improvement of 2.2% over the proposed naive baseline that assigns majority class to all instances. One limitation of this model is that it is trained on an unbalanced dataset (92% of the data is metaphoric) with specific lexical items. Although the model focused on identifying the metaphoricity of verbs on the word level identifying the arguments (*tenor*) was essential for the proposed identification process.

Dunn (2013a,b) developed an ontology-based approach to identify metaphors on the word level. The proposed approach exploit domain interaction to determine the concepts and their properties in text. First each lexical item in text is mapped to its WordNet synsets which then are mapped into concepts using the SUMO ontology (Niles and Pease, 2001, 2003). The high-level properties of these concepts, such as domain type (e.g. ABSTRACT, PHYSICAL, SOCIAL, and MENTAL) and event status (e.g. PROCESS, STATE, and OBJECT), are then extracted to form feature-vector representations. These features are used to train a logistic regression classifier to model metaphor. Dunn (2013a) evaluated his proposed model on a dataset prepared by annotating a subset from the COCA to either metaphoric, humorous and literal. The dataset consists of of 2,500 sentences out of which 500 were metaphoric. However, there is no information regarding the annotation scheme. The model achieved an F-score of 0.374 using 100-fold cross-validation. In (Dunn, 2013b), the model was evaluated on the VUAMC obtaining an F-score of 0.58 using 100-fold cross-validation. It is worth mentioning that Dunn re-implemented three other approaches (Li and Sporleder, 2009; Turney et al., 2011; Shutova et al., 2010), with modifications, in an attempt to compare their performance to his approach on the same dataset. This highlights the difficulty of cross-approaches evaluation due to the difference in the adopted level of analysis and dataset.

Hovy et al. (2013) revisited the idea of viewing metaphor as an anomaly (unusual semantic composition) through introducing a novel approach that utilised dependency-tree kernels (Moschitti et al., 2006). The authors focused on the compositional properties of metaphor by utilising lexical, part-of-speech, and WordNet super-sense representations of sentence trees as features to identify metaphors on the word level. These compositional features are used to train a support vector machines (SVM) classifier (Cortes and Vapnik, 1995). The model was trained and evaluated on a manually annotated subset of the Brown corpus prepared as part of this work focusing on 329 words, as explained in Subsection 3.2.2, obtaining an F-score of 0.75. This is a notable work that highlights the importance of syntactic information in identifying metaphors when adopting the word-level paradigm.

Schulder and Hovy (2014) employed corpus-based statistics with machine learning to identify metaphors on the word level. The authors proposed that a term from a specific target domain, focusing on the governance domain, with low frequency in a general text corpus would likely be used metaphorically. Term frequency inverse document frequency (TF-IDF) was employed to verify this hypothesis. The ClueWeb09 corpus was used, as the general text corpus, to calculate

the term relevance of targeted words from the governance domain. The authors also employed the term relevance as a feature to train a binary classifier treating the task as sequential labelling. They experimented with various classifiers but the best performing was a conditional random field (CRF) classifier. The authors annotated a dataset of 2,510 sentences drawn from 312 documents from the governance domain to evaluate the proposed model. It is worth mentioning that the dataset is unbalanced since 82.7% of it is metaphorical. Moreover, there is no information given about the annotation scheme. The authors reported a relatively low F-score of 0.373 for the best performing CRF classifier.

A series of works have been introduced by Klebanov et al. (2014a, 2015, 2016a) mainly to investigate a variety of features to identify metaphors on the word level. In (Klebanov et al., 2014a), the authors employed a set of features including unigrams, POS tags, concreteness scores and topic models from the NYT corpus to train a logistic regression classifier. The VUAMC[15] and the Essays dataset, introduced in Subsection 3.2.2, were employed to train and test the proposed model obtaining averaged F-scores of 0.3325 and 0.615 for each dataset, respectively, using 10-fold cross validation. This work was extended in (Klebanov et al., 2015) by re-weighting the training examples to remedy the dataset imbalance. The classifier performance was improved achieving averaged F-scores of 0.51 and 0.64 for the VUAMC and the Essays dataset, respectively. Improving the identification of verb metaphors was the main focus of (Klebanov et al., 2016a). Therefore, the authors investigated the effectiveness of orthographic and semantic features to identify metaphoric verbs on the word level. These features were studied under the notion of semantic generalisations and classifications to capture the regularities of verbs metaphoricity. This is done through the utilisation of semantic classes of verbs from WordNet and VerbNet besides the previously employed features to train a logistic regression classifier on the VUAMC focusing on verbs only. The training and test splits of the NAACL 2018 Metaphor Shared Task were employed. The model performance was evaluated using cross-validation on the training set obtaining an averaged F-score of 0.56. The model obtained an F-score of 0.60 on the test set.

Rai et al. (2016) investigated the effect of combining conceptual and affect-related features with lexico-syntactic ones to identify metaphors on the word level. The conceptual features include: concreteness, imageability, and meaningfulness. The affect-related features were extracted using WordNet-Affect (Strapparava and Valitutti, 2004) and include: cognitive state, physical state, trait, attitude and emotion. The VUAMC was utilised to train a conditional random fields (CRF) classifier. Since the VUAMC is annotated on the word level and in order to avoid parsing the data to obtain the arguments (*tenor*) related to the given word, the authors employed a context window of three words before and after the word itself to formulate a feature vector. The performance of the model was evaluated using 10-fold cross-validation achieving an F-score of 0.6093.

Özbal et al. (2016b) was interested in capturing the metaphoricity of proverbs. The main aim is to identify metaphors on the word level in the PROMETHEUS dataset (Özbal et al., 2016a), which is a proverbs dataset annotated for metaphors and comprises 1,054 English proverbs. The authors employed a variety of features including the ones employed by Klebanov et al. (2014a) in addition to imageability, standard/normalised domains and dense signals. Following Klebanov et al., a logistic regression classifier was employed. The performance evaluation was reported on the VUAMC obtaining an average F-score of 0.5035.

15 Klebanov et al. (2014a) reserved around 25% of the dataset for later test purposes and employed only around 75% of the VUAMC in this work which comprises 90 text fragments out of 117.

Various computational models were introduced by Jang et al. (2015a,b, 2016, 2017) to identify metaphors in social media text. Jang et al. (2015b) investigated the effect of situational factors on identifying the metaphoricity of specific candidate words. The authors employed a logistic regression classifier to identify the metaphoricity of these particular words (nouns) in a dataset of web discussions on cancer prepared as part of this work (see Subsection 3.2.2). Cancer-related events such as diagnosis, chemotherapy, etc were incorporated as features while training the model. Jang investigated the notion of *framing* in discourse for metaphor identification in the subsequent works. In (Jang et al., 2015a), frame contrasts were investigated by capturing lexical contrast around metaphorical frames. In this work, local and global contextual features, including semantic category, topic distribution, lexical chains, abstractness/concreteness and unigrams, were utilised to train a logistic regression classifier. The idea behind employing topic distribution is that non-literal words tend to have a considerably different topic distribution from the surrounding context. Therefore, topic distributions and lexical chains were employed to calculate semantic relatedness between the targeted word and the surrounding context words. Following this work, Jang investigated features of frame transition by capturing topic transition patterns occurring around metaphorical frames. Therefore, in Jang et al. (2016), sentence-level topic transition as well as emotion and cognition elements are utilised as features to train a support vector machines (SVM) classifier in order to identify the metaphoricity of a particular word. The last proposed model (Jang et al., 2017) employed frame facets in addition to lexico-syntactic features to train an SVM classifier as well to discern the metaphoricity of particular words in the cancer-related social media dataset.

RELATION-LEVEL PROCESSING

Shutova et al. (2010) introduced a nearly unsupervised approach based on verb and noun clustering to identify the metaphoricity of verbs in the BNC on the relation level. The approach employs a seed set of predefined metaphorical expressions of verb-noun pairs to learn implicit metaphorical mappings. This seed set comprises around 62 verb-noun pairs where the verb is used metaphorically and was obtained from the dataset by Shutova and Teufel (2010) which was discussed in Subsection 3.2.2. The authors based their approach on the idea that abstract nouns with similar features can be grouped (clustered) together in association with the same verbs (*vehicle*). Therefore, the authors introduced the notion of *clustering by association* to formulate this idea. The spectral clustering algorithm (Meila and Shi, 2001) was employed using lexico-syntactic features to formulate the clusters and linking them using the seed set. The RASP dependency parser (Briscoe et al., 2006) was used to parse the BNC and then corpus search is done focusing on verb-subject and verb-direct object relations to identify metaphoric expressions based on the associated clusters. Finally, a selectional preference strength filter was employed to filter out verbs exhibiting weak selectional preferences considering them as having a lower metaphorical potential such as the verbs *"assist, choose, neglect, remember"* and *"undo"*. The system was evaluated on a randomly sampled sentences with annotated metaphoric expressions from the corpus obtaining a precision of 0.79. In a follow-up work, Shutova and Sun (2013) investigated the use of hierarchical graph factorization clustering (Yu et al., 2005) to learn metaphorical associations through building a graph of concepts. The work of Shutova et al. inspired many researchers to adopt the relational paradigm of metaphor identification. Hence, focusing on proposing and experimenting with a variety of features related to the arguments (*tenor*) to improve the identification accuracy.

Turney et al. (2011) was the first to employ the idea of concreteness and abstractness to identify metaphoric expressions on the relation level focusing on predicate and attributive metaphors. They proposed that an expression with a concrete adjective or verb associated with an abstract noun is likely to be metaphoric (e.g. *"break my soul"*). Starting from a predefined set of examples, words are ranked automatically according to their abstractness/concreteness by utilising semantic similarity and the Medical Research Council psycholinguistic database (MRCP) (Coltheart, 1981; Wilson, 1988). These abstractness ratings are used as features to train a logistic regression model to predict whether the targeted adjective or verb is used metaphorically or not. The proposed model is evaluated on the TroFi dataset using 10-fold cross-validation settings achieving an F-score of 0.634 and accuracy of 0.734 for verbal metaphors. The authors prepared their own dataset from the COCA, refer to Subsection 3.2.2, to evaluate the performance of the method in identifying attributive metaphoric expressions obtaining an accuracy of 0.79. Turney et al. proposed, as a future work, employing imageability scores from the MRCP database in a similar fashion to improve metaphor prediction.

Neuman et al. (2013) built upon Turney et al.'s work and introduced three algorithms to identify metaphors of *Types-I-III*. The main idea is to employ selectional preference in addition to measuring the abstractness and concreteness to identify metaphoric expressions of subject-verb-object and adjective-noun grammar relation. Furthermore, the authors improved Krishnakumaran and Zhu's approach to identify nominal metaphors by comparing the semantic categories of nouns which were derived using positive pointwise mutual information (PPMI) on the COCA. The proposed approaches employ various lexical and semantic resources including WordNet, ConceptNet (Liu and Singh, 2004) and a dictionary. In addition, the Stanford dependency parser (De Marneffe et al., 2006) is employed to obtain the targeted grammar relations. The authors annotated two corpora, namely the Reuters and the NYT corpus, as explained in Subsection 3.2.2, to prepare their datasets of metaphors. However, they focused on the government and governance domains by analysing five target nouns (concepts), which are: *Governance, Government, God, Father,* and *Mother*. This might limit the generalisation of the proposed algorithms. The algorithms were evaluated on each corpus independently obtaining an average precision of 0.72 and an average recall of 0.80.

This work was extended by Gandy et al. (2013) by introducing a pipeline that processes metaphor in three levels. After identifying the linguistic metaphors, the authors utilised lexical resources and statistical clustering to assign domain mappings to each identified linguistic metaphors. This way conceptual metaphors are identified as well through identifying nominal analogies.

Tsvetkov et al. (2013) introduced a cross-lingual metaphor identification method to identify metaphors on the relation level focusing on the predicate type. A set of features including semantic WordNet categories, degree of abstractness, animateness, and named entities types was utilised to train a logistic regression classifier on English data. The model is then used to identify metaphors in Russian text. The system makes use of a dependency parser to extract subject-verb-object (SVO) grammar relations. The model was trained and evaluated on the benchmark TroFi dataset that was adapted in this work to support SVO relations obtaining an F-score of 0.78. This work was extended in (Tsvetkov et al., 2014) to identify attributive metaphors by utilising semantic supersenses as conceptual features. The proposed model employed a random forest classifier which was evaluated on a annotated dataset of adjective-noun grammar relations prepared as part of this work, namely the TSV dataset, achieving an F-score of 0.85 on the test

set. The authors applied model transfer to identify metaphoric expressions in other languages including Spanish and Farsi.

Tekiroğlu et al. (2015) built upon this work by employing the same feature set combined with sensorial features. The authors utilised a large sensorial lexicon, namely Sensicon, which comprises around 22,000 English words with the associated human senses. The sensorial features improved the model performance achieving an F-score of 0.875 on the same TSV test set.

Broadwell et al. (2013) was the first to employ imageability as a feature to identify metaphors. The authors define that a word is more imageable if *"it is possible to form a mental picture of its meaning"*. The imageability scoring from the MRCP database were employed based on the hypothesis that metaphors likely use highly imageable words than the surrounding context. In order to eliminate highly imageable words with literal sense, topic chaining and semantic clustering were utilised. The approach is evaluated on English and Spanish text obtaining an accuracy of 71% and 80%, respectively; however there is no information about the source of the dataset or its size.

Gargett and Barnden (2015) also built his work on the idea of employing sensory (perceptual) features such as imageability and affectiveness in addition to abstractness in order to identify metaphors. The authors employed concreteness, imageability and sentiment scores as features to identify nouns, verbs and prepositions for metaphoricity on the relation level. The Affective Norms for English Words (ANEW) dataset (Warriner et al., 2013) is utilised to obtain the affect scores such as valence, arousal and dominance. Similar to Turney et al.; Neuman et al.; Gandy et al., the concreteness scores are obtained from the MRCP database. These features are then used to train various classifiers including random forests (RF), gradient boosting machines (GBM), k-nearest neighbours (KNN), and support vector machines (SVM). The best performing was the random forest classifier. The VUAMC was parsed using the graph-based version of the Mate tools dependency parser (Bohnet, 2010) to extract particular grammar relations and then employed to evaluate the model performance.An F-score of 0.7813 was obtained using a 60-20-20% training, validation and test splits, respectively of the data.

Ben Shlomo and Last (2015) introduced metaphor identification by learning (MIL), which is a supervised approach to identify *Types-I-III* metaphors on the relation level. The approach is based on abstractness and semantic relatedness, by means of semantic relation and domain corpus frequency, to discern metaphoricity. In order to select the best performing model and features for each metaphor type, the authors experimented with various classifiers including logistic regression (LR), naïve bayes (NB), k-nearest neighbors (KNN), voting features intervals (VFI), Random Forest (RF), decision trees as well as a combination of them with the AdaBoost algorithm. The best performing models were RF, AdaBoost with NB and AdaBoost with VFI for *Types I, II,* and *III* metaphors, respectively. The same dataset introduced by Neuman et al. (2013) was employed to train the models and evaluate their performance. The best performing models obtained F-scores of 0.892, 0.706, and 0.429 for *Types I, II,* and *III* metaphors, respectively.

Gutiérrez et al. (2016) investigated the idea of modelling metaphors through compositional distributional semantics models (Baroni and Zamparelli, 2010) in order to learn linear transformations in a vector space to model the implicit cross-domain mappings of metaphors. The authors focused on identifying metaphors on the relation level by learning vector representations of targeted expressions of adjective-noun grammar relations. The distributional model was trained

using PPMI and corpus derived distributional statistics on a large corpus[16] of 4.58 billion tokens. The model was evaluated on an annotated dataset of adjective-noun pairs prepared as part of this work, as discussed in Subsection 3.2.2. The model achieved an F-score of 0.817 and an accuracy of 0.809 using 10-fold cross-validation.

Rai and Chakraverty (2017) introduced the idea of employing fuzzy rough sets to identify nominal metaphors of subject-isA-object grammar relation. A set of conceptual and affect-related features were employed to train a FuzzyRoughSet rule-based classifier (Riza et al., 2014). The conceptual features, including: imageability, concreteness, familiarity and meaningfulness, were extracted from the MRCP database. The ANEW dataset was employed to extract the affect-related features. The authors also employed Word2Vec embeddings (Mikolov et al., 2013b) to calculate the relatedness between the subject and the object. Rough sets are then used to perform feature selection. A dataset of 150 sentences sampled from a list of sentences under the *Stimulus* topic in (Thibodeau and Durgin, 2011) were employed to test the model performance in a train-test split of 66-34% obtaining an F-score of 0.8817. Rai et al. (2017) took this work a step forward by introducing an unsupervised approach based on Fuzzy c-means (FCM) clustering. Then, a rule-based classifier was implemented to classify nominal candidates of the same grammar relation into metaphors, literals and probably_metaphor. The approach employed the same set of conceptual features as well as calculating semantic relatedness between the subject and the object. The model obtained an accuracy of 71.6% over the previously employed dataset. Following this work, Rai et al. (2018) employed dependency graphs derived from a dependency parser, namely Stanford CoreNLP, to identify metaphors in the VUAMC and TroFi dataset focusing on subject-verb-object grammar relations. In addition to conceptual features, the authors employed edge features which are the assigned weights to the edges between a word from the context and the root verb in the dependency graph. These weights are based on the semantic relatedness between the nodes. The conceptual features of each node in addition to the edge features were used to train a linear SVM classifier. The model achieved F-scores of 0.7107 and 0.7484 for the VUAMC and the TroFi datasets, respectively.

SENTENCE–LEVEL PROCESSING

Sentence-level metaphor identification did not receive much attention compared to the relational and word levels. As explained earlier, approaches adopting this paradigm focus on classifying the whole sentence as either metaphoric or literal given that it contains a metaphoric word or expression without an explicit annotation of the metaphor. Mohler et al. (2013a) adopted the sentence-level paradigm and proposed the idea of *semantic signatures*. The hypothesis is that a semantic signature for a given text can be formulated as a set of highly related and interlinked WordNet senses. Signatures of different texts from the same domain can be compared; therefore a text with a semantic signature closely matches the signature of a known metaphor is likely to represent a metaphor as well. In order to detect semantically related concepts, the signatures of the source and target domains of a predefined set of metaphors are obtained through semantic expansion using WordNet, Wikipedia links, and corpus co-occurrence statistics, then the signature of an unknown targeted text can be compared to it. The authors experimented with a suite of binary classifiers, including a maximum entropy classifier, an unpruned decision tree classifier, support vector machines, and a random forest classifier. These classifier were trained and

16 The corpus comprises the English Wikipedia Dump of 2011, the UKWaC, the BNC and the English Gigaword corpus (Graff et al., 2003)

evaluated using a prepared dataset of around 500 documents from the governance domain, as explained in Subsection 3.2.2. The best performing model was the decision tree classifier which obtained an F-score of 0.70 on a balanced test set of 482 sentences.

3.2.3.3 *Topic Modelling-Based Approaches*

Few researchers were interested in employing topic modelling to identify metaphors in text. Heintz et al. (2013) employed Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to identify linguistic metaphors regardless of their syntactic structure but given that the source and target domains words are explicit. The authors focused on the governance domain. Topics were extracted through applying LDA on Wikipedia and then these topics were aligned with a predefined seed set of concepts. If a sentence contains words from both the source and target topics then it is tagged as metaphoric. The approach was evaluated on an annotated dataset of 600 sentences, from news websites and governance-related blogs, obtaining an F-score of 0.59. The same approach was applied on Spanish text in an attempt to generalise the idea to low-resource languages.
Similarly Strzalkowski et al. (2013) employed topic modelling to identify metaphors in the governance domain. However, Strzalkowski et al. employed topical chains (Broadwell et al., 2012) to identify sequences of concepts through linking semantically related words focusing on verbs and nouns. The approach was evaluated on an annotated dataset obtaining an accuracy of 71%; though there is no details about the dataset available such as the data source, annotation scheme and the dataset size. This approach was applied on three other languages which are Spanish, Russian, and Farsi. It worth mentioning that Strzalkowski et al. (2014) studied sentiment and affect conveyed by metaphors in a later work which is out of the scope of this thesis.

3.2.3.4 *Neural Networks-Based Approaches*

In the last couple of years and with the advances in deep learning researchers started to employ neural networks to model metaphor identification on the word and relation levels. This subsection reviews the different techniques that exploit neural models to identify linguistic metaphors in English text.

WORD-LEVEL PROCESSING

Do Dinh and Gurevych (2016) were the first to utilise a neural architecture to identify metaphors. They approached the problem as sequence labelling where a traditional fully-connected feed-forward neural network is trained using pre-trained word embeddings. The model is trained and tested on the VUAMC using splits of 76%, 12% and 12% for training, development and testing respectively. An F-score of 0.5614 were achieved by utilising the POS tags and concreteness ratings as features in addition to word embeddings. The authors highlighted the limitation of this approach when dealing with short and noisy conversational texts.

Gutiérrez et al. (2017) was encouraged by their earlier work on metaphor identification (Gutiérrez et al., 2016) to apply computational modelling of metaphor to aid the prediction of mental illness. In this work, the authors utilised the alteration in metaphor production and the distribution of sentiment scores as features to classify patients transcripts. A similar architecture to the one employed by Do Dinh and Gurevych (2016) was used to build a model for metaphoricity detection which was trained on the VUAMC and a subset of the BNC. The model was able to identify metaphors on the word level in the transcripts of patients with schizophrenia. Since

metaphor identification was not the main focus of this work no experiments were done to evaluate the performance of the model on the VUAMC.

As part of the NAACL 2018 Metaphor Shared Task (Leong et al., 2018), many researchers proposed neural models that mainly employ long short term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) with pre-trained word embeddings to identify metaphors on the word level in the VUAMC. The participating systems employed the VUAMC train and test splits provided by the shared task (see Subsection 3.2.2 for more details). The best performing systems are: THU NGN (Wu et al., 2018) obtaining F-scores of 0.651 and 0.672 for the *All POS* and *Verbs* tracks, respectively; OCOTA (Bizzoni and Ghanimifard, 2018) obtaining an F-score of 0.635 for the *Verbs* track; and bot.zen (Stemle and Onysko, 2018) obtaining an F-score of 0.617 for the *All POS* track and 0.642 for the *Verbs* track. In addition to employing word embeddings, THU NGN employed POS tags and character embeddings while OCOTA employed concreteness ratings as features.

Do Dinh et al. (2018a) employed multi-task learning to model the interplay between metaphor and idiom identification. The authors introduced two neural models utilising hard parameter sharing and Sluice networks (Ruder et al., 2017, 2019). The approach is used to improve four tasks focusing on non-literal language detection which are word-level metaphor identification, adjective-noun metaphor identification, idiomatic infinitive-verbs compounds identification and non-literal usage of particle verbs. The VUAMC and the TSV dataset were employed for the metaphor identification tasks. The models performance showed that the inclusion of the idiomatic data improved the performance of metaphor identification. The authors experimented on German text in addition to English.

Gao et al. (2018) were the first to employ the deep contextualised word representation ELMo (Peters et al., 2018), combined with pre-trained GloVe (Pennington et al., 2014) embeddings to train bidirectional LSTM-based models. The authors introduced two models to identify metaphoric verbs on the word level which are a sequence labelling model and a single-word classification model. They showed that incorporating the context-dependent representation of ELMo with context-independent word embeddings improved metaphor identification. The authors trained and tested their models on the VUAMC[17], the MOH-X dataset (adapted MOH) and the TroFi dataset. The sequence labelling model obtained an F-score of 0.726 on the VUAMC using train-test splits. It is worth mentioning that the evaluation on the VUAMC included closed-class words such as prepositions which are considerably easier to classify than open-class words such as verbs, nouns, adjectives and adverbs. This means that the proposed models were trained on a different subset of the VUAMC which make them not fairly comparable to the main task performance or other approaches utilising the same data splits of the shared task (Dankers et al., 2020). The verbs classification model obtained F-scores of 0.791 and 0.72 on the MOH-X and TroFi datasets, respectively, using 10-fold cross-validation. One issue with this approach is that the authors prepared the ELMo embeddings beforehand for the employed datasets, instead of using the pre-trained embeddings on a large corpus, which might limit its coverage when encountering a previously unseen word.

Mu et al. (2019) proposed a system that utilises a gradient boosting decision tree classifier. Mu et al. was the first to investigate the effect of document embeddings, namely doc2vec (Le and Mikolov, 2014), to exploit wider context to improve metaphor detection. Additionally, other word

17 Gao et al. (2018) employed around 75% of the dataset for experimentation, similar to Klebanov et al. (2014a) who held around 25% of the data for later testing purposes.

representations including GLoVe, ELMo and skip-thought (Kiros et al., 2015) were employed. The VUAMC was utilised to train and test the system, exploiting the same splits as the NAACL Shared Task in order to compare the model performance with the best performing model (Stemle and Onysko, 2018) in the shared task as well as the one proposed by Gao et al. (2018). The model utilising doc2vec obtained an F-score of 0.609 whereas the model utilising ELMo in addition to context features obtained an F-score of 0.668. Although the model performance was a bit lower compared to other approaches, investigating the effect of broader context on discerning the metaphoricity was essential. The authors suggested employing a more sophisticated neural architecture to improve the performance of the proposed approach.

Mao et al. (2018, 2019) revisited the idea of employing selectional preferences violation (Wilks, 1978) to identify metaphors by integrating it in a neural architecture. Mao et al. (2018) employed word embeddings and WordNet to develop an unsupervised model to identify metaphors on the word level. The authors trained word embeddings model on a Wikipedia dump to obtain general domain word representations. Then, for a given sentence, the target word is separated from its context to extract its synonyms and hypernyms from WordNet in order to construct a candidate set of all the possible senses of the target word. After that, metaphor identification is done by identifying the most likely sense from the candidate set. This is done by computing the cosine similarity between the candidate set and the context words based on a predefined threshold. Employing the candidate set of senses allowed the interpretation (paraphrase) of the targeted word. The approach was evaluated on the MOH-X dataset and a subset of the MOH dataset obtaining F-scores of 0.74 and 0.75, respectively. Although the proposed approach identifies metaphor on the word level, it was not compared with other word-level approaches but rather relation-level ones (Shutova et al., 2016; Rei et al., 2017). It is worth mentioning that Mao et al. evaluated the effectiveness of metaphor paraphrasing in the context of machine translation with the aim to improve the accuracy of machine translation systems through metaphor processing. Mao et al. (2019) introduced two neural network models inspired by the metaphor annotation procedure MIP and selectional preferences violation. The two models can be considered an adaptation of the model of Gao et al. (2018) in which GLoVe embeddings were employed in a neural network based on a bidirectional LSTM. One of the models employed multi-head attention to compare the targeted word representation with its context. Both models were evaluated on three benchmark datasets, namely VUAMC, MOH-X and TroFi. The same training and test splits of the VUAMC from the NAACL metaphor shared task were employed. The best models achieved an F-score of 0.743 for the *All POS* track and 0.708 for the *Verbs* track. For the MOH-X and TroFi datasets, F-scores of 0.80 and 0.724 were reported, respectively, using 10-fold cross-validation. In this work, the authors also introduced a recurrent neural model based on bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) as a baseline to show the effectiveness of employing such advanced context-dependent embeddings on metaphor identification. Mao's proposed approaches emphasised the importance of the context to identify metaphoricity by employing context-dependent and context-independent word embeddings.

An interesting approach was introduced by Dankers et al. (2019) to model the interplay between metaphor identification and emotion regression. The authors introduced multiple multi-task learning techniques that employ hard and soft parameter sharing methods to optimise LSTM-based and BERT-based models, namely hard parameter sharing, cross-stitch network and gated recurrent network. In this approach, metaphor was modelled on the word and sentence

levels. Therefore, the VUAMC was employed for word-level identification treating the task as sequence labelling. Then, the LCC dataset was employed to optimise the model and to predict a metaphoricity score at the sentence level. Pre-trained GLoVe and ELMo embeddings were employed by the introduced neural architectures. Pairwise joint learning was employed to randomly choose one of the two tasks at each step in the training. The LSTM-based model achieved an F-score of 0.745 while the BERT-based model achieved 0.769 on the VUAMC to identify metaphors.

Stowe et al. (2019) focused on filling the gap of training dataset availability and size for neural models that identify verb metaphors. The authors emphasised the importance of using syntactic information to improve the quality of annotated data and thus the identification models. They proposed employing VerbNet and syntactic patterns from Wikipedia to extract new training data that can be augmented with the dataset understudy. Three benchmark datasets were utilised to train and test a re-implementation of the model introduced by Gao et al. (2018) with and without the augmented information from VerbNet and the syntactic patterns. A slight improvement over the baseline was reported in terms of F-score to identify metaphoric verbs scoring 0.679 for the VUAMC, 0.683 for the MOH-X dataset and 0.694 for the TroFi dataset.

Another attempt to model broader discourse to improve metaphor identification is introduced by Dankers et al. (2020). The authors proposed two attention-based neural models which employ general and hierarchical attention mechanisms. The first model employs GLoVe and ELMo embeddings in a bidirectional LSTM-based network while the second one utilises a pre-trained BERT model. To model broader discourse, the authors employed a context window to include preceding and succeeding sentences. The representations of the discourse can be concatenated with the targeted word representation. The models were trained and evaluated on the same training and test splits of the VUAMC from the NAACL metaphor shared task in order to compare it with the best performing systems. The BERT-based model using hierarchical attention obtained an F-score of 0.715 while the LSTM-based model using general attention obtained an F-score of 0.673. The results are higher than the best reported system (Wu et al., 2018) in the shared task.

Building upon the success of the NAACL 2018 Metaphor Shared Task, the ACL 2020 Metaphor Shared Task (Leong et al., 2020) was introduced where many researchers proposed neural models to identify metaphors on the word level. In addition to utilising the VUAMC as the previous shared task, this task utilised the ETS Essays dataset (Klebanov et al., 2018a), also referred to as the TOEFL corpus. The majority of the proposed approaches employed deep bidirectional LSTM models architectures based on transformers, namely BERT. The best performing of these approaches are DeepMet (Su et al., 2020), Go Figure! (Chen et al., 2020), and illiniMet (Gong et al., 2020). The teams achieved F-scores on the VUAMC (*All POS* and *Verbs* tracks) and the ETS Essays dataset (*All POS* and *Verbs* tracks), respectively, as follows: 1) 0.769, 0.80, 0.715 and 0.749 for DeepMet; 2) 0.734, 0.775, 0.692 and 0.702 for Go Figure!; and 3) 0.73, 0.771, 0.703 and 0.719 for illiniMet. a variety of features were employed by the teams in addition to word embeddings including POS tags, concreteness ratings, WordNet semantic classes, etc.

More recently, Le et al. (2020) introduced two novel architectures to improve metaphor identification on the word level. The first architecture integrates dependency parse trees in a graph convolutional neural (GCN) model (Kipf and Welling, 2017) to directly connect the targeted words with the important context words in order to identify metaphors. The main hypothesis behind employing GCN is to explicitly focus on the relevant words in context surrounding the

targeted word to improve metaphor identification. GLoVe and ELMo embeddings were utilised in a bidirectional LSTM-based network. The second model employed multi-task learning to transfer the knowledge between word sense disambiguation (WSD) and metaphor detection. The main intuition behind that is to exploit the similarity between both tasks and transfer the knowledge from the reasoning process of the former to improve the performance of the latter. The authors evaluated the performance of their proposed models on the VUAMC in addition to the MOH-X and TroFi datasets. Furthermore, the Semcor dataset (Miller et al., 1994) is used for the WSD task. The same training and test splits of the VUAMC from the NAACL metaphor shared task were employed. The best models achieved an F-score of 0.751 on the *All POS* track data and 0.717 on the *Verbs* track data. F-scores of 0.796 and 0.732 were obtained on the MOH-X and TroFi datasets, respectively, using 10-fold cross-validation. This work emphasised the importance of explicitly highlighting the related arguments of the metaphoric word to discern its metaphoricity.

Rohanian et al. (2020) was interested in modelling the interplay between metaphor identification and the processing of idiomatic MWEs. The authors proposed an attention-based neural model that exploits GCN, similar to Le et al. (2020), to utilise the syntactic dependency information. The proposed model is built on top of a pre-trained BERT model. Since identifying verb metaphors on the word level was the main focus, the MOH-X and TroFi datasets, which are annotated for the metaphoricity of verbs, were employed. In order to identify the MWEs, a GCN-based system (Rohanian et al., 2019) was employed which was trained on the PARSEME English dataset (Ramisch et al., 2018). It was found that 24% of metaphoric verbs in the MOH-X dataset were automatically identified verbal MWEs while in the TroFi dataset the percentage was 16%. The performance evaluation of identifying metaphors using the proposed *MWEs-aware* metaphor identification model achieved an F-score of 0.8019 on the MOH-X dataset and 0.7278 on the TroFi dataset.

RELATION–LEVEL PROCESSING

An interesting approach was proposed by Shutova et al. (2016) that revisits the idea of employing imageability as a feature to improve metaphor identification. The authors adopted the relational paradigm focusing on identifying metaphoric expressions of adjective-noun and verb-noun grammar relations. This work was the first to employ multimodal embeddings of visual and linguistic features to detect metaphoricity in text. The proposed approach obtained linguistic word embeddings using a log-linear skip-gram model trained on Wikipedia text and obtained visual embeddings using a deep convolutional neural network trained on image data. This was done for both the full expression of adjective-noun and verb-noun pairs as well as the individual words constituting each expression to obtain phrase-level and word-level vector representations. Then, the cosine similarity was calculated to measure the distance between the phrase vector and the corresponding vectors of its constituent words. The authors used the TSV and the MOH-X datasets to train and test the proposed model achieving F-scores of 0.79 and 0.75, respectively, using a specified train-test splits for each dataset. Since this approach depends on a manually annotated dataset of images to obtain the visual features, it was designed to employ a small training dataset (only 80 annotated instances). One issue with this approach is the trade-off between the generalisability and coverage of the proposed model with respect to the complexity of annotating images for metaphoricity[18].

18 In an attempt to create a corpus of visual metaphors, Bolognesi et al. (2018) investigated the possibility of developing a repository of visual metaphors for research purposes, namely the VisMet corpus. The corpus contains around 350 images representing visual metaphors.

Bulat et al. (2017) investigated the idea of employing property-based semantic word representations to improve concept generalisation and thus improve metaphor identification. The authors proposed property-based (attribute-based) embeddings constructed from the human property-norm dataset (McRae et al., 2005). These embeddings were constructed through cross-modal mapping between skip-gram embedding vector space and the property-norm semantic space. The authors experimented with a traditional machine learning classifier, namely a support vector machine (SVM). The TSV dataset is used to evaluate the model performance obtaining an F-score of 0.77.

Rei et al. (2017) introduced a supervised similarity network to address metaphor identification on the relation level focusing on adjective-noun and verb-noun grammar relations. Their system utilises word gating, vector representation mapping and a weighted similarity function. Pre-trained word embeddings and attribute-based embeddings (Bulat et al., 2017) were employed as features. This work explicitly models the interaction between the metaphor components through gating, which is used to modify the vector of the verb/adjective based on the noun. The proposed model was trained and tested on the TSV and MOH-X datasets obtaining 0.811 and 0.742, respectively, in terms of F-score. The main limitation of this approach is that the context surrounding the targeted expression is ignored by feeding only the candidates as input to the neural model; this led to loosing important contextual information.

Bizzoni et al. (2017) was interested in employing a neural model to identify the metaphoricity of adjective-noun expressions. A simple neural model with a single fully-connected layer was developed by utilising the pre-trained Word2Vec embeddings. The proposed model was trained and evaluated on Gutiérrez et al.'s dataset obtaining an accuracy above 0.90% using 10-fold cross-validation. As discussed earlier, this dataset is developed to focus on identifying metaphoric expressions that do not need a context to establish their metaphoricity; therefore it lacks the full sentences around the annotated expressions. Moreover, this dataset focuses on 23 particular adjectives, which brings into question the generalisation and coverage of the trained model using it.

3.3 METAPHOR INTERPRETATION

Metaphor comprehension and understanding is a complex cognitive task that includes interpreting metaphors by grasping the interaction between the meaning of their target and source concepts. This is very challenging for humans, let alone computers. Thus, automatic metaphor interpretation is understudied in part due to the lack of publicly available datasets. This section reviews how previous works approached the task in order to interpret linguistic metaphors of different types. And how the variation in the task definition affected the development of datasets and computational models pertaining to metaphor interpretation.

3.3.1 Categorisation of the Task

As discussed earlier, the interpretation of metaphors focuses on "translating" a given metaphoric expression to its literal meaning. Rai and Chakraverty (2020) broadly formulated the problem of computational metaphor interpretation into either (a) extraction of properties transferred between

the source and target domains, or (b) identifying of the underlying conceptual mapping, or (c) generation of literal paraphrases. Since the main focus of this thesis is linguistic metaphors, I categorise the current computational approaches that address linguistic metaphor interpretation as follows:

1. *Lexical Substitution* (lexical paraphrasing) where the metaphoric word/phrase is replaced with its literal counterpart to clarify its semantic meaning. This task is viewed as single-word (lexical) substitution (Shutova, 2010; Shutova et al., 2012; Bollegala and Shutova, 2013);
2. *Paraphrase Generation* (inference of meaning) where the full sentence including the metaphoric expression is transformed using more literal words (Bizzoni and Lappin, 2018);
3. *Definition Generation* (interpretation or definition assignment) where a full interpretation (explanation) of the metaphoric expression is provided (Martin, 1990) in a way similar to dictionaries or lexicons.

Table 3.3 gives examples of the three aforementioned approaches of linguistic metaphor interpretation. The choice of the approach depends on the targeted application in mind. There are a variety of applications that can benefit from interpreting metaphors as lexical paraphrasing. The most straightforward one that comes to mind is machine translation (Toral and Way, 2018; Koglin and Cunha, 2019) and there is also text simplification (Barbu et al., 2015; Wolska and Clausen, 2017b; Bingel et al., 2018). In this thesis, I view metaphor interpretation as a definition generation task focusing on finding out the meaning of a given metaphoric expression and explain it in literal words. There are many applications that can benefit from this view such as language learning as well as lexical resources creation and development (Krek et al., 2018).

Table 3.3: Overview of the various approaches pertaining to interpreting linguistic metaphors providing examples from previous studies.

Approach	Metaphor	Interpretation
Lexical Substitution (Shutova, 2010) (Mohler et al., 2013b)	**brush aside** accusation You might get enough Republicans to join with Democrats to **push it through**	reject cause it to pass
Paraphrase Generation (Bizzoni and Lappin, 2018)	The crowd was a river in the street.	The crowd was large and impetuous in the street.
Definition Generation (Martin, 1990)	How do I **kill the process**?	to terminate computer process.

3.3.2 Metaphor Interpretation Datasets

The interpretation task is very important to fully understand the intended meaning of the metaphor, however it is much less explored compared to the identification task. One reason is that it is very exhausting for humans to comprehend the interaction between the target and the source components of the metaphoric expression. Although native speakers unconsciously grasp such interaction, asking a human annotator to translate such a cognitive process and interpret a metaphoric expression is a very demanding task. This is the reason behind the lack of publicly

available datasets for metaphor interpretation, which in turn hinders the development of this topic. There exist only few datasets for linguistic metaphor interpretation prepared by considering the task as either lexical substitution or paraphrase generation. To the best of my knowledge, there is no dataset created to interpret metaphors in the context of definition generation.

Shutova (2010) introduced a dataset of 46 sentences covering 62 metaphoric verbs in the form of subject-verb and verb-direct object grammar relations from a subset of the BNC to evaluate her proposed approach that viewed metaphor interpretation as lexical substitution. In order to annotate this dataset, five native speakers were asked to write down all suitable literal paraphrases for the highlighted metaphorical verbs. For example, the possible paraphrases given by the annotators for *"leak report"* are *"reveal, disseminate, publish, divulge, let out* and *disclose"*. There is no information available regarding the IAA as the final dataset is compiled by incorporating all of the annotations. This dataset is the only dataset available for single-word metaphor paraphrasing (lexical substitution) focusing on metaphoric verbs. Despite its limited size, it was used to evaluate other metaphor paraphrasing systems (Shutova et al., 2012; Bollegala and Shutova, 2013). The dataset is not directly available online but can be obtained upon request from the authors. Examples from the expressions in the dataset include: *"stir excitement, grasp theory, approach focuses,* and *ideology embraces"*.

Mohler et al. (2013b) prepared a balanced dataset of 463 documents from the transcripts of political speeches and online newspapers to interpret predicate metaphors. Two native English speakers were asked to provide a literal paraphrase of the verbal metaphoric instances in a previously prepared dataset by Mohler et al. (2013a) to identify metaphors in text (see Subsection 3.2.2). For example, the metaphoric expression *"take a backseat"* was paraphrased as *"be assigned lesser importance"*. The dataset comprises 232 valid interpretations of the verb metaphors in the given sentences and 231 invalid interpretations to act as negative examples which were generated randomly. There is no information about the IAA. The dataset is not publicly available.

Recently, Bizzoni and Lappin (2018) created a dataset to judge paraphrases of metaphoric sentences. Their dataset consists of 200 metaphorical sentences, each sentence has four ranked candidate paraphrases. The candidate paraphrases were labelled on a 1-4 scale based on the degree to which they paraphrase the metaphoric sentence. The dataset covers metaphors with various syntactic structures including: noun phrases, verbs, adjectives and multi-word metaphors. The metaphoric sentences were either selected from published sources or devised manually by the authors. Also, the provided candidate paraphrases were created manually by the authors themselves. Finally, all the sentences were revised by a native speaker. The dataset is publicly available online[19].

3.3.3 Metaphor Interpretation Approaches

Various factors affected the design of metaphor interpretation systems including the type of linguistic metaphor as well as the adopted theory of metaphor. Early systems focused on constructing knowledge about the conceptual domains in order to infer the figurative meaning of a given metaphor. Therefore, these systems relied on hand-coded rules to extract the properties and attributes of the source domain and then project this knowledge on the target domain. Since then other approaches emerged that saw the need to incorporate the results of metaphor interpretation systems in other applications. Therefore, they looked at the interpretation task as

19 https://github.com/yuri-bizzoni/Metaphor-Paraphrase

lexical paraphrasing and designed their systems to automatically generate literal paraphrases of linguistic metaphors focusing on predicate metaphors. Following these efforts, a series of research focused on interpreting nominal metaphors. As discussed in Chapter 2, this type of metaphor presents a straightforward relation between the source and the target domain. Therefore, previous works pertaining to nominal metaphor interpretation focused on extracting the properties of the source domain that are related to the target domain in order to infer the metaphor meaning. In this section, I review these approaches focusing on related work pertaining to interpreting linguistic metaphors in English text.

One of the earliest systems to interpret linguistic metaphors in text is introduced by Martin (1990). In this work a metaphor interpretation, denotation, and acquisition system (MIDAS) is introduced which formulated the task of metaphor interpretation in a way similar to definition generation by providing an actual interpretation and explanation of a given linguistic metaphor through finding the corresponding conceptual metaphor. Martin based his approach on the hypothesis that conventional metaphors are derived from more general ones. His system was able to interpret conventional as well as novel linguistic metaphors by exploiting the relation between source and target concepts. Given a metaphoric expression, MIDAS starts by searching for a corresponding conceptual metaphor that might explain it. For new "unseen" expressions, it searches for more general concepts based on the similarity between either the candidate itself, its hypernyms or its antonyms and previously stored conventional metaphors in the system's database. For example, the previously unseen metaphor *"kill the conversation"* is similar to the stored conventional metaphor *"kill the process"* and is explained similarly as *"terminate the conversation"*.

Narayanan (1997, 1999) introduced KARMA, a system for knowledge-based action representations for metaphor and aspect. This system was developed to support aspect and metaphoric reasoning of politics and economics event descriptions from newspapers represented by motion verbs. The proposed model provides an interpretation of metaphoric expressions based on the conceptual cross-domain mappings by projecting the properties and attributes of the source domain onto the target domain.

Kintsch (2000) also exploited the interaction between the source and target domains to represent the meaning of metaphoric expressions focusing on nominal metaphors. The proposed approach used Latent Semantic Analysis (LSA) to construct a semantic space of the source and target domains formulating the required knowledge for metaphor comprehension as vector representations. Semantic relatedness is employed, by means of cosine similarity, to obtain similar meanings of the source and target domain words.

Focusing on knowledge representation to interpret metaphor, the idea of fluid knowledge representation for metaphor interpretation and generation was introduced by Veale and Hao (2008). They focused on extracting conceptual properties of the source and target domains from WordNet and from the Web; these properties were extracted using lexico-syntactic patterns and are called *talking points*. A framework called *Slipnet* was then employed to allow for a number of insertions, deletions, and substitutions in these *talking points* in order to establish a connection between the mapped domains and interpret the given metaphor.

With the aim of incorporating the output of metaphor interpretation systems in other applications, Shutova (2010) introduced a corpus-based approach that addressed metaphor interpretation as a lexical paraphrasing task. She focused on predicate metaphors of subject-verb and verb-object grammar relations. The proposed approach substituted each metaphoric verb by its literal counter-

part (literal paraphrase/synonym). The system generates the paraphrases (substitutes) depending on a context-based ranking probabilistic model, which acquires paraphrases of metaphors from the BNC and ranks them according to their likelihood. The irrelevant paraphrases were then filtered out based on their similarity to the hypernymy relations of the metaphorical term from WordNet. Finally, the literalness of the paraphrases was verified by employing a selectional preference model (Resnik, 1993). The system performance is evaluated using the mean reciprocal rank (MRR) which scored 0.63 on a dataset created for this task from the BNC, as explained in Subsection 3.3.2. This work has been expanded in (Shutova et al., 2012; Bollegala and Shutova, 2013) by exploring unsupervised approaches. Instead of relying on WordNet to generate the initial candidates of substitutes as in the original system (Shutova, 2010), Shutova et al. (2012) employed a vector space model to generate these candidates in an unsupervised manner. The literalness of the candidates were verified using a selectional preference model as well. Their proposed model employs non-negative matrix factorization (Lee and Seung, 2000) trained on a subset of the WaCky corpus (Baroni et al., 2009). The mean average precision (MAP) is used to evaluate the system performance and scored 0.52. Following this work, Bollegala and Shutova (2013) employed the semantic relation between the *vehicle* and the *tenor* (i.e. the verb and the noun, respectively, in their case) to extract lexico-syntactic patterns to paraphrase a given expression. A set of candidate paraphrases were extracted from the Web using these patterns as queries. The candidate paraphrases were then scored and ranked by employing the notion of *semantic drift* using point-wise mutual information. A *lexical substitutability* test was finally performed to filter out the noisy candidates. The MRR is used to evaluate the system performance scoring 0.206 on the metaphor paraphrase dataset of Shutova (2010). This is a lower performance than the supervised system in Shutova (2010). The authors argued that this result is in line with the performance of other unsupervised lexical substitution approaches compared to supervised ones.

Mohler et al. (2013b) addressed the issue of giving natural language understanding applications, such as question answering, textual entailment, lexical substitution, and word-sense disambiguation, the ability *"to grasp the semantic content of metaphors"*. The proposed approach applied textual entailment to interpret metaphor. The system is built upon the *Groundhog* textual entailment system introduced in Hickl et al. (2006) which exploits various lexical and contextual features and various machine learning algorithms for entailment classification. The system was evaluated on a dataset of 232 interpretations of verb metaphors prepared as part of this work.

Ovchinnikova et al. (2014) introduced an abduction-based metaphor interpretation system. The system starts by parsing and converting the input text containing a linguistic metaphor to a logical representation. These representations in addition to a knowledge base were used as inputs to an abductive inference component to produce a source-target mapping which was then used to obtain the final interpretations. The performance of the pipeline[20] was evaluated on linguistic metaphors extracted from English and Russian datasets.

Su et al. (2017) employed semantic relatedness between the source and target domains to interpret nominal metaphors. First the properties of the source domain are extracted then transferred to the target domain through *dynamic transfer*. The relatedness between the extracted properties of the source domain and the target domain is estimated by calculating the cosine distance between their pre-trained Word2Vec embeddings. In order to extract the source domain properties, Su et al. employed a database of properties of entity concepts, namely the Property Database, as well as a taxonomy of adjectives, namely Sardonicus. These properties were

20 https://github.com/eovchinn/ADP-pipeline

expanded using their synonymys from WordNet in order to avoid data sparsity. It is worth mentioning that this work was an extension of the work introduced by Su et al. (2016) focusing on interpreting Chinese nominal and verbal metaphors based on latent semantic similarity.

Since Bizzoni and Lappin (2018) addressed the task of metaphor interpretation as full sentence paraphrase, they designed their model to focus on two sub-tasks which are the binary classification of paraphrases and paraphrase ranking. The model employs convolutional neural networks (CNN) as well as LSTM networks to encode and learn the representation of the metaphoric sentence and its candidate paraphrase. The authors created their own dataset of 200 sentences, which is explained in Subsection 3.3.2, to assess the performance of the proposed system.

An Emotion-driven Metaphor Understanding (EMU) system was proposed by Rai et al. (2019) based on the hypothesis that metaphors can have various interpretations which vary based on the receiver's perception. The authors adopted the interaction view of metaphor and emphasised that emotion-infused interpretation is better than relying solely on the source–target similarities and properties. Their unsupervised approach relied on pre-trained Word2Vec embeddings to capture the emotional properties of the source domain. Their system was able to give six interpretations of a given metaphor along the six basic emotional categories of Ekman and Friesen (1971).

3.4 WHERE IS THE GAP?

From the previous discussion on the related work to metaphor processing, the existing research gaps are identified and can be summarised as follows:

METAPHOR DATASETS CREATION AND ANNOTATION

1. Existing Metaphor Datasets
   - Many datasets are not publicly available since they were created to evaluate specific approaches and none of them address metaphor identification in the user-generated text of tweets either on the word or relation levels.
   - The majority of available datasets lack coverage of metaphors and text genres as they rely on predefined examples of metaphors from a specific domain during the creation process.
   - A common issue of all the available datasets is that they are specifically designed for a certain task definition focusing on a certain level of metaphor analysis which makes their annotation scheme difficult to generalise.
2. Level of Metaphor Analysis
   - The majority of large datasets for metaphor identification are designed to support word-level metaphor analysis. On the other hand, some datasets that are designed to support relation-level metaphor identification ignore the context of the metaphoric expression.
   - The available datasets are specifically designed to evaluate certain approaches focusing on a certain level of metaphor analysis this entails a difficulty of performing cross-systems comparisons and thus having a conclusive performance interpretation.

  - Direct mapping from word-level to relation-level annotation is not straight forward and requires extra annotation effort

3. Metaphor Interpretation Datasets

  - The task definition and categorisation affected the preparation of the datasets for metaphor interpretation. Therefore, only two datasets are publicly available that support metaphor interpretation, one prepared for lexical substitution and the other for paraphrase generation. These available datasets have important limitations in terms of size, representativeness and quality.

METAPHOR IDENTIFICATION

- Previous research did not study the effectiveness of the minimally supervised approaches to accurately identify metaphors in short texts.
- Different architectures and experimental settings have been developed to utilise various features to identify metaphors either on the word or relational levels. None of the existing approaches assessed the effectiveness and scalability of these features to identify metaphors under a unified architecture.
- The majority of existing approaches pertaining to metaphor identification adopt the word-level paradigm without explicitly modelling the interaction between the metaphor components. On the other hand, while existing relation-level approaches implicitly model this interaction, they ignore the context where the metaphor occurs.

METAPHOR INTERPRETATION

- The variation in the task definition affected the development computational models pertaining to metaphor interpretation. The majority of previous approaches have treated the task as lexical paraphrasing and none of them addressed the task as definition generation to find out the meaning of a given metaphoric expression and explain it in literal words.

3.5 SUMMARY AND CONCLUSION

This chapter traced the development of metaphor processing in the past few decades highlighting the change in concerns regarding the adopted processing paradigms and theories of metaphor and how this affected the choice of the employed approaches and the selection of features. An overview of the two metaphor processing tasks of interest in this thesis is given, namely metaphor identification and interpretation. The chapter began by discussing metaphor identification highlighting the adopted paradigms of analysis, developed datasets and the proposed approaches under each paradigm. After that, an overview of metaphor interpretation is presented focusing on the task categorisation in the current literature followed by the available datasets and developed approaches.

Metaphor identification is considered the first step towards metaphor processing. This chapter highlighted the difference between the various paradigms adopted by previous research to analyse metaphor in text including annotating and identifying metaphor on the sentence, relation or word

levels. The level of processing is among the factors that need to be determined before designing a computational approach to metaphor identification; another factor is the type of metaphor. These choices will also entail choosing the dataset (or the annotation scheme in case of preparing a new one) for evaluation and comparison purposes.

The chapter then provided extensive details about existing datasets for metaphor identification in English text. The annotation scheme, type of metaphor and level of processing were discussed for each dataset. While the majority of available datasets employed news text or known corpora such as the BNC, Klebanov et al. (2018a) introduced a dataset of non-native English writings annotated for metaphors on the word level and made it available for research usage under a non-disclosure agreement. As shown in Table 3.2, the majority of researchers carried out their annotations by leveraging their in-house team of annotators, with various backgrounds and expertise, and only few employed crowd-sourcing platforms. Additionally, many researchers developed their own annotation procedures depending on the type of metaphor and the level of processing. Further, the majority relied on the annotator's intuition to define metaphor. These variations in metaphor corpora design considerations pose a limitation on cross-systems comparisons and the possibility of a unified performance evaluation and interpretation. Another issue that adds to this limitation is that many datasets are not publicly available since they were created to evaluate specific approaches. However, there are few publicly available datasets that are currently employed by researchers focusing on metaphor identification on the word or relational levels. These widely used benchmark datasets are the TroFi (Birke and Sarkar, 2006) dataset, the VUAMC (Steen et al., 2010) and the MOH (Mohammad et al., 2016) dataset for word-level metaphor identification, whereas the TSV (Tsvetkov et al., 2014) and the adaptation of MOH by Shutova et al. (2016) (MOH-X) datasets are utilised for relation-level metaphor identification.

The majority of earlier approaches pertaining to metaphor identification adopted the selectional preference violation view of metaphor. They utilised hand-coded rules and knowledge bases to build their systems which limited their coverage and generalisation. Many approaches in the early 2000s employed lexical resources such as WordNet to identify metaphors in an attempt to remedy the lack of annotated data. In order to avoid the limitations of these early attempts, researchers started to shift their focus towards statistical and machine learning approaches. Hence, putting effort to prepare and annotate datasets for metaphor to train and evaluate the proposed approaches. Looking at metaphor identification as a classification task through the lens of machine learning allowed the researchers to focus more on the utilised features to discern the metaphoricity of a given sentence, expression or word. Of course, this required spending more effort on features selection and engineering. Conceptual (also referred to as psycholinguistic) features such as abstractness, concreteness, imageability and affectiveness were introduced by researchers adopting the relational paradigm to identify metaphoric expressions of certain grammar relations. Since these features are related to the *tenor*, they were rarely employed by approaches adopting the word-level paradigm which focused on employing lexical features of the word to be classified.

Recently, researchers started to exploit the advancement in deep learning to develop neural models in order to identify metaphors on either the word level or the relation level. Few approaches employed feed-forward neural networks while the majority opt for recurrent neural networks such as LSTM which seemed a more appropriate choice to deal with sequential data. The main benefit of using neural techniques is their ability to learn salient features automatically from the raw data. This helped in reducing the efforts of features engineering which were made

when employing traditional machine learning techniques. Context-independent pre-trained word embeddings such as GLoVe and Word2Vec were among the commonly used features by these advanced approaches. Many approaches utilised context-dependent representations such as ELMo and BERT pre-trained embeddings. Moreover, various attention mechanisms have been explored to improve metaphor identification (specially on the word level). However, these advanced models require large datasets for training and hyper-parameter optimisation. This forced the majority of approaches to process metaphors on the word level, since the large (yet unbalanced) available dataset annotated for metaphor is the VUAMC and it is annotated on the word level. These approaches employ the sequential nature of the text to capture the metaphoricity of a given word by treating the task as either sequential labelling or single-word classification. And thus, implicitly model the information from previous words (local context). On the other hand, relation-level approaches employ the contextual-interaction between the *tenor* and *vehicle* explicitly through techniques such as similarity projection or gating. While the majority of word-level approaches typically rely on the lexico-syntactic features of the local context, recent approaches started to explicitly model the broader context around the targeted word which improved metaphor identification on the word level. A recent trend of research is to model the interplay between metaphor identification and other NLP tasks such as emotion analysis (Dankers et al., 2019), word sense disambiguation (Le et al., 2020), idioms identification (Do Dinh et al., 2018a) and the processing of MWEs (Rohanian et al., 2020). The majority of them emphasised the importance of employing syntactic dependencies to improve the performance of the proposed models.

One common tradition between all approaches regardless of the employed paradigm is that they conducted their performance evaluation in terms of precision, recall, and F-score (some of them reported accuracy). This is because all approaches viewed the task broadly as a classification task assessing whether a given word, expression or sentence is metaphoric or not. However, since not all of them employ the same unified dataset sometimes it is hard to fairly compare the results of different approaches. Therefore, the tradition of previous work in this area is to compare approaches addressing the task on the same level against each other on level-specific annotated benchmark datasets.

The second part of the chapter focused on metaphor interpretation highlighting how previous approaches categorised the task as either lexical substitution, paraphrase generation or definition generation. This task categorisation affected the preparation of the datasets. Only two datasets were prepared in the context of metaphor interpretation. This lack of annotated datasets hindered the progress in this area. Additionally, it affected the adopted evaluation strategies with the majority of approaches employed human judgements of system output as a way to evaluate interpretation systems.

Earlier approaches pertaining to metaphor interpretation relied on hand-coded rules to extract the properties of the *vehicle* and then project this knowledge on the *tenor*. The output of these systems was difficult to be incorporated in other applications. Therefore, a line of research emerged that treats the interpretation task as lexical paraphrasing. These systems focused mainly on generating literal paraphrases of linguistic metaphors automatically; the majority of approaches addressed predicate metaphors (i.e. metaphoric verbs). The employed features are based on acquiring domain properties from knowledge bases or large corpora. WordNet was the most utilised lexical resource by the majority of metaphor interpretation approaches that treat the task as lexical substitution.

4 TWEET DATASETS FOR METAPHOR PROCESSING

"Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets."

(Steve Lohr, 2014, The New York Times)

This chapter discusses the preparation of gold standard datasets of tweets for metaphor processing focusing on the main tasks of interest in this thesis, namely metaphor identification and interpretation. I am interested, in the context of this thesis, in identifying and thus annotating linguistic metaphors in English text on the relation level focusing on verb-noun and adjective-noun grammar relations. Furthermore, I am interested in obtaining their interpretation by formulating the task as definition generation. The chapter therefore covers the first research theme of this thesis, which is *Resource Preparation*, where I seek to answer research questions RQ1.a, RQ1.b and RQ1.c. that are presented in Chapter 1 and are formulated as follows:

RQ1.a Can metaphor be defined in such a way that results in a high inter-annotator agreement?

RQ1.b How to adapt existing benchmark datasets to better suit relation-level metaphor identification using minimal annotation effort while maintaining annotation accuracy and consistency?

RQ1.c How can lexical resources be employed to prepare a dataset of reliable definitions (interpretations) of metaphoric expressions?

The chapter first introduces the main challenges that face research into metaphor processing in terms of corpora preparation. Then, Section 4.2 discusses my approach to address these challenges by providing details on the proposed annotation methodology to create a high-quality dataset of tweets annotated for linguistic metaphors. Additionally, the evaluation of the newly created metaphor dataset of tweets and its linguistic analysis will be provided. Section 4.3 revisits the gap in the existing metaphor datasets and provides the details of the adaptation of existing benchmark datasets to better suit relation-level metaphor identification. Finally, Section 4.4 explains the work done in this thesis to create a gold standard dataset for metaphor interpretation by treating the task as definition generation. Details will be given on how lexical resources are utilised to facilitate the annotation process.

4.1 INTRODUCTION

Among the main challenges of the computational modelling of metaphors is their pervasiveness in language, which means they do not only occur frequently in our everyday language but they are also often conventionalised to such an extent that they exhibit no defined patterns. Therefore, designing an annotation scheme to annotate metaphors in text has been deemed difficult due to the nebulous nature of metaphoricity. Furthermore, achieving consistent annotations with higher inter-annotator agreement has been difficult and as such previous work has introduced restrictions, such as limiting the study to only a few chosen words of a certain syntactic type (Birke and Sarkar, 2006; Turney et al., 2011; Jang et al., 2015a) or particular predefined metaphors (Hovy et al., 2013; Tsvetkov et al., 2014). As discussed in Chapter 3, there have been a number of datasets created for metaphor identification, however, the majority have focused on formal well-structured text selected from a specific corpus. A common issue of all the available datasets is that they are specifically designed to evaluate certain approaches focusing on a certain level of metaphor analysis. This entailed several limitations such as 1) the difficulty of generalising the adopted annotation scheme; 2) the variations of the adopted metaphor definition; 3) the difficulty of performing cross-systems comparisons and thus having a conclusive performance interpretation. Furthermore, as explained in Table 3.2, only few datasets are publicly available and none of them address metaphor identification in the user-generated text of tweets either on the word or relation levels.

Twitter presents a challenging type of social media content due to the unstructured text, vagueness of topics and brevity of the tweets. Metaphor processing in tweets can be very useful in many social media analysis applications including political discourse analysis (Charteris-Black, 2011) and health communication analysis (Semino et al., 2018). However, to the best of my knowledge, there has been no attempt to identify metaphors in tweets either on the word level or the relation level in part due to the lack of tweet datasets annotated for linguistic metaphor. Although Ghosh et al. (2015a) introduced a dataset of ironic, sarcastic and metaphoric tweets annotated on the sentence level, the main focus of their work was not the metaphor itself but the figurative nature of the whole tweet. As discussed in the previous chapter, the dataset was prepared in the context of the SemEval 2015 Task 11 on Sentiment Analysis of Figurative Language in Twitter therefore the main aim of the study was the sentiment analysis of figurative tweets not metaphor processing.

This thesis addresses the aforementioned limitations of existing datasets and the lack of metaphor datasets of tweets by proposing an annotation methodology to create a high-quality dataset of tweets annotated for linguistic metaphor focusing on verb-noun grammar relations. The proposed approach is driven by achieving balance, sense coverage and verbs representativeness as well as high annotation accuracy. I validate this methodology by creating a set of 2,500 manually annotated tweets in English, for which substantial inter-annotator agreement scores are achieved among five annotators, who are native speakers of English. This methodology is based on the use of a minimally supervised system for metaphor identification, which is introduced as part of this thesis and will be explained in Chapter 5. This system is used to assist in the identification and the selection of the examples for annotation, in a way that reduces the cognitive load for annotators and enables quick and accurate annotation. I selected a corpus of both general language tweets and political tweets relating to Brexit and I compare the resulting corpus on these two domains. As a result of this work, the first dataset of tweets annotated for linguistic metaphors is published,

which I believe will be invaluable for the development and evaluation of approaches for metaphor identification in tweets.

Another further issue that I address here is the availability, quality and size of large benchmark datasets for relation-level metaphor processing in English text. As discussed in the previous chapter, the majority of large datasets for metaphor identification are designed to support word-level metaphor analysis. On the other hand, some datasets that are designed to support relation-level metaphor identification ignore the context of the metaphoric expression. The conversion from word-level to relation-level annotation is non-trivial. In this work, I take a step towards filling this research gap by introducing an adapted version of benchmark datasets to better suit relation-level metaphor identification. Two benchmark datasets, namely the VUAMC and the TroFi dataset, were adapted to identify verb-noun metaphoric expressions. Moreover, the benchmark TSV dataset, which is originally designed to identify metaphoric expressions on the relation level, is extended by providing context for the adjective-noun relations in its training set. The adapted version of these datasets are published according to the licensing type of each of them to facilitate research on metaphor processing.

Finally, I investigate the issue of the lack of publicly available datasets for metaphor interpretation. The creation and manual annotation of such datasets is a demanding task which requires huge cognitive effort and time. Moreover, there will always be a question of accuracy and consistency of the annotated data due to the subjective nature of the problem. This work addresses these issues by presenting an annotation scheme to interpret verb-noun metaphoric expressions in English text by treating the task as definition generation. The proposed approach is designed with the goal of reducing the workload on annotators and maintaining consistency. The proposed methodology employs an automatic retrieval approach which utilises external lexical resources, word embeddings and semantic similarity to generate possible interpretations of identified metaphors. This enables quick and accurate annotation. I validate the proposed approach by interpreting around 1,500 metaphors in tweets for definition generation which were annotated by six native English speakers. As a result of this work, the first gold standard dataset for metaphor interpretation is published, which will facilitate research in this area.

The following contributions are made in this chapter, which cover the first research theme of this thesis and seek answers to **RQ1.a**, **RQ1.b** and **RQ1.c**, as follows:

- Proposing an annotation scheme that results in an expert annotated dataset of tweets for metaphor identification on the relation level with the aim of achieving large coverage of metaphoric usages and text genres while maintaining high annotation accuracy.

- Adapting existing word-level benchmark datasets for relation-level metaphor identification by employing a semi-automatic approach to avoid the need for extensive manual annotation and to facilitate future research in relation-level metaphor processing.

- Introducing an annotation scheme for metaphor interpretation by casting the task as definition generation and employing this scheme to create a dataset of metaphor interpretation focusing on verb-noun expressions.

Publications: Parts of this chapter have been published in Zayed et al. (2018, 2019, 2020a,c)

4.2 TWEETS DATASET FOR METAPHOR IDENTIFICATION

Metaphor is one of the most important elements of human communication, especially in informal settings such as social media. The widespread nature of Twitter communication has led to a growing interest to analyse the usage of the figurative language in such a context. As discussed in Chapter 3, there are different factors that affect the creation of datasets annotated for linguistic metaphors and their annotation scheme. Among these factors are the level of metaphor analysis and the type of metaphor, in addition to the adopted view of metaphor and the targeted application. The majority of previous approaches pertaining to metaphor identification have focused on formal well-structured text selected from a specific corpus to create datasets to model and evaluate their approaches. A common issue of all the available datasets is that they are specifically designed for a certain task definition focusing on a certain level of metaphor analysis which makes their annotation scheme difficult to generalise. Additionally, the majority of available datasets lack coverage of metaphors and text genres as they rely on predefined examples of metaphors from a specific domain during the creation process. This section discusses how I addressed these issues and seek to answer **RQ1.a**. My main aim is to create a dataset of tweets annotated for metaphors on the relation level that offers comprehensive coverage of metaphoric usages as well as text genre. In order to achieve this, an annotation methodology needs to be designed that guarantees high inter-annotator agreement at a large scale. Accordingly, the resulting dataset can be used for the development and evaluation of metaphor processing approaches in tweets. The following subsections discuss the proposed annotation approach using MTurk to create this dataset which is designed to ensure the dataset balance, coverage as well as high accuracy. I employ a minimally supervised metaphor identification classifier that I designed to facilitate quick and accurate annotations as well as maintaining consistency among the annotators. Subsection 4.2.1 will outline the data preparation procedure including the data sources as well as a brief overview of the employed identification system (for detailed discussion of this see Section 5.2). The annotation methodology along with the evaluation of the resulting dataset will be discussed in Subsection 4.2.2. Finally, the dataset analysis will be given in Subsection 4.2.3.

4.2.1 Data Preparation

My goal is to prepare a high-quality dataset of tweets annotated for linguistic metaphors with the aim of ensuring balance, coverage, and representativeness. These factors (Evans, 2007) are central to building a corpus so it is important to considered them when creating a metaphor dataset. Further, the other factors discussed earlier such as the metaphor type and level of analysis should be considered as well. This section discusses the data sources and the preparation steps for creating the dataset which comprises two phases: an initial annotation scheme and an improved one utilising a minimally supervised classifier to prepare the data and MTurk to host the experiment.

4.2.1.1 *Sources*

The availability of Twitter data has motivated many researchers to create datasets of tweets for various social media mining applications. This is usually done by querying Twitter APIs to collect

data over a specific period of time. The main interest of this thesis is to process metaphors in tweets which therefore requires collecting and then annotating them for metaphors. I have two options to collect and use data from Twitter. The first one is to query Twitter and collect a new dataset of tweets. A second (and quicker) option is to utilise already existing datasets of tweets that have been developed earlier as part of other NLP applications. I decided to choose the latter option and explore the feasibility of using already existing datasets of tweets as my data source.

In the first phase of this study, I employed the figurative dataset of tweets which was introduced in Ghosh et al. (2015a). This dataset is referred to as the SemEval 2015 SAFL dataset in this thesis. As discussed in Subsection 3.2.2, this dataset was initially created by querying Twitter using lexical markers such as the words *"figuratively, virtually"* and *"literally"* to collect metaphoric tweets. The tweets were collected over a period of one month, June 2014. I employed this dataset to develop an initial annotation scheme for metaphor identification as will be discussed in Subsection 4.2.1.2.

In the second phase of this study, my main goals are 1) to avoid the limitations of the SemEval 2015 SAFL dataset including targeting specific topic genres or domains (e.g. ironic tweets); 2) to refine the annotation task based on the conclusions from the initial study including the definition of metaphor, the method of analysis and the annotation guidelines. In this phase, I utilised two data sources to prepare the main dataset in this work covering two categories of tweets. The first category is general domain tweets which is sampled from tweets pertaining to sentiment and emotions from the SemEval 2018 Task 1 on Affect in Tweets (Mohammad et al., 2018). The second category of data is of a political nature which is sampled from tweets around Brexit (Grčar et al., 2017). The data collection for each of these categories is detailed as follows:

Emotional Tweets. People tend to use figurative and metaphoric language while expressing their emotions. This part of the dataset is prepared using emotion related tweets covering a wide range of topics. The data used is a random sample of the Distant Supervision Corpus (DISC) of the English tweets used in the SemEval 2018 Task 1 on Affect in Tweets, hereafter SemEval 2018 AIT DISC dataset[1]. The original dataset is designed to support a range of emotion and affect analysis tasks and consists of about 100 million tweets[2] collected using emotion-related hashtags such as *"angry, happy, surprised, etc"*. The text of around 20,000 tweets is retrieved given their published *tweet-ids* using the Twitter API[3]. The tweets are then preprocessed to remove URLs, elongations (letter repetition, e.g. verrrry), and repeated punctuation as well as duplicated tweets. After that, around 10,000 tweets are arbitrary selected.

Political Tweets. Metaphor plays an important role in political discourse which motivated me to devote part of the dataset to political tweets. My goal is to manually annotate tweets related to the Brexit referendum for metaphor. In order to prepare this subset of the dataset, I employed the Brexit Stance Annotated Tweets Corpus[4] introduced by Grčar et al. (2017). The original dataset comprises 4.5 million tweets collected in the period from May 12, 2016 to June 24, 2016 about Brexit and manually annotated for stance. The text of around 400,000 tweets on the referendum day is retrieved from the published *tweet-ids*. These tweets contained a lot of duplicated tweets and re-tweets. The data is cleaned and preprocessed similar to the emotional tweets as discussed above, yielding around 170,000 tweets.

1 Available online at:https://competitions.codalab.org/competitions/17751#learn_the_details-datasets
2 Only the *tweet-id*s were released and not the tweet text due to copyright and privacy issues.
3 Twitter API: https://developer.twitter.com/en/docs/api-reference-index
4 available online at: https://www.clarin.si/repository/xmlui/handle/11356/1135

4.2.1.2 *Initial Annotation Scheme*

I developed a preliminary annotation scheme and tested it through an in-house pilot annotation experiment before migrating the annotation experiment to the paid crowdsourcing platform MTurk. The goal of this initial annotation experiment is to 1) evaluate the proposed annotation scheme; 2) examine the clarity of the annotation guidelines given to the annotators; 3) assess the effect of the chosen dataset topic genre on the annotation quality. In this initial scheme, the annotators are asked to highlight the words which are used metaphorically relying on their own intuition, and then mark the tweet depending on the presence of a linguistic metaphor as *"Metaphor"* or *"NotMetaphor"*. In this experiment, 200 tweets were extracted from the SemEval 2015 SAFL dataset (more details about this dataset were given in Subsection 3.2.2). The tweets are sarcastic and ironic in nature. The annotation is done by three native speakers of English from Australia, England, and Ireland. One of the annotators has a background in computational linguistics while the others have a more general computer science background (with some knowledge of NLP). The annotators were given several examples to explain the annotation process. A set of guidelines are developed for this annotation experiment in which the annotators were instructed to, first, read the whole tweet to establish a general understanding of the meaning. Then, mark it as metaphoric or not if they suspect that it contains a metaphoric expression(s) based on their intuition taking into account the given definition of a metaphor. A tweet might contain one or more metaphors or might not contain any metaphors. Finally, the annotators were asked to highlight the word(s) that according to their intuition has a metaphorical sense.

The annotators achieved an inter-annotator agreement of 0.41 in terms of Fleiss' kappa (Fleiss, 1971). Although the level of agreement was quite low, this was expected as the metaphor definition depends on the native speaker's intuition. Moreover, the annotators have to examine the whole tweet carefully to identify a possible metaphor which adds a cognition load on them. The number of annotated metaphors varies between individual annotators with a maximum percentage of metaphors of 22%. According to the annotators, the task was seen as quite difficult because it was very hard to pick the boundary between metaphoric and literal expressions. A reason for this is perhaps the ironic nature of the tweets, with some authors deliberately being ambiguous. Sometimes the lack of background knowledge adds extra complexity to the task. Another important challenge is the use of highly conventionalised language where discerning the metaphoricity of an expression depended, in part, on the annotators' cultural background given that the annotators have different nationalities. The question that poses itself here is how to draw a strict line about which word/expression should be considered as a metaphor and which not.

I concluded from this experiment that this initial annotation scheme is difficult to generalise and employ on a larger dataset due to the previously mentioned limitations. It would be still an expensive task in terms of the time and effort consumed by the annotators. This initial annotation scheme is in a way similar to the MIP scheme but without relying on dictionaries and only relying on the annotators intuition to define metaphor. Thus, similar to MIP, the annotators have to go through a series of decisions to discern a metaphoric expression in the given tweet. In an attempt to reduce such cognitive load on the annotators, I explored the usage of WordNet as a reference for sense distinction on 100 tweets. An inter-annotator agreement (IAA) agreement of 0.21 was achieved which is extremely low due to the annotators' disagreement on what they believed to be metaphors in their initial judgement, therefore they checked WordNet for different expressions. This initial pilot study and the informal discussion with the annotators in addition to the aforementioned raised concerns revealed that 1) this dataset is not suitable for the annotation;

2) the annotation scheme needs to be further improved to reduce the cognitive load on the annotators and maintain consistency. Therefore, first, I proceeded with the other data sources discussed in Subsection 4.2.1.1 to help improve the quality of the proposed dataset. Secondly, I improved the annotation scheme as will be discussed in the next subsections.

4.2.1.3 *Weakly Annotated dataset*

In order to address the limitations of the initial annotation experiments, I refined the annotation scheme focusing on the adopted definition of metaphor and the way adopted to analyse (scan) the tweet to identify metaphors. I proposed preparing a weakly annotated dataset using a metaphor identification system, to reduce the cognitive load for annotators and maintain consistency. This system will be used to identify potential metaphoric expressions in tweets. Then, MTurk will be employed to ask a number of annotators to identify the expressions that are metaphorical in their judgement from these pre-identified ones. This way, the cognitive load on the annotators will be reduced while maintaining consistency. Figure 4.1 shows the process of creating the tweets dataset.

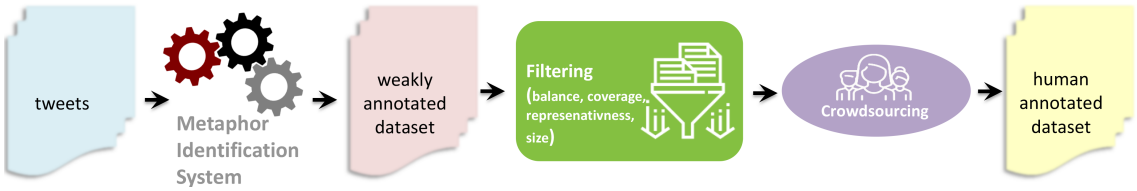


Figure 4.1: The proposed approach to create a dataset of tweets for metaphor identification.

As depicted, the tweets are first processed using a weakly supervised system. The system, which will be discussed in depth in Chapter 5, makes use of distributed representations of word meaning to capture metaphoricity focusing on identifying verb-noun pairs where the verb is used metaphorically. These verb-noun pairs are highlighted using the Stanford dependency parser (Chen and Manning, 2014). Then pre-trained word embeddings are employed to measure the semantic similarity between the candidate pair and a predefined seed set of metaphors. The given candidate is classified using a previously optimised similarity threshold. As a result, a weakly annotated dataset is prepared which comprises the emotional and political tweets discussed in Subsection 4.2.1.1.

Since the employed system is weakly supervised and is not trained on a specific dataset or text genre, it does not limit the final dataset. The arbitrarily selected tweets from both the emotional tweets and the political tweets subsets are used individually as input to the system which highlights the verb-direct object pairs using the parser, as mentioned earlier, as potential candidates for metaphor classification. A candidate is classified as a metaphor or not by measuring its semantic similarity to a predefined small seed set of metaphors which acts as an existing known metaphors sample. Metaphor classification is performed based on a previously calculated similarity threshold value. The system labelled around 42% and 48% of the candidates as metaphorical expressions from the emotional tweets subset and the political tweets subset, respectively.

4.2.1.4 *Dataset Compilation*

Now that I have prepared two weakly annotated subsets of emotional and political tweets, my approach for selecting the final subset of each category of tweets is driven by achieving the following factors:

1. **Verbs Representativeness (Verb Coverage)**: the dataset should cover a wide range of verbs and a variety of associated nouns.
2. **Sense Coverage**: ideally each verb should appear at least once in its metaphoric sense and once literally. If the verb does not have one of these senses, more examples should be included. Moreover, unique object arguments of each verb should be represented.
3. **Size**: to ensure usability in a machine learning setting, the dataset should be sizeable.
4. **Balance**: in order to avoid bias towards a certain class while training an identification model, it is preferable to have a balanced dataset that equally represents positive and negative examples.

To ensure verbs representativeness, I employed a set of 5,647 expressions of verb-direct object grammar relations to obtain the final tweets dataset. Part of this set is collected from the MOH-X dataset, which contains 647 verb-direct object pairs. The other part is collected by utilising the training split of the VUAMC (*Verbs* track) from the NAACL 2018 Metaphor Shared Task. I retrieved the original sentences of the annotated verbs in this subset from the VUAMC, which yielded around 8,000 sentences. I then parsed these sentences using the Stanford dependency parser and extracted 4,526 verb-direct object pairs.

For each verb in the set[5], all the tweets that contain this verb are extracted without regard to the associated noun (object) argument or the initial metaphoric/literal classification of the weakly supervised system. This step yielded around 3,000 instances from the emotional tweets subset and 38,000 instances from the political tweets subset. For each verb, at least one metaphoric instance and one literal instance are randomly selected using the initial classification by the system to ensure balance, e.g. *"find love"* vs *"find car key"* and *"send help"* vs *"send email"*. Also I ensured the uniqueness of the noun argument associated with each target verb to ensure sense coverage within each subset and across both subsets meaning that the same verb appearing in both subsets has different nouns in order to cover a lot of arguments. Each instance should not exceed five words such as *"send some stupid memory"* or *"abandon a humanitarian approach"*. I observed that the parser more frequently made errors on these longer phrases and thus removing them eliminated many erroneous sentences. Moreover, from preliminary experiments, I realised that the annotators got confused when there are multiple adjectives between the verb and the direct object in a given expression and focused on them instead of the object. Although it might be argued that a random set of the tweets could have been selected but this will not achieve the goal of verb and sense coverage. Moreover, another approach to ensure verb representativeness would have been employing VerbNet (Schuler, 2006) but I wanted to be sure that the majority of selected verbs have metaphoric usages.

The final dataset comprises around 2,500 tweets of which around 1,100 tweets are emotional tweets of general topics and around 1,400 tweets are political tweets related to Brexit. Each tweet has a highlighted verb-direct object expression that needs to be classified according to the

5 The number of unique verbs (lemma) in this set is 1,134 covering various classes.

metaphoricity of the verb given the accompanying noun (direct object) and the given context. The next step is to employ the paid crowdsourcing platform MTurk to setup the annotation experiment in order to manually annotate the expressions in these tweets. Table 4.1 shows examples of the different instances that appeared in the emotional and political tweets subsets.

Table 4.1: Examples of the instances appearing in the emotional and political tweets subsets and the corresponding classification of the employed weakly supervised system. *The human annotation disagrees with the system annotation on these examples.

Emotional Tweets	System Classification	Political Tweets	System Classification
accept the fact	metaphor	add financial chaos	not metaphor*
attract hate	metaphor	back #brexit cause	metaphor
break ego	not metaphor*	blame heavy rain	not metaphor
deserves a chance	metaphor*	claim back democracy	metaphor
have time	metaphor	claiming expenses	metaphor*
bring happiness	metaphor	have a say	metaphor
grab chance	metaphor	gain more seats	not metaphor*
grab bike	not metaphor	grab a black biro	not metaphor
hold phone	not metaphor	hand over britain	not metaphor*
join team	not metaphor	make history	metaphor
win game	not metaphor	write your vote	not metaphor

4.2.2 Annotation Process

Having completed the data compilation as discussed in the previous section, the dataset is ready to be used in the main annotation experiment which will be set up on the paid crowdsourcing platform MTurk. The following sections will address the employed annotation scheme highlighting the adopted view of metaphor and the annotation procedure and guidelines. Finally the annotation evaluation will be explained in detail.

4.2.2.1 *Metaphor Definition*

In this thesis, I follow the conceptual metaphor theory (CMT) to view metaphor where a concept (represented by a word sense) can be borrowed to represent another concept by exploiting common or single properties of both concepts. Further, this thesis views a word or an expression as metaphoric if it has at least one basic/literal sense and a secondary metaphoric sense. The literal sense is more concrete and used to perceive a familiar experience while the metaphoric sense is abstract. As discussed in Chapter 2, the metaphoric sense should resonate semantically with the basic sense which means that the metaphorical sense corresponds closely with the literal sense sharing similar semantic features. For example, the metaphoric expression *"launch a campaign"* aligns with (resonates with) more literal, more concrete expressions such as *"launching a boat"* (Hanks, 2016). The basic sense (meaning) of a given word could be determined using a dictionary similar to the MIP approach or by relying on the native annotator intuition such as in Hovy et al. (2013). The latter is the approach adopted in this work. For preparing this dataset of tweets, I focus only on linguistic metaphors of verb-direct object grammar relations. I made some distinctions as follows:

Idioms and Similes. As explained in Chapter 2, I make a distinction between metaphors and other figures of speech that they might be confused with, including idioms and similes. Idioms such as *"blow the whistle, call the shots, pull the rug out, turn a blind eye, etc."* were filtered manually. I did not encounter any similes in this dataset.

Verbs with No Metaphorical Potential. Following Shutova and Teufel (2010), Klebanov et al. (2014a) and Mohammad et al. (2016), I excluded modal verbs such as *"can, may, etc"* from the dataset in addition to the auxiliary verbs *"be* and *will"* assuming that they exhibit no potential of being used metaphorically.

Verbs with Weak Metaphorical Potential. In addition to verbs that exhibit strong potential of being metaphors, I am interested in investigating the metaphoricity of light verbs such as *"get, give, make, take, etc"* and aspectual verbs such as *"begin, end, finish, start, stop, etc"* as well as other verbs such as *"accept, choose, cause, remember, etc"* in addition to the auxiliary verbs[6] *"do* and *have"*. Subsection 4.2.3 presents an analysis of these verbs as they appeared in the proposed dataset.

4.2.2.2 *Annotation Task*

The annotation task is concerned with the identification of linguistic metaphors in tweets. The main goal is to discern the metaphoricity of a target verb in a highlighted verb-noun expression in a given tweet. I set up the annotation task on MTurk. Five native English speakers were hired to annotate the dataset whose field of study is bachelor of arts with either English, journalism or creative writing. The annotation task is formulated as follows:

Task Definition. The annotators were asked to review the tweets and classify the highlighted expression as being used metaphorically or not, based on the provided definition of metaphor and their intuition of the basic sense of the verb. Given the annotators' background in English studies and that they already knew what a metaphor is from a linguistic point of view, it was not hard for them to understand the CMT view of metaphor.

Guidelines and Task Design. Each tweet has a highlighted expression of a verb-noun expression. The annotators were instructed to follow a set of guidelines in order to annotate the highlighted expression for metaphoricity, which are:

1. Read the whole tweet to establish a general understanding of the meaning.
2. Determine the basic meaning of the verb in the highlighted expression. Then, examine the noun (object) accompanying the verb and check whether the basic sense of the verb can be applied to it or not. If it can not, then the verb is probably used metaphorically.
3. Select how certain they are about their answer.

These steps were represented in the task as three questions appearing to the annotators on MTurk. Figure 4.2 shows the designed annotation interface which ensured annotation speed through simple questions.

Reading the whole tweet is important as giving a decision based on reading the highlighted expression only is not enough and leads to inaccurate results. The annotators can skip the

6 Previous research including Shutova and Teufel (2010); Klebanov et al. (2014a); Mohammad et al. (2016) excluded the auxiliary verbs from their datasets.

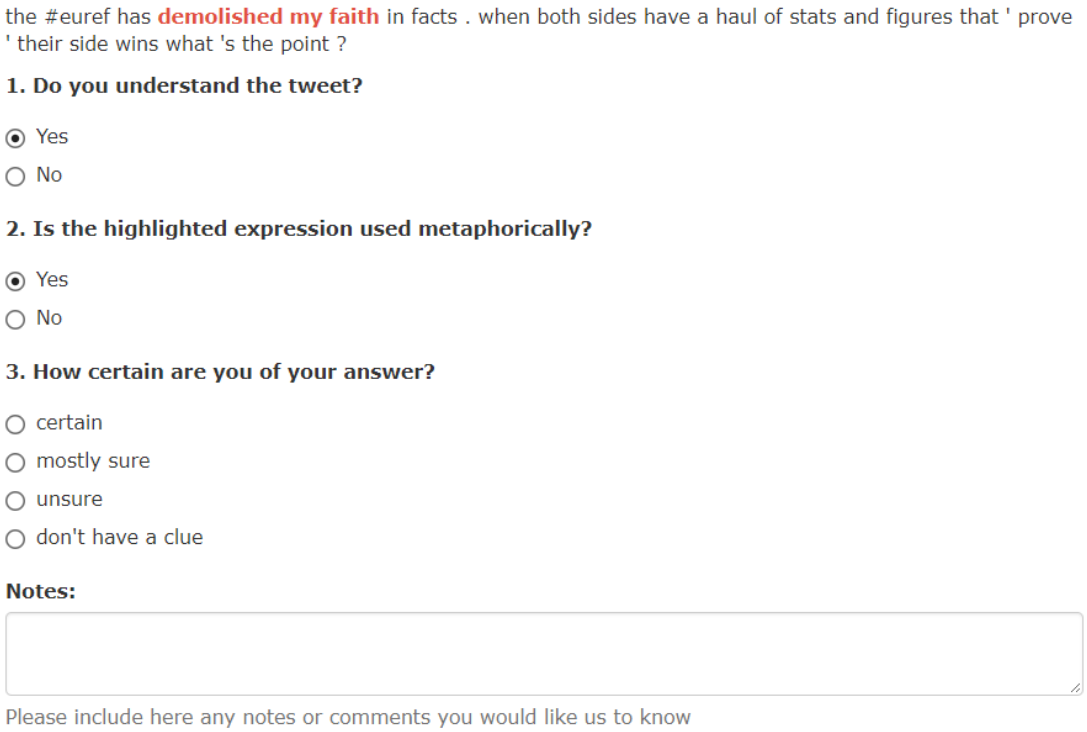


Figure 4.2: A screenshot of the questions in the annotation task given to the annotators on MTurk to identify metaphors in tweets.

tweet if they do not understand it but a threshold is set for skipping tweets. If the annotator is confused about whether an expression is a metaphor or not they were asked to select the *"don't have a clue"* option in question 3. However, a limit was implemented for this choice in order to avoid priming. The annotators were encouraged to add some notes regarding their confusion or any insights they would like to share. The annotators were provided with several examples to explain the annotation process and to demonstrate the definition of metaphor adopted by this work as well as showing how to discern the basic sense of a verb.

Task Settings. I created three annotation tasks on MTurk. The first task is a demo task of 120 tweets from the emotional tweets subset. These tweets included 20 gold tweets with known answers which were obtained by searching the emotional tweets subset for metaphoric expressions (positive examples) from the MOH-X dataset as well as including some negative examples. This task acted as a training demo to familiarise the annotators with the platform and to measure the understanding of the task. Moreover, it acted as a test for selecting the best preforming annotators among all applicants. The efficiency of each applicant is measured in terms of: 1) the time taken to finish the task, 2) the amount of skipped questions, and 3) the quality of answers which is measured based on the gold tweets. The top five candidates were selected to proceed with the main tasks. The second task is the annotation of the emotional tweets subset and the third task was devoted to annotating the political tweets subset. Having two tasks after the demo task allowed for discussing the main annotation disagreements between the annotators. Main discrepancies were brought to open discussion between the annotators and led to some observations which will be highlighted in Subsection 4.2.3.

The tasks are designed as pages of 10 tweets each. Pages are referred to as a human intelligence tasks (HITs) by MTurk and annotators were paid per HIT (page). The time taken to annotate around 200 tweets was estimated to be one hour; therefore, 60 cents were paid for each page. This comes down to $12 per hour, which aligns with the minimum wage regulations of the country where the annotators resided at the time of this experiment.

4.2.2.3 *Evaluation*

In order to assess the reliability of the annotations, the IAA is measured in terms of Fleiss' kappa. Typically, the scores are interpreted using the Landis and Koch (1977) scale where scores of over 0.6 are considered substantial (for more details on this see Section 2.2.1.1). Moreover, the points of agreement and disagreement between the annotators are investigated.

Inter-annotator Agreement. Table 4.2 shows the obtained IAA scores in terms of Fleiss' kappa between the five annotators. As discussed earlier, I wanted to have a deeper look into light and aspectual verbs, as well as verbs with weak metaphoric potential, so I computed the IAA with and without these verbs for each subset of the tweets dataset. As observed from the results, the annotators were able to achieve a substantial agreement (as for Landis and Koch (1977) scale[7]) on the demo task as well as the emotional tweets and the political tweets subsets. After the demo task, the annotators were instructed to pay extra attention to light verbs and to be consistent with similar abstract nouns as much as they can, meaning that an expression such as *"give hope"* would often have the same annotation as *"give anxiety/faith"*. To ensure better performance and avoid distraction, the annotators were advised to annotate around 300 tweets per day and resume after reading the instructions again. Since there is no automatic control of this rule, I verified that all annotators adhered to this rule by manually checking the time stamps of the annotated tweets.

Points of (Dis-)agreement. Tables 4.3 and 4.4 list examples of the agreements and disagreements between the five annotators. The majority of disagreements centred around light verbs and verbs with weak metaphoric potential. The next subsection discusses the annotation results in detail and presents the statistics of the dataset.

Table 4.2: Inter-Annotator Agreement scores of the metaphor dataset of tweets in terms of Fleiss' kappa among the five annotators. *The excluded verbs are light verbs, aspectual verbs, in addition to weak metaphoric potential verbs including *"accept, choose, enjoy, imagine, know, love, need, remember, require,* and *want"*

	Fleiss' kappa		
	partial exclusion (keeping light verbs)	**full exclusion***	**no exclusion**
Demo Task (120 tweets)	0.627 (106 annotated instances)	0.715 (85 annotated instances)	0.623 (108 annotated instances)
Emotional Tweets (1,070 tweets)	0.742 (884 annotated instances)	0.732 (738 annotated instances)	0.701 (1,054 annotated instances)
Political Tweets (1,391 tweets)	0.806 (1,341 annotated instances)	0.805 (1,328 annotated instances)	0.802 (1,389 annotated instances)

7 Landis and Koch (1977) consider fair agreement for kappa scores over 0.2, moderate agreement for scores over 0.4 and substantial agreement for scores over 0.6.

Table 4.3: Examples of agreements among the five annotators (100% majority vote) to identify metaphors in tweets.

		majority vote
Emotional Tweets	its great to be happy, but its even better to **bring happiness** to others.	metaphor
	make memories you will look back and smile at.	
	as long as the left stays so ugly, bitter, annoying & unlikeable, they will not **win any elections**...	not metaphor
	they forget this when they **have money** and start tweeting like they have all the answers	
Political Tweets	make or **break moment** today! together we are stronger! vote remain #strongerin #euref	metaphor
	... cameron can not win this #euref without your support. how many will **lend their support** to...	
	person's details taken by police for offering to **lend a pen** to voters, what a joke.	not metaphor
	in just a couple of days, no one will ever have to **utter the word** 'brexit' ever again	

Table 4.4: Examples of disagreements among the five annotators (60% majority vote) to identify metaphors in tweets.

		majority vote
Emotional Tweets	someone should **make a brand** based off of triangle noodles that glow in the dark. call it illuminoodle...	metaphor
	smile for the camera, billy o. if you **need a smile** every day then #adoptadonkey @donkey-sanctuary	
	cities are full of mundane spaces. **imagine the potential** to transform them into catalysts for positive emotions	not metaphor
	our captors are treating us well and we are very happy and well **enjoying their kind hospitality**	
Political Tweets	perhaps we can **achieve a cohesive society** when the referendum is over, but it does not feel like that is possible. #euref	metaphor
	#euref conspiracy theories **predict people's voting intentions**. will they sway today's vote?	
	democracy works there's still time. british people can not be bullied do not **believe the fear** #voteleave	not metaphor
	what's interesting here is not the figure but that it was from an online poll which has always **favoured the leave**.	

4.2.3 Dataset Statistics and Linguistic Analysis

An exploratory analysis is conducted to better understand the properties of the dataset and highlight the main linguistic findings. The following subsections discuss this analysis in detail in addition to highlighting the statistics of the annotated dataset.

4.2.3.1 *Statistics*

The statistics of each subset of the dataset is presented in Table 4.5 focusing on different statistical aspects of the dataset. It is worth mentioning that the political tweets subset contains 431 unique verbs that did not appear in the emotional tweets subset. The text of the political tweets was more understandable and structured. The emotional tweets subset contains some tweets about movies and games that sometimes the annotators found hard to understand.

Table 4.5: Statistics of the proposed metaphor dataset of tweets, namely the ZayTw dataset.

	Demo Task	Emotional Tweets	Political Tweets
# of tweets	120	1,070	1,390
# of unique verb-direct object (noun) pairs	119	1,069	1,390
Average tweet length	23.82	22.14	21.12
# of unique verbs (lemma) (in the annotated verb-noun pairs)	71	321	676
# of unique nouns (in the annotated verb-noun pairs)	102	725	706
% instances annotated as metaphors	63.15%	50.47%	58.16%
% instances annotated as not metaphors	36.84%	49.54%	41.84%
% instances annotated with agreement majority vote of 60%	20.17%	10.39%	12.29%
# of non-understandable tweets (skipped)	5.2%	1.68%	0.14%

4.2.3.2 *Linguistic Analysis*

As observed from the IAA values listed in Table 4.2, light and aspectual verbs as well as some other verbs represent a major source of confusion among the annotators. Although other researchers pointed out that they exhibit low potential of being metaphors and excluded them from their dataset, the dataset covers different examples of these verbs with different senses/nouns. The majority vote of the annotators on such cases could give some insight on the cases where these verbs can exhibit metaphorical sense.

In the following paragraphs, I provide a linguistic analysis of the proposed dataset performed by manual inspection.

- The majority of annotators tend to agree that the verb *"have"* exhibits a metaphoric sense when it comes with abstract nouns such as *"anxiety, hope* and *support"* as well as other arguments including *"meeting, question, theory, time, skill* and *vote"* . On the other hand, it is used literally with nouns such as *"clothes, friend, illness, license* and *money"*. The annotators find the light verbs *"get, give* and *take"* to be more straightforward while discerning their metaphoric and literal usages. They agreed on their metaphorical usage with abstract nouns such as *"chance, happiness, smile, time* and *victory"* and their literal usage with tangible concepts including *"food, job, medication, money, notification* and *results"*.

- Regarding the verb *"make"* the annotators agreed that as long as the accompanied noun cannot be *crafted* then it is used metaphorically. Metaphors with this verb include *"difference, friends, money, progress* and*time"*, while literal ones include *"breakfast, mistake, movie, noise* and *plan"*.

- The nouns occurring with the verb *"start"* in metaphoric expressions include *"bank, brand* and *friendship"*. Moreover, there are some rhetorical expressions such as *"start your leadership journey/living/new begining"*. The nouns appearing in the expressions classified as literal include *"argument, car, course, conversation* and *petition"*. The verb *"end"* occurred with *"horror* and *feud"* metaphorically and *"thread* and *contract"* literally according to the majority vote.

- The annotators agreed that nouns such as *"food, hospitality, life* and *music"* occurring with the verb *"enjoy"* form literal expressions while the only metaphoric instance is *"enjoy immunity"*. In the case of the verb *"love"*, the majority of annotators agreed that it is not used metaphorically as one can love/hate anything with no metaphorical mapping

between concepts. The disagreements revolve around the cases when the expression is an exaggeration or a hyperbole e.g. *"love this idea/fact/book"*.

- Expressions with stative verbs of thought such as *"remember* and *imagine"* are classified as non-metaphoric. The only debate was about the expression *"... remember that time when ..."* as, according to the annotators, it is a well-known phrase (fixed expression).

- I looked at the verbs *"find* and *lose"* and they were easy to annotate following the mapping between abstract and concrete senses. They are classified as metaphors with abstract nouns such as *"love, opportunity* and *support"* as well as something virtual such as *"lose a seat (in the parliament)"*. However, it was not the case for the verb *"win"*. The majority agreed that expressions such as *"win award/election/game"* are literal expressions while the only disagreement was on the expression *"win a battle"* (three out of five annotators agreed that it is used metaphorically).

- Annotating the verbs *"accept* and *reject"* was intriguing as well. The majority of annotators classified the instances *"accept the fact/prices"* as literal while in their view *"accept your past"* is a metaphor.

- An issue is raised regarding annotating expressions that contain the verbs *"cause, blame, need* and *want"*. Most agreed that *"need apology/job/life"* can be considered as metaphor while *"need date/service"* is not.

From this analysis, it can be concluded that following the adopted definition and view of metaphor helped the annotators to discern the sense of the highlighted verbs. Additionally, highlighting the targeted expression focused the annotators attention on discerning its metaphoricity instead of trying to figure out where the metaphor is in the whole tweet. This, in turn, reduced the time and cognitive load experienced by annotators and decreased the number of decisions that need to be made. On the other hand, relying on the annotators' intuition (guided by the given instructions) to decide the basic meaning of the verb led to some disagreements but it was more time and effort efficient than other options. Light verbs are often used metaphorically with abstract nouns. There are some verbs exhibiting weak metaphoric potential and classifying them is not as straightforward as other verbs. However, they might be used metaphorically on occasions, but larger data is required to discern these cases and find a solid pattern to define their metaphoricity. Hyperbole and exaggerations and other statements that are not meant to be taken literally need further analysis to discern their metaphoricity. Sharing and discussing the disagreements after each annotation task among the annotators helped to have a better understanding of the task.

4.3 ADAPTED WORD-LEVEL BENCHMARK DATASETS

As discussed in the introduction, among the challenges that face research adopting the relation-level metaphor processing paradigm is the availability and size of large benchmark datasets when compared to the ones developed to support the word-level paradigm. This impedes cross-paradigm comparisons and therefore makes it hard to have a conclusive performance interpretation of the proposed state-of-the-art systems. This problem was also discussed in Shutova

(2015) where she called for a unified large corpus in order to have a reliable evaluation and informative performance interpretation. In this work, I attempt to fill this gap by adapting the VUAMC and the TroFi benchmark datasets to better suit relation-level metaphor identification. The conversion from word-level to relation-level annotation is not straightforward and requires extra annotation effort, as discussed in Subsection 3.2.1. I therefore proposed and applied an adaptation method using minimum annotation effort while maintaining annotation accuracy and consistency which answers **RQ1.b**. This work is inspired by Tsvetkov et al. (2014) and Shutova et al. (2016) who attempted to adapt existing word-level metaphor identification datasets to suit their relation-level identification approaches (refer to Subsection 3.2.2.3 for more information). Furthermore, in this work also, I looked at the limitations of current datasets that are originally designed to support relation-level metaphor processing. Therefore I proposed an improved version of the TSV benchmark dataset to allow its usage by approaches that utilise the full context around the metaphoric expressions.

Although, the VUAMC is the most well-known and widely used corpus for metaphor identification, it is not possible to apply it to relation-level metaphor identification without further annotation effort. To the best of my knowledge, there is no attempt to adapt the benchmark VUAMC to suit relation-level metaphor identification. This has discouraged other researchers focusing on relation-level approaches from employing this dataset such as the work done by Rei et al. (2017), Bulat et al. (2017), Shutova et al. (2016) and Tsvetkov et al. (2014) who did not evaluate or compare their approaches using this dataset. This is also the case for the TroFi dataset which is one of the earliest balanced datasets annotated to identify metaphoric verbs on the word level. Although, Tsvetkov et al. (2014) adapted the TroFi dataset focusing on subject-verb-object (SVO) grammar relations, researchers who focused on verb-noun relational analysis were not able to utilise this adapted version. On the other hand, the TSV dataset is the only available annotated dataset for relation-level metaphor identification that addresses adjective-noun grammatical relations. However, the main issue with this dataset is the absence of full sentences in the training set leaving a relatively small test set that has full sentences which limits its usage for state-of-the-art approaches that rely on using the full context. In this work, I introduce the first adapted version of the VUAMC to better suit relation-level metaphor processing. Furthermore, I adapt the TroFi dataset focusing on verb-direct object relations and I extend the TSV datasets to support approaches that rely on using the full context around the annotated metaphors.

As mentioned in Chapter 3, the widely used benchmark datasets for word-level metaphor identification are the TroFi, VUAMC and MOH datasets, while the TSV and MOH-X datasets are commonly used for relation-level metaphor identification. Table 4.6, adapted from Table 3.2, revisits the properties of each dataset. Since the MOH dataset was already adapted by Shutova et al. (2016), and referred to as the MOH-X dataset, to support relation-level metaphor identification of verb-noun grammar relations, I focus, in this work, on the other three datasets (see Section 3.2.2 for extensive details of each dataset).

4.3.1 Dataset Adaptation Methodology

In this section, I discuss the methodology of adapting the VUAMC, TroFi and TSV datasets to better suit relation-level metaphor processing. I employed a semi-automatic approach in order to avoid extensive manual annotation.

Table 4.6: Statistics of widely used benchmark datasets for linguistic metaphor identification.

Level of analysis	Dataset	Syntactic structure	Data Source	Size	% Metaphors
word level	TroFi (Birke and Sarkar, 2006)	verb	WSJ	3,727 sentences	57.5%
	VUAMC (Steen et al., 2010)	all POS	BNC	~16,000 sentences (~200,000 words)	12.5%
	MOH (Mohammad et al., 2016)	verb	WordNet	1,639 sentences	25%
relation level	TSV (Tsvetkov et al., 2014)	adjective–noun	Web	~2,000 adj-noun pairs	50%
	MOH-X (Shutova et al., 2016)	verb-direct object; subject-verb	WordNet	647 sentences	48.8%

4.3.1.1 *VUAMC and TroFi dataset Adaptation*

As discussed earlier, relation-level metaphor identification focuses on a specific grammatical relation that represents the source and target domains of the metaphor. The datasets that are initially annotated for word-level processing have the word that represents the source domain (the *vehicle*) labelled as metaphoric regardless of its *tenor* since it is word-by-word classification. Therefore, in order to adapt them to suit relation-level processing, the associated word(s) that represent the target domain (the *tenor*) need to be identified.
My approach towards adapting the datasets annotated on the word level is as follows:

1. select the benchmark dataset which is originally annotated on the word level;
2. extract particular grammatical relations focusing on the *vehicle* as the head of the relation (e.g. the verb in a *dobj* or adjective in *amod* relation);
3. retrieve the gold labels from the original dataset based on the metaphoricity of the *vehicle*;
4. verify the correctness of the retrieved relations and the assigned gold label.

In this work, the Stanford dependency parser (Chen and Manning, 2014) is employed to identify the grammar relations (dependencies). Specifically, the recurrent neural network (RNN) parser, pre-trained on the WSJ corpus, is used from the Stanford CoreNLP toolkit (Manning et al., 2014).

For the VUAMC adaptation, the training and test splits provided by the NAACL Metaphor Shared Task in the *Verbs* track are utilised. I focus on this track since I am interested in verb-noun relations where the verb can be used metaphorically. The verbs dataset consists of 17,240 annotated verbs in the training set and 5,874 annotated verbs in the test set. First, the original sentences of these verbs are retrieved from the VUAMC since the shared task released their *ids* and the corresponding gold labels. This yielded around 10,570 sentences in both sets. Then, these sentences are parsed using the Stanford dependency parser and extracted the verb-direct object (i.e. *dobj*) relations, discarding the instances with pronominal or clausal objects[8]. The extracted relations are then filtered to exclude parsing-related errors. Manual inspection is done to ensure that, in a given *dobj* relation, the verb is metaphoric due to the associated object (more details will be given in Subsection 4.3.2). The final adapted dataset comprises 4,420 sentences in the training set and 1,398 in the test set. I kept the train and test splits of the shared task to facilitate the future usage of the adapted dataset in such venues.

8 This is done automatically using regular expressions to select the grammatical relations with certain POS tags.

For the TroFi dataset adaptation, the 3,737 manually annotated English sentences from Birke and Sarkar (2006)[9] are utilised. Each sentence contains either literal or non-literal use for one of 50 English verbs. These sentences are parsed to extract dependency information using the Stanford dependency parser. Then, the extracted relations are filtered to only select the *dobj* relations that include verbs from the 50 verbs list and to eliminate mis-parsing cases. This resulted in a dataset of 1,535 sentences.

Table 4.8 shows the statistics of the adapted VUAMC and TroFi dataset after applying the quality assessment in Subsection 4.3.2. Examples of the annotated sentences from the adapted VUAMC and TroFi dataset are listed in Table 4.9 as they appear in the adapted relation-level version.

4.3.1.2 *TSV Dataset Adaptation*

My main goal when adapting the TSV relation-level dataset is to provide a context for the balanced training set of 1,768 metaphoric and non-metaphoric adjective-noun pairs. Table 4.7 gives examples of the adjective-noun expressions appearing in the original TSV training set[10]. This will allow the computational models to benefit from the contextual knowledge that surrounds the expression. The method used to achieve this goal is to query the Twitter Search API using the adjective-noun pairs and retrieve tweets as the context around these expressions. Among the main motivations behind selecting the user-generated text of tweets, to expand this dataset, are: 1) to encourage and facilitate the study of metaphors in social media contexts; 2) the availability of Twitter data as well as the ease of use of the Twitter API.

Table 4.7: Examples of the annotated adjective-noun expressions in the TSV training dataset.

Metaphor	Non-metaphor
academic gap	abandoned building
big ego	bad kids
blind faith	blind patient
deep sorrow	deep cut
empty life	empty house
fishy offer	frozen food
heated criticism	heated oven
lost freedom	lost hat
raw idea	raw vegetables
shallow character	shallow water
unspeakable power	uninformed citizen
warm smile	warm day

For each expression in the training set, a tweet is retrieved given that its length is more than 10 words and it does not contain more than four hashtags or mentions to ensure that the retrieved context has enough information. Then, the tweets are preprocessed to remove URLs and duplicate tweets. This yielded an adapted training set of 1,764 tweets[11] that contains metaphoric and non-metaphoric expressions of adjective-noun relations. The next step is to ensure the quality of the retrieved content in terms of keeping the metaphoricity of the original expression. This is done manually as will be discussed in the next section. Table 4.8 provides the statistics of the

9 The TroFi dataset is available online at: http://natlang.cs.sfu.ca/software/trofi.html
10 The TSV dataset is available online at: https://github.com/ytsvetko/metaphor
11 Almost a tweet for each expression in the original TSV training set except for the expressions *"colliding contradiction, crisscrossed chaos, hope-sapping poverty, seductive greenery* and *unforgiving policeman"* where there were no corresponding tweets were found.

adapted TSV training dataset after expanding it with full sentences (tweets). Examples of the annotated tweets from the adapted TSV training dataset are given in Table 4.9.

Table 4.8: Statistics of the adapted VUAMC, TroFi and TSV benchmark datasets to better suit relation-level metaphor identification. *The training and test sets from the NAACL Metaphor Shared Task (the *Verbs* track).

	VUAMC*		**TroFi Dataset**	**TSV Dataset**
	training set	**test set**		**training set**
targeted grammar relation	verb-direct object		verb-direct object	adjective-noun
# sentences	4,420	1,398	1,535	1,764
# metaphoric instances	1,675	586	908	881
# non-metaphoric instances	2,745	812	627	883
% metaphors	37.96%	41.92%	59.15%	49.94%

Table 4.9: Examples from the adapted VUAMC, TroFi and TSV benchmark datasets showing the targeted expression and the provided label (1:metaphor; 0:non-metaphor).

	ID	**Text**	**Expression**	**Label**
VUAMC	a1e-..._1_4	Latest corporate unbundler reveals laid-back approach: Roland Franklin, who is leading a 697m pound break-up bid for DRG, talks to Frank Kane	reveals laid-back approach	1
	fpb-..._1150_5	I want you to break the news gently to Gran.	break the news	1
	fpb-..._1117_12	Half an hour after the inspector left, as if to prove his point, the lavatory refused to flush.	prove his point	0
	crs-..._35_12	The Community Health Team had major responsibility for assessing children and recommending provision.	recommending provision	0
TroFi	wsj13:9766_16	And even when that loophole was closed, in 1980, the Japanese decided to absorb the tariff rather than boost prices.	absorb the tariff	1
	wsj77:9805_9	But to improve its profitability , it recently targeted mid-sized businesses as well.	targeted mid-sized businesses	1
	wsj27:5617_6	This time, the ground absorbed the shock waves enough to transfer her images to the metal in bas-relief.	absorbed the shock waves	0
	wsj67:11208_14	Because they 're so accurate, cruise missiles can use conventional bombs to destroy targets that only a few years ago required nuclear warheads.	destroy targets	0
TSV (Training)	1248174...	atsukara Kukuku. How astounding... You've captured my ancient heart. How will you atone for this sin...?	ancient heart	1
	1248238...	@sacrebleu141 @FasslerCynthia But it's exactly what the left wants. Trains the people into blind obedience	blind obedience	1
	1245923...	Bakugou jumped backwards as the warm handle suddenly turned hot and more smoke started pouring into his apartment from the gaps in the door.	warm handle	0
	1248271...	Still have nightmares about waiting tables many years later. Hands down the hardest, most stressful job I've ever had.	stressful job	0

4.3.2 Quality Assessment and Enhancement

In order to assess the quality of the adapted datasets, I proposed a preliminary quality assessment scheme and tested it through an initial experiment on a randomly sampled subset from each dataset. I then employed this scheme to ensure the quality of the complete datasets.

4.3.2.1 *Initial Quality Assessment Experiment*

In this pilot experiment, 100 sentences are randomly drawn from each dataset. Two native English speakers with background in (computational) linguistics were then asked to manually identify the quality of the retrieved sample. Since the datasets were previously annotated and the adopted semi-automatic approach did not alter the original annotations, the main concerns for evaluation are as follows:

For each instance in the VUAMC and the Trofi dataset:

1. to check that the *dobj* dependency is syntactically valid;
2. to ensure that the verb is metaphoric due to the associated object;
3. check if the expression is really a metaphor.

For each instance in the TSV dataset:

1. to ensure that the tweet is in understandable English;
2. to check that the *amod* dependency is syntactically valid;
3. to ensure that the provided context (scrapped tweets) preserves the metaphoricity of the expression.

For the VUAMC, the annotators agreed that in 81.1% of the metaphoric cases, the metaphoricity of the verb is due to the complement direct-object. However, the annotators raised some issues regarding the original annotation of the VUAMC using the MIPVU procedure. Their main concerns were: 1) the quality of the original annotation which is done on the word level without explicitly highlighting the *tenor* or the implicit *ground* of the metaphor; 2) the consistency of the annotations across the corpus which relied on the annotators' judgement of the basic meaning of the given word using a dictionary.

The annotators highlighted that the TSV and TroFi datasets have more reliable annotations that align well with the linguistic definition of metaphor than the VUAMC. This can be attributed to the following reasons: 1) the TSV dataset was originally annotated on the relation level with explicit labelling of the *tenor*; 2) the TroFi dataset comprises carefully selected examples of literal and non-literal usages for 50 particular verbs. For the TroFi dataset, the annotators agreed that all the verbs in the random set were used metaphorically due to the associated direct-object without raising any concerns regarding the original annotation of the dataset.

The manual inspection of the random subset of the TSV dataset revealed that, surprisingly, the provided context for the adjective-noun expressions preserved the meaning and the metaphoric sense of all the queried expressions. I suspected that some ambiguous cases might lead to ambiguous contexts. For example, the expression *"filthy man"*, which is marked as a metaphor in the dataset, could be used literally to describe the hygienic state of a person; however, the retrieved tweet preserved the metaphoric sense of this expression that describes the morality of a person. This might be due to the following reasons: 1) the conventionality and frequency of adjective-noun metaphoric expressions; 2) the nature of the user-generated (conversational) text

of the tweets allows the usage of figurative and metaphoric expressions more frequently than their literal counterparts; 3) the nature of the expressions in the TSV dataset itself in terms of abstractness and concreteness. Further corpus studies are required to investigate this finding.

4.3.2.2 *Data Filtering and Quality Enhancement*

Based on the conclusions of the initial quality assessment, an expert annotator[12] is asked to review the three adapted datasets for quality enhancement following the same scheme. Table 4.10 includes detailed statistics of this quality assessment.

To enhance the **TSV** dataset and ensure its quality, if any of the aforementioned problems is detected the annotator provided another tweet by manually searching Twitter. This is done in a similar way to that adopted by Tsvetkov et al. (2014) while preparing the TSV test set. The annotator noticed that sometimes the tweets contain code-mixed text in English and other language written in Latin letters. These instances are replaced by understandable ones. For the **TroFi** dataset and the **VUAMC**, the annotator corrected the detected parsing errors if possible otherwise the erroneous instances are discarded. Moreover, if the expression is metaphoric due to the associated subject (not the direct object), the expression is corrected and labelled as having an *nsubj* dependency. These expressions are not excluded from the data. Finally, when the annotator disagrees about the metaphoricity of a given instance, it has to be checked first in the original VUAMC dataset and if no annotation error is detected then the instance is flagged to have an annotation disagreement with what the annotator believed to be a metaphor. Aligning with the other two annotators of the pilot experiment, quality and consistency issues are raised about the VUAMC annotation. For example, the verb *"commit"* is labelled five times as a metaphor with the nouns *"acts, bag, government,* and *offence(s)"* and three times as literal with the nouns *"rape,* and *offence(s)"* in very similar contexts. As shown in Table 4.10, the annotator flagged around 5% of the data for annotation doubt or inconsistency. The majority of the inconsistent annotations revolves around the verbs *"receive, form, create, use, make, recognise, feel, enjoy, win* and *reduce"*.

Table 4.10: Statistics of the quality assessment of the three adapted datasets showing the total percentage of instances accepted by the annotator.

Dataset		Total % accepted by annotator
TSV	the Tweet is in understandable English?	70%
	the relation is syntactically valid?	82.75%
	did the context (tweet) keep the metaphoric sense of the expression?	99.36%
TroFi	the relation is syntactically valid?	98.52%
	the verb is metaphoric or literal due to the associated object?	100%
VUAMC	the relation is syntactically valid?	98.42%
	the verb is metaphoric or literal due to the associated object?	98.5%
	annotation disagreement or inconsistency	5.45%

12 "expert annotator" means having a computational linguistic background and extensive experience in metaphor processing.

4.4 DEFINITIONS DATASET FOR METAPHOR INTERPRETATION

As hinted in the introduction, automatic metaphor interpretation is much less explored in part due to the lack of publicly available datasets. Previous approaches to metaphor interpretation cast the task as either lexical substitution, paraphrase generation or definition generation (more details were provided in Section 3.3.2). The creation of the dataset depends on the task definition which in turn is determined by the end application. In this work, I look at metaphor interpretation as a definition generation task with the aim to aid language learners and non-native speakers to understand metaphors as well as enrich the development process of lexical resources.

Manually annotating a dataset for metaphor interpretation (either to provide a definition/-explanation or to paraphrase the expression) is a very demanding task which requires effort and time from a human annotator to figure out the meaning of a given metaphor and provide a literal explanation (if possible) for it. Moreover, it is a highly subjective task; the meaning of an expression can vary from one annotator to the other depending on the context and the cultural background of the annotator. This will introduce a question of accuracy and consistency of the created dataset and the submitted annotations.

Only two datasets are publicly available that support metaphor interpretation, one prepared for lexical substitution and the other for paraphrase generation. These datasets are introduced by Shutova (2010) and Bizzoni and Lappin (2018) as discussed in Section 3.3.2. These available datasets have important limitations in terms of size, representativeness and quality. Both datasets are relatively small which limits their usage for machine learning applications. Also, they are restricted to a small subset of metaphors which limits their metaphoric coverage and representativeness. Moreover, their annotation technique influences their quality as both datasets are not evaluated in terms of inter-annotator agreement. The work presented in this section attempts to address these issues and seeks an answer to **RQ1.c** by introducing an annotation scheme that employs lexical resources to assist in the creation of the interpretations. I considered several aspects to ensure the dataset quality including:

- data selection to ensure metaphoric coverage and representativeness;
- data compilation to ensure annotation consistency and quality;
- native human annotators' training and expertise;
- clear annotation scheme and guidelines.

The proposed annotation scheme is designed with the goal of reducing the cognitive load for annotators while maintaining accuracy and consistency based on my previous experience and conversations with expert annotators. Dictionaries are employed to automatically compile a list of possible definitions for a given metaphoric expression. These possible candidates of interpretations are generated by employing semantic similarity based on word embeddings. As a result, the first gold standard dataset of metaphor interpretations is produced. The following subsections discuss the data preparation process, the proposed annotation scheme and the evaluation process in detail.

4.4.1 Data Preparation

This section discusses the preparation steps behind the proposed interpretation dataset. The criteria that were followed to select a dataset of already identified metaphors is described. The main concern while choosing a dataset of metaphors is to ensure wide coverage and representativeness. The data compilation process will be then demonstrated where existing lexical resources are employed with the goal to reduce the cognitive load on the annotators while maintaining accuracy and consistency.

4.4.1.1 *Data Source*

The first step towards creating the interpretation dataset is to have a manually annotated dataset where the metaphors are identified. In this work, I am interested in interpreting verb-noun metaphoric expressions in the context of the user-generated text of tweets, therefore I will focus on interpreting the metaphoric instances in the tweets dataset introduced in Section 4.2, hereafter the ZayTw dataset. Table 4.11 revisits some examples from the ZayTw dataset which comprises around 1,500 metaphoric instances.

Table 4.11: Examples of instances appearing in the ZayTw dataset, introduced in Section 4.2, showing verb-direct object metaphoric expressions that can be used as targets for interpretation.

Tweet	Metaphoric Expression
its great to be happy, but its even better to **bring happiness** to others.	bring happiness
make memories you will look back and smile at.	make memories
make or **break moment** today! together we are stronger! vote remain #strongerin #euref	break moment
...cameron can not win this #euref without your support. how many will **lend their support** to...	lend their support

4.4.1.2 *Data Compilation*

Now that I have a set of sentences (tweets) with identified metaphors (verb-noun pairs) that needs to be interpreted, the direct approach would be to ask human annotators to write down a definition of each metaphoric expression. As discussed earlier, this task will be very demanding and highly subjective. It will require a lot of time and cognitive effort from the annotators to interpret the metaphor after understanding the interaction between its components (the *tenor* or the noun and the *vehicle* or the verb). With the aim to reduce this cognitive load and maintain consistency, I bootstrap an initial list of possible interpretations for the highlighted metaphor (targeted verb-noun pair) from lexical resources and provide it to the annotators.

The idea comes from the question: what would a language learner (a non-native speaker) do when encountering a new[13] metaphoric expression in a given text? One way could be to look it up in a dictionary. Since there is no specific dictionary for metaphors, sometimes the full expression could be found in a dictionary where very conventionalised metaphors are labelled as idioms[14]. For the majority of cases, where there is no direct match of the whole metaphoric expression (verb-direct object pair) in a dictionary, the user could start looking for the verb in the

13 By "new" here I do not mean "novel" in the absolute sense but I mean that the language learner did not know the metaphoric expression beforehand.

14 I argue against this generalisation from a linguistic point of view and I clarified the difference between metaphors and idioms in Section 2.1.4.

dictionary. Then, try to find the nearest definition that can match the metaphoric sense of the verb and at the same time represent its interaction with the accompanying noun. To automate this idea, there are two approaches to pursue; first, to check out metaphors that are labelled as idioms in lexical resources and extract their definitions. Second, to check out the nearest definition of the verb in a dictionary that could be applied to the noun to convey a metaphoric sense. Both methods should be validated by human annotators.

METAPHORS IN WIKTIONARY IDIOMS

As discussed in Section 2.1.4, an idiom is an inseparable lexical unit which, unlike a metaphor, its figurative meaning cannot be guessed from the meanings of the individual words. Commonly used metaphors which became conventionalised in the language found their way into lexical resources (dictionaries) under the idioms category. As discussed earlier, it is understandable to assign conventionalised metaphors to an already existing label rather than creating a new one. Wiktionary[15] is a multilingual online lexicon (dictionary) edited and maintained by volunteers in a collaborative way and is considered an important resource for natural language processing research (Meyer and Gurevych, 2012). The lexicon has a large set of idioms under the *English Idioms Category*[16]. In this work, Wiktionary's API[17] is used to query the idioms category in order to automatically get the definition of the metaphoric expressions in the dataset. Table 4.12 shows examples of the metaphors labelled as idioms and their retrieved definition.

Table 4.12: Examples of the metaphoric expressions from the ZayTw dataset found under Wiktionary's *English Idioms Category*.

Metaphor	Definition
blow someone's mind	to astonish someone, to flabbergast someone.
break a law	to violate a law.
build bridges	to establish links or friendly relations.
cast one's vote	to vote for something.
take a chance	to risk doing something; to try something risky.

Although this category contains around 8,000 idioms, only around 10% of the identified metaphors in the ZayTw dataset were found under this category. It means that the dataset contains only around 140 conventionalised metaphors which are considered fixed expressions and labelled as idioms in Wiktionary. This motivated me to proceed with the second idea of finding the nearest definition of the metaphoric expression in a dictionary as will be discussed in the next section.

NEAREST DEFINITIONS IN A DICTIONARY

Consider the highlighted metaphoric expression in the following tweet:

(4.1) `I want him to participate in Presidential Elections so we can defeat him and` ***`break his ego`*** `[...]`

In this example, the concrete (physical) concept of a brittle object represented by the verb *"break"* is borrowed to express an abstract (emotional) concept represented by the noun *"ego"*. Although

15 https://www.wiktionary.org
16 https://en.wiktionary.org/w/index.php?title=Category:English_idioms
17 https://www.mediawiki.org/wiki/API:Query

the metaphoric expression *"break ego"* is not directly found in a dictionary, there will be a sense for the verb *"break"*, in almost any dictionary, that is related to destroying emotions or people's spirit, will or determination which is, in a sense, related to the concept of the noun *"ego"*. Table 4.13 shows the definition of the verb *"break"* related to emotional concepts in several dictionaries.

Table 4.13: The definition of the verb *"break"* that is related to *"destroying emotions"* in various dictionaries.

Dictionary	Definition
Wiktionary[18]	to cause (a person or animal) to lose spirit or will; to crush the spirits of.
WordNet[19]	weaken or destroy in spirit or body.
Oxford[20]	crush the emotional strength, spirit, or resistance of.
Oxford Learner's[21]	to destroy something or make somebody/something weaker; to become weak or be destroyed.
Longman[22]	to make someone feel that they have been completely defeated and they cannot continue working or living.
Macmillan[23]	to destroy someone's confidence, determination, or happiness.

My hypothesis is that measuring the semantic similarity of the noun of the metaphoric expression against each sense of the verb retrieved from a dictionary can reflect the interaction between the meaning of the components of the metaphor and, in turn, reveal the nearest definition of the metaphoric expression. To experimentally examine this hypothesis, I propose a computational model that employs a dictionary API, pre-trained word embeddings and cosine similarity.

In this work, a sense of a verb is represented by its definition in a dictionary along with the accompanied contextual examples (example sentences). The Oxford Learner's Dictionary is used to retrieve the definitions of a given verb and the example sentences. The reason behind choosing this dictionary is that it offers many contextual examples for each word compared to other examined dictionaries including Wiktionary and WordNet. More contextual examples will help to better model the sense of the verb. Other factors are also considered while choosing the dictionary including the number and granularity of senses that a word has.

The first step, in this proposed approach, is to retrieve the definitions and the sentence examples of each verb in the dataset of metaphors in order to represent the different senses of the verb. Given a metaphoric expression, GloVe pre-trained word embeddings were then used to calculate the cosine similarity between the sense of the verb (represented by the definition and the contextual examples) and the noun of this given metaphoric expression. The proposed approach can be formulated as follows:

- Given a set of metaphoric expressions of verb-noun pairs $M = \{(V, N)\}$, suppose that each verb in M has a set of senses S_v in the dictionary represented by its definition and the sentences examples.

- Each sense is represented by a sequence of words $w_{v,i,1}, w_{v,i,2}, ..., w_{v,i,l}$ where l is the number of words in the i^{th} sense of the verb v in the dictionary for each $i \in S_v$.

- The cosine similarity between the embeddings of the noun n in the metaphoric expression represented as x_n and the embeddings of the words of the verb sense combined into a single vector by mean pooling as $x'_{v,i}$ can be calculated as follows:

$$Similarity = cos(x_n, x'_{v,i}); \forall i \in S_v \tag{4.1}$$

This gives a list of senses (definition and example sentences) ranked according to the similarity score.

– The top three definitions are then obtained as possible candidates to interpret the given metaphor (v, n) according to the highest similarity score. Initial evaluations demonstrated that selecting the top three definitions was a sufficient trade-off between reducing cognitive load and maintaining accuracy.

The Gensim Python library (Rehurek and Sojka, 2010) was used to calculate the similarity. The uncased 300-dimensional GloVe embeddings pre-trained on the Common Crawl dataset were utilised. Table 4.14 lists the nearest three definitions from Oxford Learner's Dictionary, ranked by the cosine similarity score, which could interpret the given metaphoric expressions based on the similarity between the noun of the metaphoric expression and the sense of its verb.

The dataset now comprises around 1,500 tweets with highlighted metaphoric expressions and a list of possible interpretations for each highlighted expression. The next step is to manually annotate this dataset. The annotators will be asked to select one interpretation from the list or provide their own interpretation in case no applicable definition can be found.

Table 4.14: Examples of the nearest definitions from Oxford Learner's Dictionary that could interpret the given metaphoric expressions based on the cosine similarity between the noun of the metaphoric expression and the verb sense.

Metaphoric Expression	Definition	Cosine Similarity
bind country	to unite people, organizations, etc. so that they live or work together more happily or effectively	**0.620**
	to force somebody to do something by making them promise to do it or by making it their duty to do it	0.573
	to tie somebody/something with rope, string, etc. so that they/it cannot move or are held together firmly	0.422
hit economy	to have a bad effect on somebody/something	**0.524**
	to reach a particular level	0.519
	to experience something difficult or unpleasant	0.414
lend support	to give or provide help, support, etc.	**0.743**
	to give money to somebody on condition that they pay it back over a period of time and pay interest on it	0.404
	to give a particular quality to a person or a situation	0.375
meet fear	to experience something, often something unpleasant	**0.669**
	to be in the same place as somebody by chance and talk to them	0.557
	to touch something; to join	0.535
promote intolerance	to help something to happen or develop	**0.385**
	to move somebody to a higher rank or more senior job	0.116
	to move a sports team from playing with one group of teams to playing in a better group	0.085

4.4.2 Annotation Process

The annotation task is set up on the MTurk platform. Six native English speakers were hired to annotate the dataset whose field of study is English. It is worth mentioning that all annotators have the same nationality to rule out cultural background bias.

Task Definition. Given a tweet with a highlighted metaphoric expression, the main goal of the task is to select the most probable definition/interpretation (if exists) of the highlighted expression among the given definitions (similar to manual sense disambiguation but for the metaphoric expression). If the given list does not contain a definition that correctly interprets the metaphor, the annotator is asked to provide a simple definition that explains both the verb and the noun of the metaphor. The annotators are encouraged to consider explaining the meaning of the metaphoric expression to a child, a language learner or a person with a learning difficulty.

Guidelines and Task Design. Each tweet has a highlighted metaphoric expression of a verb-direct object syntactic structure. The annotators were instructed to follow the following set of guidelines:

1. Read the whole tweet to establish a general understanding of the meaning.
2. Focusing on the highlighted expression, read the given definitions and determine which one is the most probable (nearest) definition of the highlighted metaphor. In case no applicable choice is found, select "not applicable".
3. In case of choosing "not applicable", provide a definition to interpret and explain the metaphor in a few words.

These steps were represented in the task as three questions appearing to the annotators on MTurk as shown in Figure 4.2. The list of possible interpretations for each highlighted expression was shuffled before giving it to the annotators in order to avoid the bias and priming where the annotators select the first choice every time. A free text area was provided under each tweet to allow the annotator to write their comments, insights or any confusing issues about the tweet content. The annotators went through a training phase by taking a demo task to familiarise them with the platform and to clarify the annotation process.

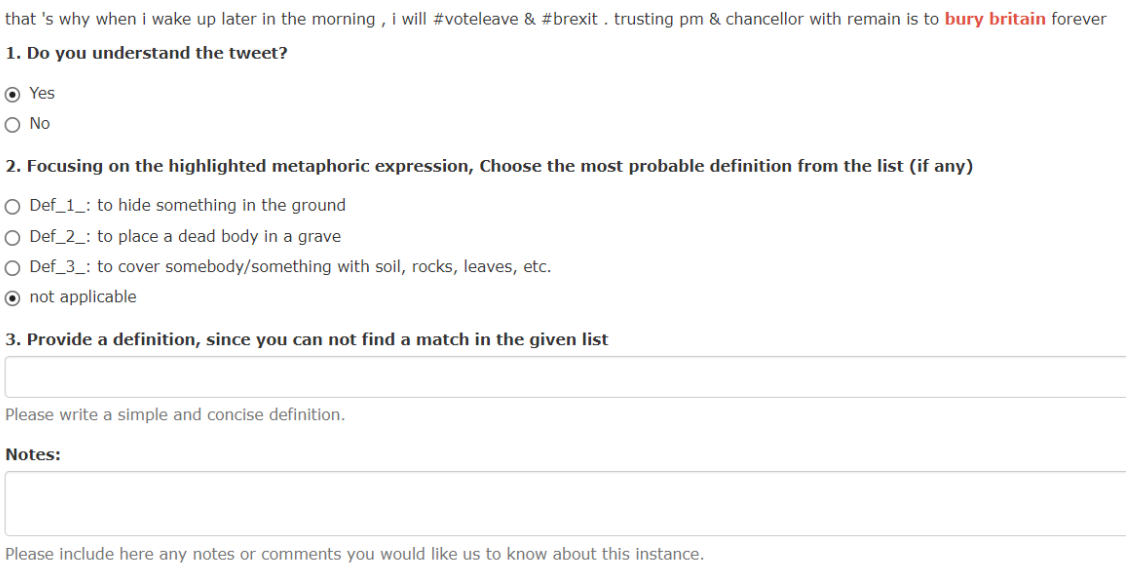


Figure 4.3: An example from the annotation task given to the annotators on MTurk to interpret a highlighted metaphoric expression.

Task Settings. The annotation task is designed as pages of 10 tweets each. The time taken to annotate around 60 tweets was estimated to be one hour; therefore, $1.80 was paid for each page. This comes down to $12 per hour, which aligns with the minimum wage regulations of the country where the annotators resided at the time of this experiment.

4.4.3 Dataset Evaluation and Analysis

This section provides a description of the assessment of the annotation results. The main observations and analysis of the dataset will also be discussed. Moreover, the points of agreement and disagreement between the annotators will be highlighted along with statistical analysis of the dataset.

4.4.3.1 *Evaluation*

The IAA was measured among the six annotators in order to evaluate the reliability of the annotation scheme. Since this task does not allow multiple correct choices, Fleiss' kappa (Fleiss, 1971) was a sensible choice (for more discussion on this see Section 2.2.1.1). Each definition in the list is considered as a category and the annotator's definition as a category, so in total there are four categories. Among the six annotators, the IAA averaged 0.272 for four categories on 1,301 annotated instances. This is a fair agreement based on Landis and Koch (1977) scale, despite the subjectivity of the task.

I was interested to analyse the best obtained IAA by varying the number of annotators depending on the majority of the annotated (non-skipped) instances. Therefore, I calculated the IAA between the best (top) five, four and three annotators, respectively, who tend to agree the most as shown in Table 4.15. From this analysis I observed that: 1) in case of the five annotators who agreed the most, the discarded annotator was the one who tend to choose the customised definition more often; 2) while in the case of the three annotators who agreed the most, the discarded two annotators were the ones who tend to choose the dictionary definition more often (as will be discussed in detail in Subsection 4.4.3.2). Having such versions of the dataset will allow the users to choose the subset that better suits their application. A higher quality dataset can be obtained from the instances which have majority vote over 60% with a moderate agreement strength of 0.48 in terms of Fleiss' kappa.

Table 4.15: Metaphor interpretation dataset analysis based on the agreement strength in terms of Fleiss' kappa per number of annotators.

	Annotated Instances	**Fleiss' Kappa**	**Agreement Strength**
Top three annotators	1,353	0.436	Moderate
Top four annotators	1,352	0.425	Moderate
Top five annotators	1,304	0.386	Fair
All six annotators	1,282	0.27	Fair
high quality subset with majority vote >60% (six annotators)	**676**	**0.48**	**Moderate**

4.4.3.2 *Analysis*

Definition Choice: In 70.82% of the cases, the annotators preferred to choose a definition from the suggested ones. On the other hand, they opt to provide their own definition of the

metaphoric expression either in the cases of encountering uncommon usage of the verb in a metaphoric way such as *"wash off all your sadness"*, *"open your heart"* and *"bring cheers"* or if the suggested definitions from the dictionary do not accurately reflect the metaphoricity of the expression such as *"take a stand"*, *"make a conscious effort"* and *"reduce anxiety"*. Figure 4.4 illustrates the percentage of choosing to provide an interpretation for each annotator. One of the annotators always preferred to write his own interpretations (definitions) of the metaphoric expressions; he provided an interpretation for 88.16% of the instances.

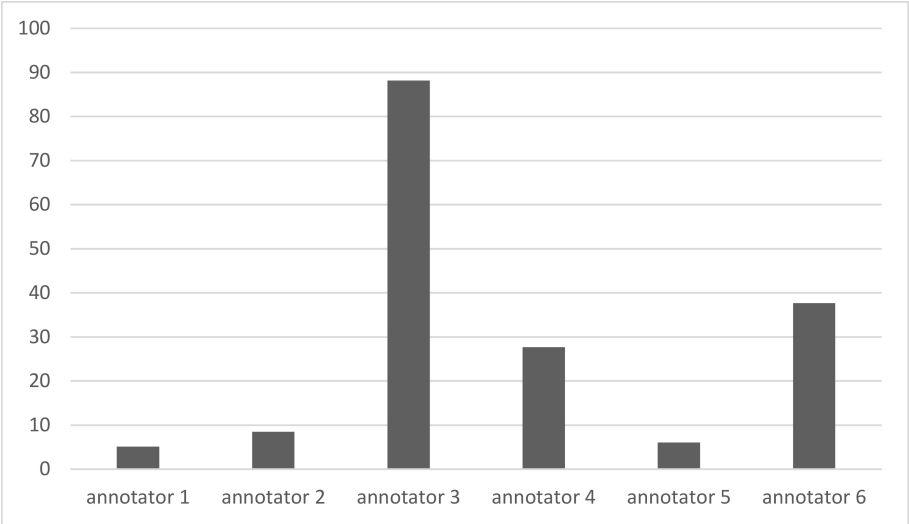


Figure 4.4: Percentage of providing a customised interpretation (definition) per annotator.

Points of (Dis-)agreements: Almost half of the provided annotations have a majority vote greater than 60% which yields a moderate IAA of 0.48 in terms of Fleiss kappa. The majority of disagreements centred around whether the suggested definition in the dictionary is enough to represent the metaphoric sense of the expression or not. Tables 4.16 and 4.17 shows examples of the agreements and disagreements between the six annotators. For example, the six annotators agreed that the suitable definition for the metaphoric expression *"release pain"* is the one from the Oxford Learner's Dictionary as shown in Table 4.16 whereas they opt for providing their own definition for the metaphoric expression *"brushing up my german"*. Table 4.18 gives more information about the statistics of the annotated dataset.

The effect of tweets: Although the context where the metaphoric expression appears is important to understand the expression, the noisy ungrammatical text of the tweets affected the annotation process. I observed that two annotators find it difficult to understand around 50 tweets, therefore, they skipped them which affected the overall agreement. The rest of the annotators did not skip them but they provided some notes about them. According to the annotators the reasons behind skipping these tweets were: 1) they do not understand the topic of the tweet at all (sarcasm, science fiction or games); 2) there is not enough information about the noun to give a definition; 3) the tweet is not grammatically correct to convey a meaning.

Annotators' Experience: Some of the annotators raised the issue of using metaphors while defining a metaphor. The annotators had to make sure not to use metaphors when writing their own definitions, which they found difficult. For instance, one annotator encountered the metaphoric expression *"stand a chance"* and she wanted to write *"to take/have an opportunity"*

Table 4.16: Examples of agreements among all annotators (100% majority vote) to interpret the highlighted metaphoric expressions.

Metaphoric Expression	Definition	Source
repay the tremendous support	to give something to somebody or do something for them in return for something that they have done for you	Oxford
release old emotional pain	to express feelings such as anger or worry in order to get rid of them	
ruin all the fun	to damage something so badly that it loses all its value, pleasure, etc.; to spoil something	
brushing up my german	to improve on something that one used to excell at	annotator provided
defeating brexit	to defeat the opposing group, argument, party etc.	
ramp up production	to increase the rate of production of somethings	

Table 4.17: Examples of disagreements among all annotators (less than 60% majority vote) to interpret the highlighted metaphoric expressions.

Metaphoric Expression	Definition	Source
take control	to capture a place or person; to get control of something	Oxford
take a minute	to need or require a particular amount of time	
finds fear	to have a particular feeling or opinion about something	
checked out this new friend	to look at information showing or pictures of a new supporter	annotator provided
wash off all your sadness	to stop feeling a particular emotion	
brings cheers	to make someone/group of people to feel a certain emotion	

Table 4.18: Statistics of the annotated dataset of metaphor interpretation.

Aspect	Value
total # of tweets	1394
# of unique (lemmatised) verb-direct objects	1394
average tweet length	22.36
# of skipped (non-understandable) tweets by all annotators	5
maximum # of skipped tweets by one annotator	50
minimum # of skipped tweets by one annotator	1
total # of annotated instances by six annotators	1,301
total # of annotated instances by three annotators	1,353
maximum # of instances with annotator's provided definition	1,147
% of instances annotated with agreement majority vote greater than 60%	52.02%
% of instances annotated with agreement majority vote less than 60%	47.9%
% of customised definitions by all annotators	29.2%

which is another metaphor; therefore she had to think of another definition using literal words. The majority of annotators agreed that sometimes using a metaphor is the easiest way to express what the author wants to say and here lies the difficulty of the metaphor interpretation task itself. It is worth mentioning also that the genre of the tweets affected the annotators' experience. Some annotators found many of the metaphoric expressions in the political tweets very straightforward and obvious, but when it came to emotional or motivational metaphors they found them slightly harder to define in simple terms.

4.4.4 Dataset Publication as Linked Data

I believe that this resource can be used to enrich lexical resources such as Wiktionary and WordNet by including a metaphor category similar to the idioms one. Therefore, in order to provide access to the data and promote reusability, the dataset is provided as a linked open dataset. As the original annotators chose the definitions from the provided suggestions obtained from the Oxford Learner's Dictionary, which is not possible to republish due to copyrights, the links will be instead provided by reference to the website. In particular, the sense *IDs* will be referred to as links and then the annotations are published in the Resource Description Framework (RDF) as linked data as shown in Figure 4.5.

```
<#anno1>
  <#tweetId> 746095727118532608;
  <#metaphor> "ignited a new passion"@en ;
  <#interpretation> [
    dc:source
    <https://www.oxfordlearnersdictionaries.com/definition/english/ignite_1#ignite_sng_1>;
    <#hash> "70B6783C04E770A02409174F97089E58";
    <#annotators> 2;
    <#majorityVote> 0.334;
    <#cosineSim> 0.520;
  ],
  [
    <#hash> "B501696811F1198BCFF3435E3822B571", "04BA979E7D9900B23321CE7318265E5F",
"433578FE1D3F6301616A61D732927B54", "EA9FA1DB0B8050D611715B92E7567B12";
    <#annotators> 4;
    <#majorityVote> 0.667;
    skos:definition "to cause something to happen or begin", "to make someone
start feeling a particular way", "made people more interested than ever", "to start
something/feelings"
  ]
```

Figure 4.5: Section of the metaphor interpretation dataset published as linked data.

In this case, a direct link is provided to the definition and a hash of the definition, which can be used to verify the definition has not changed. A script is provided with the download that fetches the definitions, verifies that they match the required hash and produces the results as comma-separated values. The customised definitions by the annotators will be provided as well. Since Twitter restricts the distribution of its content[24] (i.e. tweets text and metadata), the tweets text are not provided and only the *tweet-id* is shared for each annotated instance.

24 https://developer.twitter.com/en/developer-terms/policy

4.5 SUMMARY

This chapter is concerned with discussing the work done in this thesis to prepare the necessary resources for metaphor processing focusing on the identification and interpretation of linguistic metaphors on the relation level in English text. The first part of this chapter discussed creating a dataset of tweets annotated for metaphors. This dataset was a result of a proposed annotation methodology that was developed with the aim of reducing the cognitive load on the annotators and maintaining consistency. Although, the methodology is employed to annotate linguistic metaphors of the predicate type (i.e. verb-direct object pairs) in tweets, it can be applied to any text type, metaphor type or level of analysis. The tweets selection process was driven by achieving balance, sense coverage and verbs representativeness as well as high accuracy. The resulting metaphor dataset consists of various topic genres focusing on tweets of general topics and political tweets related to Brexit. A substantial inter-annotator agreement was achieved among five annotators, who are native speakers of English, despite the difficulty of defining metaphor, the conventionality of metaphors, and the noisy nature of the user-generated text of tweets. The dataset will be published online to facilitate research on metaphor processing in tweets.

This chapter also discussed how I adapted word-level benchmark dataset to suit relation-level metaphor identification. This step was essential towards filling the gap of the availability of large benchmark datasets for relation-level metaphor processing in English. A semi-automatic approach was employed to adapt the VUAMC to better suit identifying metaphors on the relation level avoiding the need for extensive manual annotation. The TroFi dataset, one of the earliest word-level datasets for metaphor identification of verbs, was also adapted to support verb-noun metaphor identification. Furthermore, the TSV dataset, which was originally annotated on the relation level focusing on adjective-noun relations, was extended by assigning context to its expressions from Twitter. This will encourage research in this area to work towards understanding metaphors in social media. As a result of this work, an adapted version of these benchmark datasets will be made publicly available according to the licensing type of each of them which will facilitate research on relation-level metaphor identification focusing on verb-direct object and adjective-noun relations.

Finally, this chapter discussed the work on creating the first gold standard dataset for metaphor interpretation along the more complex "definition generation" approach which provides full explanation of a given metaphoric expression. The methodology of preparing the dataset was demonstrated which combines an automatic retrieval approach with manual annotation to ensure wide coverage, accuracy and consistency. Lexical resources, word embeddings and semantic similarity were employed to assist in the annotation process with the aim to reduce the cognitive load on the annotators and to address the subjectivity of interpreting metaphoric expressions. As a result, around 1,500 metaphoric verb-direct object expressions in tweets were annotated. The methodology and annotation scheme can be generalised to annotate metaphors of any syntactic structure in any text genre/type. I believe that this dataset will be invaluable for the development and evaluation of approaches for metaphor interpretation. The full set of the annotated instances, including the annotators customised definitions, will be published as linked data in RDF format to promote reusability and to facilitate its incorporation into other lexicons such as Wiktionary and WordNet.

5 METAPHOR IDENTIFICATION IN TWEETS

"The questions I should like to see answered concern the "logical grammar" of "metaphor" and words having related meanings. It would be satisfactory to have convincing answers to the questions: "How do we recognize a case of metaphor?", "Are there any criteria for the detection of metaphors?" [...]"

(Black, 1962)

This chapter focuses on the second research theme in this thesis, which is *Metaphor Identification in Tweets*. My goal is to identify linguistic metaphors in English tweets by adopting the relational paradigm. I start my investigation of this theme by looking at distributional approaches to metaphor identification with the aim to design and employ a minimally supervised one to aid in the data annotation process (as discussed in the previous chapter). After that I turn my attention to the various features employed in the literature to identify metaphor in text. I therefore investigate meta-embedding learning methods in order to study the effectiveness of an ensemble of features to identify metaphoric expressions on the relation level. Finally, inspired by works in visual reasoning, I propose a novel approach for context-based textual classification that utilises affine transformations. I applied this approach that is based on contextual modulation to identify metaphoric expressions focusing on verb-noun and adjective-noun dependency relations in tweets.

The work presented in this chapter, under the aforementioned research theme, is formulated as three research questions, which were discussed in detail in Chapter 1, as follows:

RQ2.a To what extent can a minimally supervised approach based on distributional representation accurately identify metaphors in short texts?

RQ2.b Can employing an ensemble of linguistic and advanced contextual features to learn meta-embeddings in a neural architecture improve metaphor identification in tweets?

RQ2.c Can contextual modulation improve the performance for relation-level metaphor identification?

The chapter begins by introducing the motivation behind the proposed approaches for metaphor identification in this work. Section 5.2 presents the use of distributional word embeddings in the detection of verbal metaphors on the relation level. Then, Section 5.3 explains the methods for learning meta-embeddings that are explored in this work in order to identify verb-noun metaphoric expressions. Finally, Section 5.4 describes the work done to employ contextual modulation through affine transformations to identify metaphors on the relational level in tweets.

5.1 INTRODUCTION

Processing and understanding user-generated content on social media have attracted a growing attention over the past few years. Twitter, which is widely used by people to express their ideas and emotions, is considered a valuable resource for online conversational data. This has attracted many NLP applications to process this noisy, less informal and short text. Over the last couple of years, there has been an increasing interest towards metaphor processing and its applications, either as part of NLP tasks such as machine translation (Koglin and Cunha, 2019), text simplification (Wolska and Clausen, 2017a; Clausen and Nastase, 2019) and sentiment analysis (Rentoumi et al., 2012) or in more general discourse analysis use cases such as in analysing political discourse (Charteris-Black, 2011), financial reporting (Ho and Cheng, 2016) and health communication (Semino et al., 2018).

The main goal of this thesis is to process linguistic metaphors in tweets focusing on the identification and interpretation tasks. This chapter presents the proposed approaches to metaphor identification in this thesis. As discussed in Chapter 3, this problem can be addressed at different levels of granularity, namely at the sentence level (a sentence is metaphorical or not) or at the word level (a given word is used metaphorically or literally) or at the relational level (the grammatical/semantic relationship between a pair of words, such as a verb and its noun object or a noun and its adjectival modifier, is metaphoric or literal). In this thesis, I adopt the relational paradigm by modelling the interaction between the metaphor components (the *tenor* and the *vehicle*) in order to capture the metaphoricity in a way that mimics the human comprehension of metaphors.

The first part of the work presented in this chapter, focuses on exploring the use of distributional semantics in the identification of metaphors on the relational level. I propose an approach that makes use of distributional representations in a minimally supervised way to capture the metaphoricity of verb-noun expressions. My aim is to facilitate the ease of adaptation of the proposed approach to new text genres and thus it is designed to exploit a limited number of lexical resources and avoid the need to large annotated metaphor dataset. The question that I attempt to answer is that how effective this approach would be when dealing with relatively short text. Furthermore, as explained in Section 4.2, one of the main challenges that faces the computational modelling of metaphors in tweets is the lack of annotated datasets of tweets for metaphors on either the word or relation levels. The first step in this thesis is to create such a dataset by developing an annotation scheme that maintains accuracy and consistency. In order to reduce the cognitive load on the human annotators and enable quick and accurate annotation, I proposed the utilisation of this minimally supervised metaphor identification classifier to assist in the identification and the selection of the examples for annotation.

This work further investigates the features employed in the literature to identify metaphors in text. Recent works pertaining to metaphor identification experimented with a wide range of linguistic and advanced contextual features focusing on well-structured text. The key question is what are the effective features to consider while designing a new system and how to generalise feature selection beyond the level of metaphoric analysis and the text genre. I therefore study the employment of meta-embedding learning methods in a neural architecture with the aim to improve metaphor identification in tweets. This will allow the investigation of the effectiveness of each proposed feature by examining two strategies of learning meta-embeddings from multiple embedding sets, namely concatenation and dynamic meta-embeddings.

As discussed earlier in this thesis, the majority of existing approaches pertaining to metaphor identification adopt the word-level paradigm by treating the task as either *single-word classification* or *sequence labelling* without explicitly modelling the interaction between the metaphor components. Such explicitly marked relations facilitate other downstream tasks such as metaphor interpretation and cross-domain mappings. On the other hand, while existing relation-level approaches implicitly model this interaction, they ignore the context where the metaphor occurs. In the final part of the work presented in this chapter, I address these limitations by introducing a novel architecture for identifying relation-level metaphoric expressions of certain grammatical relations based on contextual modulation. In a methodology inspired by works in visual reasoning, I propose an approach based on conditioning the neural network computation on the deep contextualised features of the candidate expressions using feature-wise linear modulation. I demonstrate that the proposed architecture achieves state-of-the-art results on benchmark datasets. The proposed methodology is generic and could be applied to other textual classification problems that benefit from contextual interaction.

The following contributions are made in this chapter, which cover the second research theme of this thesis and seek answers to **RQ2.a**, **RQ2.b** and **RQ2.c**, as follows:

- Employing distributional semantics to introduce a semi-supervised approach to identify metaphors in text with the aim of aiding in the creation and annotation of a dataset of tweets annotated for linguistic metaphors on the relation level.
- Studying the effectiveness of an ensemble of features to identify linguistic metaphors of certain grammatical relations by utilising meta-embedding learning methods.
- Proposing a novel approach for context-based textual classification based on contextual modulation through affine transformation and apply it on relation-level metaphor identification.

Publications: Parts of this chapter have been published in Zayed et al. (2018, 2020b)

5.2 METAPHOR IDENTIFICATION USING DISTRIBUTIONAL SEMANTICS

Distributional semantics models have been widely used in a variety of NLP applications that focus on representing the meaning of a word such as word sense disambiguation (Navigli and Martelli, 2019), sarcasm detection (Ghosh et al., 2015b) and metaphor identification (Gutiérrez et al., 2016). Large corpora are utilised by these models to capture the relative meaning of words based on their distribution across different contexts. As discussed in Section 3.2, one of the prominent approaches to identify relation-level metaphors, focusing on verb-noun dependency relations, is the minimally supervised one in Shutova et al. (2010) that is based on distributional clustering. This approach was developed with the aim to facilitate the ease of adaptation to new domains and text genres. Therefore, inspired by this work, I introduce an approach that makes use of distributional representations in a minimally supervised way to capture the metaphoricity of verb-noun expressions. The performance of this approach is evaluated on relatively short text in order to provide an answer to **RQ2.a**. Given that this approach employs fewer lexical resources and does not rely on large annotated datasets of a specific domain or text genre it can be easily

used in the preparation and annotation process of a metaphor dataset of tweets. As discussed in Section 4.2, this approach was used as part of the proposed annotation scheme to facilitate the recognition and selection of the annotated instances.

The approach employs vector representations of word meanings and semantic similarity to classify a verb-noun expression in a given sentence for metaphoricity. The verb-noun expressions are highlighted using a dependency parser. A highlighted expression is classified as a metaphor by measuring its semantic similarity to a predefined small seed set of metaphors which acts as an existing known metaphors sample. Metaphoric classification is performed based on a previously calculated similarity threshold value on a development dataset. The following subsections discuss the proposed approach in detail. The use of different word embedding models will be investigated to identify verb-noun pairs where the verb is used metaphorically. Several experiments will be presented to show the performance of the proposed approach on benchmark datasets of relatively short text.

5.2.1 Distributional Semantics-based Proposed Approach

The idea behind this proposed approach is based on the distributional hypothesis (Harris, 1954; Firth, 1957) which states that *"words with similar meanings tend to have similar distributions in language"*. Word meanings can be represented as a distributed vectors (when using context-counting models) or word embeddings (when using context-predicting models). Then, the semantic similarity between two words, represented in the vector space, can be obtained using a similarity measure such as cosine similarity. With this idea in mind, therefore, a given metaphoric candidate should have common semantic features with some positive examples of metaphors[1]. However, simply calculating the similarity between a candidate verb-noun expression and a metaphoric seed is not enough due to the effect of each of the verb and the noun on the overall similarity score. For example, consider a metaphoric seed such as *"break agreement"* and two given candidates such as *"break promise"* and *"break glass"*. The semantic similarities between the word embeddings of the seed and the two candidates measured by the cosine similarity function are 0.5304 and 0.6376, respectively, using the pre-trained Word2Vec (Mikolov et al., 2013b) model on the Google News dataset. This indicates that both candidates are similar to the seed and there is not enough information to identify which one should be classified as a metaphor. Table 5.1 shows the similarity values of the two candidates and the most similar metaphoric seeds from the predefined seed set employed in this work. As shown, the two candidates obtained comparable consine similarity scores with the same seed set.

I decided to look into the individual words of the candidate considering the fact that semantically similar or related words will be placed near each other in the embeddings space while unrelated words will be far apart. Therefore, it is expected that the noun *"promise"* will be in the neighbourhood of *"agreement"* in the semantic space, while *"glass"* will not. So if both candidates share similar verbs, classification could be done based on the similarity of the nouns; in that case, *"break promise"* can be classified as metaphor due to the vicinity of its noun to the noun of the metaphoric seed while *"break glass"* will not. Since using one positive (metaphoric) example is not enough for precise classification, a small set of verb-noun pairs are used, hereafter referred to

1 This also resonates with the conceptual metaphor theory (CMT) where linguistic metaphors can be grouped under their conceptual metaphor.

Table 5.1: The cosine similarity between the candidates *"break promise"* and *"break glass"* and the top 10 metaphoric seeds in the seed set. Word embeddings are obtained using a Word2Vec model pre-trained on Google News dataset.

Candidate	Metaphoric Seed	Cosine Similarity	Candidate	Metaphoric Seed	Cosine Similarity
break promise	break agreement	0.6376	break glass	break agreement	0.5304
	hold back truth	0.4560		hold back truth	0.3435
	fix term	0.3653		frame question	0.3109
	spell out reason	0.3385		face hour	0.2949
	seize moment	0.3384		block out thought	0.2701
	glimpse duty	0.3224		seize moment	0.2677
	grasp term	0.3019		throw remark	0.2583
	frame question	0.2959		skim over question	0.2509
	accelerate change	0.2927		mend marriage	0.2375
	throw remark	0.2776		spell out reason	0.2354

as the seed set, where the verb is used metaphorically. The specification of the seed set will be explained in detail in Subsection 5.2.1.2.

5.2.1.1 *Technique*

I define a seed set of metaphoric verb-noun pairs as $S = \{(V, N)\}$. Given a target verb-noun candidate (v_t, n_t) that needs to be classified, the distance between every verb v_s in S and the verb of the candidate v_t is calculated using the cosine distance measure as follows:

$$D_{ts} = d(v_t, v_s) \;\; \forall v_s \in S \tag{5.1}$$

This gives a list of verbs ranked according to the distance to the verb of the candidate; the top n nearest verbs are then selected and the nouns associated with them in the seed set are obtained as follows:

$$Y_{vt} = top_n\{n_s : (v_s, n_s) \in S\} \;\text{ by }\; D_{ts} \tag{5.2}$$

Finally, the average of the distances between these nouns and the target noun in the candidate expression is calculated. If this average is less than a threshold δ then the candidate expression will be classified as a metaphoric expression as follows:

$$\frac{1}{|Y_{v_t}|} \sum_{n_s \in Y_{v_t}} [d(n_t, n_s)] \leq \delta \tag{5.3}$$

Table 5.2 shows the cosine distance between the verbs and the nouns of the candidates *"break promise"* and *"break glass"* versus the verbs and the nouns of the top 10 metaphoric seeds from the seed set using a pre-trained Word2Vec model on the Google News dataset; those 10 seeds have the most similar (nearest in terms of distance) verbs to the candidate verb. The table is sorted based on the cosine distance of the verb of the seed to the verb of the candidate expressions which is *"break"*. The top-10 expressions in this example are: *"break agreement, hold back truth, mend marriage, fix term, catch contagion, throw remark, seize moment, impose decision, impose control* and *frame question"*,

Table 5.2: The cosine distance between the verbs and nouns of the candidates *"break promise"* and *"break glass"* versus the verbs and the nouns of the top 10 metaphoric seeds in the seed set using a pre-trained Word2Vec model on the Google News dataset.

Candidate Verb	Seed's Verb	Cosine Distance	Candidate Noun	Seed's Noun	Cosine Distance	Candidate Noun	Seed's Noun	Cosine Distance
break	break	**0**	**promise**	agreement	**0.7479**	**glass**	agreement	1.0093
	hold back	0.6591		truth	0.7736		truth	**0.8872**
	mend	0.6935		marriage	0.9381		marriage	0.9419
	fix	0.6952		term	0.8085		term	1.0252
	catch	0.6966		contagion	1.0126		contagion	0.9089
	throw	0.7035		remark	0.8513		remark	0.9559
	seize	0.7201		moment	0.8556		moment	0.9510
	impose	0.7350		decision	0.8207		control	0.9506
	impose	0.7350		control	0.9107		decision	0.9987
	frame	0.7371		question	0.8462		question	0.9424
average distance					0.8565			0.957

5.2.1.2 *System Architecture*

As described in Figure 5.1, the proposed system consists of three main components, which are a dependency parser, a seed set of metaphors and pre-trained word embeddings, as follows:

Dependency Parser: This work focuses on identifying metaphors on the relation level, therefore the Stanford dependency parser (Chen and Manning, 2014) is used to extract the dependency relations in a given sentence. Specifically, the recurrent neural network (RNN) parser is used from the Stanford CoreNLP toolkit (Manning et al., 2014) to extract dependencies focusing on verb-subject and verb-direct object grammatical relations.

Seed Set: I used the seed set of Shutova et al. (2010) to act as the predefined set of existing known metaphoric expressions (positive examples). The seed set consists of 62 verb-subject and verb-direct object phrases where the verb is used metaphorically. These seeds were extracted originally from a subset of the BNC corpus which contains 761 sentences. The specified dependency relations were extracted from these sentences which are then filtered and manually annotated for metaphoricity (more details were given in Chapter 3). Examples of the metaphors in the seed set are *"break agreement, cast doubt, mend marriage,* and *stir excitement"*.

Word Embedding Model: This work utilises distributional vector representation of word meaning to calculate semantic similarity between a candidate and a seed set. Word2Vec (Mikolov et al., 2013a,b) and GloVe (Pennington et al., 2014) are two widely used context-predicting models for learning word embeddings from large corpora. In this work, I investigated the effect of using different pre-trained word embedding models and similarity measures as shown in detail in the next section.

5.2.2 Experiments

5.2.2.1 *Datasets*

Two different test sets are used to evaluate this approach; I briefly discuss them as follows:

The MOH-X dataset: As discussed in Section 3.2.2, the MOH-X dataset has relatively short sentences (an average sentence length of 11 words); therefore it will be very suitable to assess the

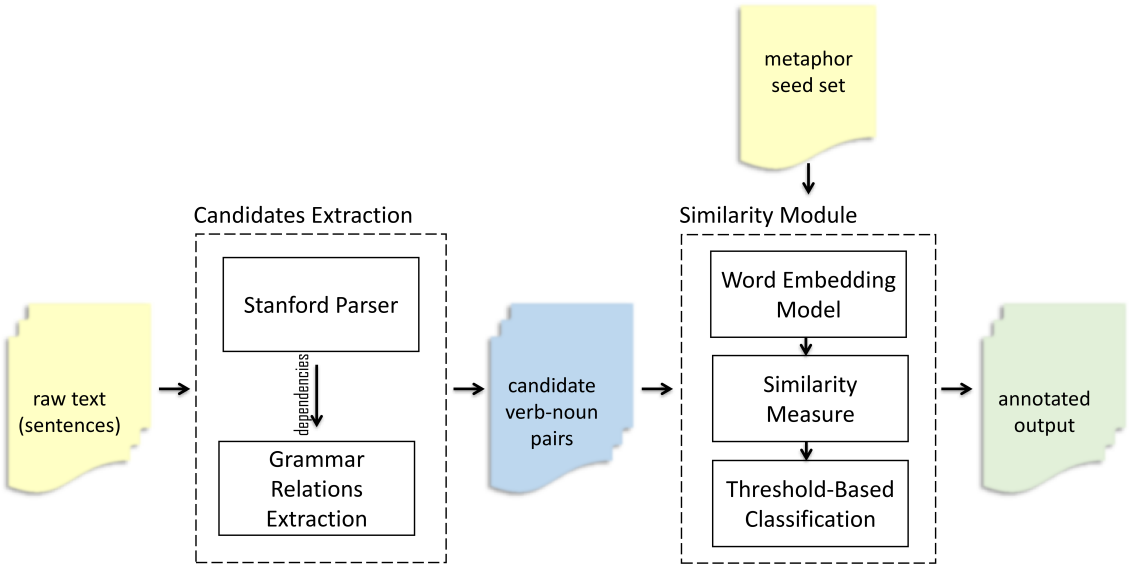


Figure 5.1: The proposed minimally supervised system to classify verb-noun expressions for metaphoricity based on distributional semantics.

performance of the proposed approach to deal with short text. The dataset, which is adapted from the original MOH dataset (Mohammad et al., 2016), comprises around 647 verb–noun pairs. The sentences were originally sampled from the example sentences of each verb in WordNet (Fellbaum, 1998).

The VUAMC Test Set: Around 300 verb-noun pairs from the VUAMC are used to test this approach. This subset[2] is drawn from the training split of the VUAMC (*Verbs* track) from the NAACL 2018 Metaphor Shared Task. I retrieved the original sentences of the annotated verbs in this subset from the VUAMC, which yielded around 8,000 sentences. I then parsed these sentences using the Stanford dependency parser and extracted 4,526 verb-direct object pairs. A balanced subset of these pairs is arbitrarily selected to form the test set. Table 5.3 shows some examples from this test set.

5.2.2.2 *Experimental Setup*

Similarity Measures: In order to calculate the similarity between two distributed representations, I examined two similarity measures as follows:

– Cosine Distance: The cosine similarity function measures the cosine of the angle between two vectors. Given the vectors u and v, the cosine similarity can be calculated as follows:

$$\cos(u, v) = \frac{u \cdot v}{||u||||v||} \tag{5.4}$$

Then, the cosine distance can be defined as:

$$cosine\_distance = 1 - \cos(u, v) \tag{5.5}$$

2 This study was done in the early stages of this thesis where I was investigating the usage of different datasets for metaphor identification. More particularly, I wanted to study the feasibility of employing the VUAMC to identify metaphors on the relation level and examine its limitations. Therefore, I employed a sample from it in this study.

Table 5.3: Examples from the VUAMC balanced test set of 300 verb-noun pairs.

Metaphor	Not Metaphor
reveal approach	collect passport
break corporation	use power
make money	abolish power
see language	perform shuffle
make error	decorate wall
face criticism	put stage
give access	read book
lay foundation	research joke
make time	tell story
abuse status	give key
accept reform	celebrate anniversary
capture power	need pilot
demonstrate danger	provide job

– Word Mover's Distance (WMD) (Kusner et al., 2015): It also captures the semantic similarity between two words. It could be defined as the minimum travelling distance from one embeddings vector to the other.

Word Embedding Models: Two pre-trained word embedding models are employed to obtain the vector representations of the candidate and seed set words as follows:

– Word2Vec: I used the 300-dimensional Word2Vec embeddings pre-trained on around 100 billion words from the Google News dataset[3]. The vectors cover around 3 million words.

– GloVe: I used the 300-dimensional GLoVe embeddings pre-trained on the Common Crawl dataset[4]. The vectors cover around 840 billion tokens of web data (about 2 million words).

For simplicity, a single vector representation is used for each word ignoring multi-word combinations such as phrasal verbs, examples of which include e.g. *"hold back* and *flip through"*.

System Parameters: I performed experiments on a development set to select the values of the parameters top_n and the threshold δ. The best value obtained for n is found to be 10 which means to average the similarity of the top 10 nearest verbs. The suitable distance average threshold δ is found to be 0.80 for the employed GloVe pre-trained embeddings and 0.85 for the employed Word2Vec pre-trained embeddings. These values give a good trade-off between false positives and false negatives.

5.2.2.3 *Distributional Clustering Baseline*

I consider the system introduced by Shutova et al. (2010) that is based on distributional clustering as a baseline. This system consists of four main components which are: a seed set, a clustering component, a candidate extraction component, and a filtering component. Extensive details of each component were given in Section 3.2.3. I re-implemented this system as closely as possible. However, I employed the Stanford dependency parser instead of the RASP parser to extract the dependency relations.

3 https://code.google.com/archive/p/word2vec/
4 https://nlp.stanford.edu/projects/glove/

5.2.2.4 *Results*

I conducted several experiments to asses the performance of the proposed approach that is based on distributional semantics, referred to as DistSemant. The results are compared to the distributional clustering-based baseline system Shutova et al. (2010). Tables 5.4 and 5.5 show the system performance on the MOH-X dataset and the VUAMC test set, respectively.

Table 5.4: Evaluation of the proposed distributional semantics-based model, *DistSemant*, on the MOH-X dataset of 647 verb-noun pairs and a performance comparison to the baseline system.

			Precision	Recall	F-score	Accuracy
The Distributional Clustering Baseline (Shutova et al., 2010)			**1.0000**	0.0095	0.0189	0.5148
The *DistSemant* approach	Word2Vec	WMD	0.5321	0.8413	0.6519	0.5599
		cosine distance	0.8727	0.1524	0.2595	0.5739
	GloVe	WMD	0.5243	**0.8571**	0.6506	0.5490
		cosine distance	0.6317	0.7460	**0.6841**	**0.6625**

Table 5.5: Evaluation of the proposed distributional semantics-based model, *DistSemant*, on the VUAMC test set of 300 verb-noun pairs and a performance comparison to the baseline system.

			Precision	Recall	F-score	Accuracy
The Distributional Clustering Baseline (Shutova et al., 2010)			**0.7500**	0.0197	0.0385	0.4915
The *DistSemant* approach	Word2Vec	WMD	0.556	0.8487	0.6719	0.5729
		cosine distance	0.7455	0.2697	0.3961	0.5763
	GloVe	WMD	0.5565	**0.9079**	0.6900	0.5797
		cosine distance	0.6377	0.8684	**0.7354**	**0.6780**

5.2.3 Discussion

Overall performance. The proposed approach performs better using GloVe pre-trained embeddings and the cosine distance as the similarity metric. It is noted that the system suffers from a low recall when using the Word2Vec pre-trained embeddings with the cosine distance function. I manually examined the system output on the employed datasets. The system was able to correctly detect conventionalised metaphoric expressions such as *"absorb knowledge, attack cancer, blur distinction, buy story, capture essence, swallow word, visit illness* and *wear smile"*. It also did a good job in figuring out literal ones such as *"attack village, build architect, leak container, steam ship* and *suck poison"*.

Error analysis. An error analysis is performed to determine the system flaws by inspecting the misclassified instances. Some of the false positives include *"ascend path, blur vision, buy love, communicate anxiety, jam mechanism, lighten room, line book* and *push crowd"* which could be argued as being used metaphorically depending on the context. The system was able to spot some inconsistency in the annotations of the VUAMC test set. For example, the verb *"win"* in the expression *"win election"*, which is classified as metaphor by the proposed system, has three different annotations across the rest of the VUAMC. It is annotated once as metaphor and twice as literal in very similar contexts. It is also annotated as metaphor with similar abstract concepts

such as the nouns *"match"* and *"bid* in *"win match"* and *"win bid"*, respectively. This resonates well with the later findings about the annotation inconsistency of the VUAMC as discussed in Section 4.3.2.

Baseline analysis. The results also indicate that the distributional clustering baseline has a very low recall on the employed test sets. The reason behind that is that it utilises clusters developed using the BNC corpus, which likely limit the coverage of the system. A candidate expression is either in the clusters or not. For example, if the candidate's verb appeared in a verb cluster but this cluster was not mapped to the cluster where the associated noun occurs the whole candidate expression will be discarded. As a result, out of the 300 pairs in the VUAMC test set only 7 candidates were included in the final classification as the rest of the words were not seen before in the clusters. Similarly, out of the 647 pairs in the MOH-X dataset only 4 were recognised by the baseline.

Comparisons and Limitations. One of the limitations of the proposed approach is that it has limited coverage when compared to other fully supervised approaches. This can be attributed to the relatively small size of the seed set (similar to the baseline). The proposed system performance will not be directly comparable to the state-of-the-art systems such as Shutova et al. (2016) and Rei et al. (2017) on the MOH-X dataset, since they used different test settings. Shutova et al. (2016) separated a random subset of around 87% of the MOH-X dataset for testing while Rei et al. (2017) used 10-fold cross-validation to evaluate their model. Therefore, Table 5.6 shows an approximate comparison to the best results obtained by utilising the GLoVe pre-trained embeddings and the cosine distance as similarity metric. Both systems are based on fully supervised approaches in which literal as well as metaphoric examples are employed to train their systems, whereas the proposed approach is minimally supervised (similar to Shutova et al. (2010)) which uses only the metaphoric examples. This limits the generalisation of the proposed approach to classify relatively new (not conventional) metaphors.

Table 5.6: Approximate comparison of the proposed distributional semantics approach with the state-of-the-art approaches Shutova et al. (2016) and Rei et al. (2017). The results are not strictly comparable due to different dataset experimental settings as discussed.

	MOH-X (all)			
	Prec.	**Recall**	**F1-score**	**Acc.**
Shutova et al. (2016) (multimodal)	0.65	**0.87**	**0.75**	-
Rei et al. (2017) (SSN)	**0.736**	0.761	0.742	0.748
The *DistSemant* Approach (GLoVE + Cosine)	0.6317	0.7460	0.6841	0.6625

The work introduced in this section is also related to Gutiérrez et al. (2016) who explicitly modelled metaphor in a compositional distributed semantic model and then employed this model to classify the metaphoricity of adjective-noun expressions where the adjective could be used metaphorically. However, Gutiérrez et al. employed a context-counting model to learn vector representations from a large corpus as discussed in Section 3.2.3. This approach requires a sizeable training dataset of metaphoric/literal examples which limits its extensibility.

Findings. From the experiments, this approach achieves better results when compared to other minimally supervised approaches. Although the results seem lower when compared with full supervised approaches, the aim of this work is not to outperform the state-of-the-art approaches as much as introducing an approach with a good performance which exploits a limited number

of lexical resources and does not rely on complex linguistic analysis or feature extraction from a large annotated corpus. This will facilitate the application of the system to new domains and text genres. As discussed earlier, the main goal of the work presented in this section is to introduce a minimally supervised approach that performs well on contexts with short text in order to be able to utilise it in the creation and the annotation process of a metaphor dataset of tweets (as shown in Section 4.2).

5.3 METAPHOR IDENTIFICATION USING META-EMBEDDINGS

Over the last few decades, a variety of approaches has been introduced to identify metaphors in text on different levels of processing. The common factor among these various approaches is that they employed a wide range of features as discussed in Chapter 3. These features include lexical, syntactic, semantic and psycholinguistic ones. Recently, researchers have been exploring more advanced features such as context (in-)dependent word embeddings, visual embeddings, property-based semantic word representations and document embeddings. Different architectures and experimental settings have been developed to utilise these features which makes it difficult to assess their effectiveness and scalability to identify metaphors. The key questions are to what extent can the available models scale to noisy and short informal contexts such as tweets, and what is the effectiveness of the introduced features in different settings? Therefore, I propose the investigation of these features under a unified neural architecture. My aim is to combine the traditional linguistic features that have been proven to be effective in identifying metaphors with the advanced deep contextualised embeddings through employing advanced ensembling learning methods. This investigation will seek an answer to **RQ2.b**.

The representation of word meaning through word embeddings has been proven useful in a lot of NLP application including metaphor identification (for more details see Section 3.2.3). There exists a variety of approaches to prepare pre-trained embeddings on different datasets such as Mikolov et al. (2013b) and Pennington et al. (2014). Creating an ensemble of different embedding sets leads to having an enhanced embeddings with better coverage of words (Yin and Schütze, 2016). This ensemble, which is referred to as meta-embeddings, does not require large text corpora during its learning process and it only exploits the pre-trained word vectors. There are a variety of methods for learning meta-embeddings including Yin and Schütze (2016); Coates and Bollegala (2018); Bollegala et al. (2018); Bao and Bollegala (2018) and Kiela et al. (2018). In this work I employ two methods, a straightforward one and a more advanced one, which are concatenation (Concat) and dynamic meta-embeddings (DME) (Kiela et al., 2018), respectively. The reason behind choosing DME is that it learns the meta-embeddings in a dynamic way which allow them to be applied directly on downstream tasks.

As discussed earlier in Chapter 3, context is essential when it comes to metaphor identification as it facilitates inferring the intended meaning of a word and disambiguate its sense. A lot of labelled data is needed to exploit such wider context to train deeper neural networks and tune their parameters. One of the main problems that faces the computational modelling of metaphors is the lack of large datasets. The majority of the available metaphor datasets are relatively small to train deep learning models (for more discussion on datasets limitations see Chapter 3). The work presented in this section attempts to bridge the gap between the current limitations of the small-data settings and the advantage of exploiting context to identify metaphors

by leveraging meta-embeddings. I develop an attention-based neural network architecture that combines the strengths of both the pre-trained context (in-)dependent word embeddings as well as traditional linguistic features to identify metaphors in tweets. Further, I investigate the effectiveness of employing an ensemble of state-of-the-art pre-trained embeddings, namely the deep contextualised representations BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018) as well as the context-independent GloVe embeddings in addition to other lexical and semantic features. As mentioned earlier, I experiment with two strategies of combining feature sets, namely concatenation and DME. The next sections explain the developed model and the selected features in depth. I present detailed experiments on the metaphor dataset of tweets, the ZayTW dataset, in addition to other benchmark datasets.

5.3.1 Meta-Embeddings-based Proposed Approach

My main goal is to study the effectiveness of various features on the task of metaphor identification under a unified neural architecture. I adopted a similar architecture to the neural one introduced by Gao et al. (2018). However, I formulated it using the neural network conceptual framework for NLP introduced by Honnibal (2016) which comprises four steps, namely embed, encode, attend and predict, hence it was given the name EEAP. Honnibal proposed that any general neural NLP pipeline can be constructed from these four basic operations. In this thesis, I used his abstraction on the use case of identifying relation-level metaphoric expressions when using neural models. The proposed sequence neural model, presented in this section, utilises pre-trained context (in-)dependent word embeddings on large datasets. This will allow the use of the generated knowledge in a setting where only little labelled data is available.

The main component in this architecture is a bidirectional long short term memory (BiLSTM) (Hochreiter and Schmidhuber, 1997) encoder which produces a sentence representation from the provided features. I experiment with multiple features, as will be explained in the next subsection, to evaluate the effectiveness of each of them for the task of metaphor identification. Then, I combine all representations as an ensemble of features to learn meta-embeddings by exploiting two ensembling strategies which are simple concatenation and dynamic meta-embeddings (DME). Furthermore, I explore the importance of employing an attention mechanism after the encoding process.

My idea is that the shared knowledge from various embeddings can be crucial to capture different levels of complex features that can, in turn, improve the metaphor identification task. In that essence, pre-trained context-independent word embeddings such as GloVe which encapsulate global syntactic and semantic features about words can be coupled with context-dependent word embeddings such as ELMo and BERT to capture richer semantic representation. Furthermore, I posit that including traditional linguistic features that are well-established in the literature in a neural architecture with other advanced features can further improve the model performance.

5.3.1.1 *Examined Features*

I intend in this work to examine the effectiveness of the following features for metaphor identification focusing on their scalability to identify metaphors in tweets. I plan to study the effect of each feature individually in a simple framework and then collectively to examine whether the

shared knowledge from various embeddings improves relation-level metaphor identification or not. The features under study are:

Global Vectors for Word Representation (GloVe): It is a widely used word embeddings algorithm introduced by Pennington et al. (2014) to construct a dense vector representation of a word. It has been employed by previous works to identify metaphor in text on the word level as discussed in Section 3.2.3. In the work presented in this section, I used the uncased 100-dimensional pre-trained GloVe embeddings trained on 2 billion tweets and contains 1.2 million words.

Embeddings from Language Models (ELMo): Modelling deep contextualised word representations from language models was introduced by Peters et al. (2018). The representations from these models are obtained from deep bidirectional language models that are based on LSTM networks to learn context-dependent aspects of word meanings along with syntactic aspects. This results in better representations of a word depending on its context. The usage of ELMo has been first introduced to improve the identification of word-level metaphors by Gao et al. (2018). I used the 1,024-dimensional pre-trained ELMo embeddings trained on the One Billion Word benchmark corpus (Chelba et al., 2014).

Bidirectional Encoder Representations from Transformers (BERT): Devlin et al. (2018) took the idea of ELMo further by using the recent attention transformer architecture (Vaswani et al., 2017) to encode context. As discussed in Section 3.2.3, Mao et al. (2019) introduced the use of BERT in an RNN-based architecture to show the effectiveness of employing such advanced context-dependent embeddings on metaphor identification. A lot of work followed in their foot steps after that. I used the uncased 768-dimensional embeddings that were pre-trained on Wikipedia and the BooksCorpus (Zhu et al., 2015).

Index embeddings. As discussed in Chapter 3, several previous models to metaphor identification adopted the word-level paradigm and treated the task as *single-word classification*. These models have to highlight the word being examined. Gao et al. (2018) proposed the idea of using a binary index to perform this in which a low dimension representation of a given sentence length is used to give the model a sense of position awareness. A one-hot vector is used where 1 is assigned in the position of the targeted word that needs to be classified for metaphoricity and 0 is assigned otherwise. Mu et al. (2019) took this idea a step forward by plugging this binary vector in a trainable embedding layer. I adopted a similar approach in this work to study the effectiveness of this feature for relation-level metaphor identification. Given a targeted expression of a grammatical relation such as verb-noun or adjective-noun, I highlight both the verb/adjective and the noun together using a binary representation (a one-hot vector) which is then plugged in a trainable embedding layer.

Concreteness. Conceptual features such as concreteness, abstractness, affectiveness and imageability are employed by machine learning based approaches to identify metaphors on the relation level (as discussed in Chapter 3). However, to the best of my knowledge, investigating their effectiveness under an advanced neural architecture has not been explored. In this work, I study the effect of employing the level of lexical concreteness or abstractness to improve metaphor identification under the proposed neural architecture. Therefore, I experimented with the concreteness ratings obtained from Brysbaert et al. (2014).

Attribute features. Bulat et al. (2017) experimented with attribute-based semantic representations obtained from the property-norm dataset (McRae et al., 2005). The authors concluded that attribute-based embeddings can outperform the context independent word embeddings from

pre-trained models such as Word2Vec (Mikolov et al., 2013b). However, this was not tested in a neural architecture. Therefore, I experimented with 2,526-dimensional attribute vectors to study whether they can still improve the metaphor identification task under this setting or deep contextualised vectors can supersede them.

WordNet Supersenses. The supersense of a word, obtained from lexicographer file names in WordNet (Fellbaum, 1998), can be a useful coarse-grained representation of word meaning. Tsvetkov et al. (2013, 2014) showed that such a feature, referred to as coarse semantic feature, can capture high-level properties of concepts which help in identifying metaphoricity. I investigated the effectiveness of this feature by incorporating a 45-dimensional vector that represents the supersenses of a given word in the proposed recurrent neural model.

5.3.1.2 *System Architecture*

As discussed earlier, the proposed architecture, depicted in Figure 5.2, goes through four main steps under the EEAP formulation of Honnibal (2016) which was discussed earlier. The system takes as an input the raw text of a sentence (or a tweet) which has a highlighted expression[5] such as a verb-noun that needs to be classified as metaphor or literal by following these steps:

Embed: Given a labelled dataset of sentences, the model begins by embedding the tokenised sentence S into vector representations depending on the chosen embedding method. Since the intent is to give the model access to multiple types of features/embeddings as discussed in the previous subsection, this step is followed by selecting the strategy of combining these features to learn meta-embeddings. I, therefore, examined two ensembling strategies which are:

- **Simple concatenation (Concat):** Feature concatenation is a simple yet reliable technique that has been used in the literature for combining different embedding types which results in a larger representation covering shared knowledge. It can be represented mathematically as:

$$E_i^{Concat} = [e_{1,i}, e_{2,i}, ..., e_{n,i}] \tag{5.6}$$

  where i is the i^{th} word in S and n is the number of embedding types.
- **Dynamic meta-embeddings (DME):** Kiela et al. (2018) introduced this technique for improving sentence representations. This is a supervised learning method of embedding ensembles by which a network can dynamically learn which embedding type has stronger effect. The idea is to predict a weight for each embedding type by projecting it into a common dimensional space (as described in Equation 5.7) then combine the projected embedding types by taking the weighted sum (as described in Equation 5.8). This is done through a module that is appended at the beginning of the computational neural model of the downstream task. The method utilises a self-attention mechanism 5.9; therefore, it can be either context-independent or context-dependent. I experimented with both approaches.

5 The highlighted expression will be given as an input to the system in terms of the indices of its tokens in the given sentence.

$$E'_{i,j} = g(W_j E_{i,j}) \tag{5.7}$$

$$E_i^{DME} = \sum_{j=1}^{n} \alpha_{i,j} E'_{i,j} \tag{5.8}$$

where $j = (1, 2, ..., n)$ are the embedding types, W are trainable weights and α are the weights from the self-attention mechanism which is obtained as follows:

$$\alpha_{i,j} = g(E'_{i,j}) = \phi(aE'_{i,j} + b) \tag{5.9}$$

where a is a randomly initialised weight matrix and b is a bias vector; ϕ could be either a softmax, sigmoid or tanh.

Employing an ensemble of features can be considered as combining different resolutions learned from the multiple pre-trained embedding models. Figure 5.3 shows the DME method in detail.

Encode: The next step is to train a neural network with the obtained representations. Since context is important for identifying metaphoricity, a sentence encoder is a sensible choice. I use bidirectional LSTM to obtain a contextual representation (sentence matrix) which summarises the syntactic and semantic features of the whole sentence on different levels.

Attend: After encoding the sentence, all the words would not contribute equally to the final sentence representation, thus, the next step would be reducing the obtained sentence matrix to a sentence vector by selecting important features. Recently, attention mechanisms have become a useful method to select the most important elements while minimising information loss. I employ the attention mechanism presented in Lin et al. (2017). It takes the output from the sentence encoder as an input (i.e the sentence matrix H that corresponds to the output of the bidirectional LSTM layer), as well as a randomly initialised weight matrix W, a bias vector b and a context vector u to produce the attended output as follows:

$$e_i = tanh(WH_i + b) \tag{5.10}$$

$$\alpha_i = softmax(ue_i) \tag{5.11}$$

$$o = \sum_{i=1}^{S} \alpha_i H_i \tag{5.12}$$

Predict: The last step is to make the final prediction using the attended output from the previous step. I use a fully-connected feed-forward layer with a sigmoid activation that returns a single (binary) class label to identify whether the targeted expression is metaphoric or not.

5.3.2 Experiments

5.3.2.1 *Datasets*

This system requires labelled data where a targeted expression of a specific grammatical relation is highlighted as explained in the previous section. Therefore, a benchmark relation-level metaphor

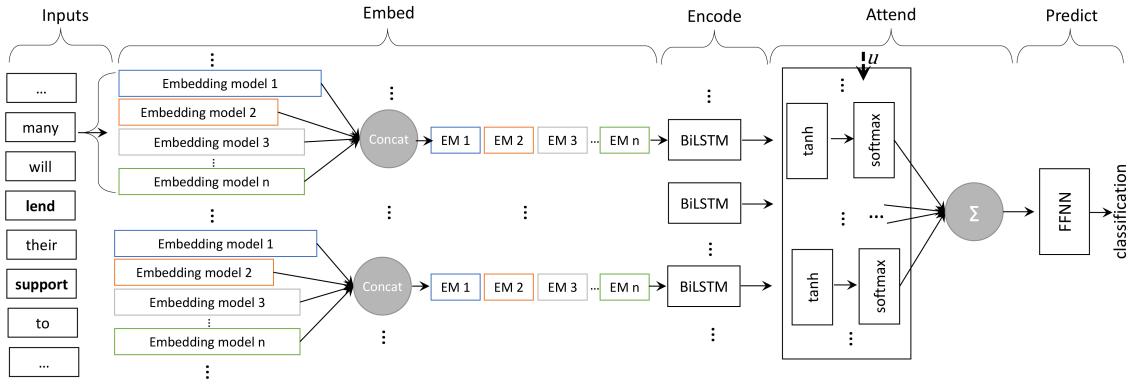


Figure 5.2: The proposed approach for metaphor identification based on meta-embedding learning methods. Concatenation is depicted as a way to create features ensemble. The framework is formulated into four steps: embed, encode, attend and predict. The output of the *"Concat"* step can be considered as a horizontal stacking of vectors, preserving the dimensions of the original embeddings.

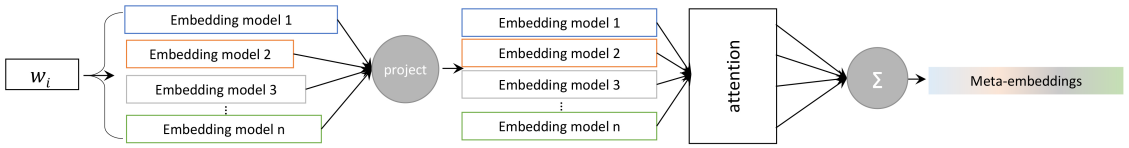


Figure 5.3: Illustration of the dynamic meta-embeddings (DME) technique that allows the model to choose the important embedding types automatically. The output of the *"project"* step can be considered as a vertical stacking of features after unifying the dimensions of the original embeddings in the new dimensional space.

dataset is employed to train and evaluate the proposed approach, namely the MOH-X dataset. Since this dataset is relatively small to train a neural model, I employed the ZayTw dataset that was prepared as part of this thesis to train the system and evaluate the effectiveness of each feature to identify metaphors in such informal context. I briefly discuss the properties of each dataset as follows:

The MOH-X dataset: As discussed in Section 3.2.2, the MOH-X Dataset consists of 647 verb-noun pairs which are extracted by Shutova et al. (2016) from the original word-level MOH dataset. This dataset contains short and simple sentences that are originally sampled from the example sentences of each verb in WordNet. Since this dataset has a relatively small size, it is often used in a 10-fold cross-validation settings.

The ZayTw Dataset: As explained in Section 4.2, the ZayTw dataset is introduced as part of this thesis to identify metaphoric expressions of verb-noun dependency relations in tweets. The dataset comprises around 2,500 tweets annotated for metaphor focusing on emotional tweets of general topics as well as political tweets related to Brexit. I divided this dataset randomly into training, development (validation) and test sets of 70%, 10% and 20% splits, respectively.

5.3.2.2 *Experimental Setup*

The proposed model utilises a bidirectional single-layer LSTM encoder with 100 hidden units. Several experiments are carried out on the ZayTw dataset to study the effectiveness of different features to identify metaphors on the relation level. All the hyper-parameters were optimised on

the randomly separated validation set. The Adam algorithm (Kingma and Ba, 2015) is used for optimisation during the training phase and binary cross-entropy as a loss function to fine tune the network. L_2-regularisation weight of 0.01 is used to reduce overfitting. I experimented with various batch sizes and number of epochs. The reported results are obtained using batch size of 10 instances for the MOH-X dataset and 30 instances for the ZayTw dataset. In all experiments, the input sentences are zero-padded to the longest sentence length in the dataset. For the concatenation of the different embedding types, I experimented with both context-dependent and context-independent DME. Moreover, I did various experiments to choose the new dimension of the projected embeddings, the best reported results are using an average of the dimensions of all the embedding types. The models are implemented using Keras (Chollet, 2015) with the TensorFlow backend. All experiments are done on a NVIDIA Quadro M2000M GPU of of 4GB memory and the average running time for the proposed models is around two hours for maximum of 100 epochs.

5.3.2.3 *Results*

The main goal of the work presented in this section is to study the effectiveness of combining the state-of-the-art features for metaphor identification in a neural architecture; therefore, I compared the performance of each feature individually. I then combined the features using the two aforementioned strategies of learning meta-embeddings. Table 5.7 shows the obtained results in terms of the micro-averaged precision, recall, F1-score and accuracy. All the results presented in this table are obtained after running the experiments three times with random seeds and taking the average; variance is within the range of 0.005-0.01.

I further compared the performance of the proposed model using the best performing features, which are the GLoVe, ELMo and index embeddings in addition to the attribute features, to the current work that addresses the task on the relation level on the benchmark dataset MOH-X. The selected approaches for comparisons are: the multimodal system of linguistic and visual features by Shutova et al. (2016) and the supervised similarity network (SSN) by Rei et al. (2017). I consider the SSN system as a baseline. It is worth mentioning that Shutova's (2016) results on the MOH-X dataset are not strictly comparable as they used different experimental settings where around 87% of the dataset were randomly separated and used for testing. Since the source code of Rei's (2017) system is available online[6], I trained and tested their model using the ZayTw dataset. Table 5.8 shows the obtained results.

5.3.3 Discussion

Features effectiveness and scalability. As shown in the previous subsection, several experiments were done to investigate the effectiveness of the features under study to identify metaphoric expressions on the relation level in tweets using a neural sequence model. In order to statistically interpret the significance of the obtained results, the one-tailed paired *t-test* (Yeh, 2000) at $p\text{-}value<0.05$ is carried out to compare each feature set against the GLoVe embeddings as a baseline feature. Table 5.7 highlights the most significant set of features for the models based on concatenation and the DME learning methods, referred to as *"Concat"* and *"context-dep DME"* models, respectively. As concluded, employing the index embeddings boost the performance

6 https://github.com/marekrei/ssn

Table 5.7: Effectiveness of the different feature sets under study using the concatenation and the DME learning strategies on the ZayTw dataset. (WN=WordNet; Concat.=the concatenation method of learning meta-embeddings; context-dep=context-dependent attention in the DME) *The best performing features based on the statistical significance test with a p-value<0.05 compared to GloVe embeddings as a baseline feature.

		Concat.				DME (context-dep)			
	Features	**Prec.**	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**
	GloVe	0.633	0.635	0.634	0.602	0.673	0.610	0.640	0.627
	ELMo	0.668	0.668	0.668	0.639	0.656	0.698	0.676	0.637
	BERT*	0.670	0.697	0.683	0.649	0.640	0.675	0.657	0.618
	GloVe+ELMo*	0.689	0.689	0.689	0.663	0.693	0.668	0.680	0.659
advanced	GloVe+BERT	0.665	0.646	0.656	0.631	0.660	0.632	0.646	0.623
	GloVe+ELMo+BERT	0.669	0.664	0.667	0.639	0.658	0.675	0.667	0.633
	GloVe+ELMo+Index (GEI)*	**0.785**	**0.740**	**0.762**	**0.749**	0.688	**0.726**	**0.706**	0.672
	GloVe+ELMo+BERT+Index (GEBI)*	**0.759**	**0.751**	**0.755**	**0.735**	**0.723**	0.687	**0.704**	0.687
traditional	attribute+concrete+WN Supersenses	0.608	0.628	0.618	0.578	0.624	0.588	0.606	0.584
	traditional+GloVe	0.659	0.671	0.665	0.633	0.690	**0.700**	0.695	0.667
	traditional+GEI*	**0.801**	**0.711**	**0.753**	**0.747**	**0.724**	**0.700**	**0.711**	0.692
advanced +	traditional+GEBI*	**0.765**	**0.751**	**0.758**	**0.739**	0.682	0.675	0.678	0.653
traditional	GEBI+attribute*	**0.776**	**0.765**	**0.771**	**0.753**	**0.699**	0.679	0.689	0.667
	GEBI+concrete*	**0.775**	**0.733**	**0.753**	**0.739**	**0.719**	0.693	**0.706**	0.686
	GEBI+WN Supersenses*	**0.754**	**0.765**	**0.759**	**0.737**	**0.720**	**0.707**	**0.714**	0.692

Table 5.8: Results of the proposed models using the concatenation and the DME learning methods compared to the state-of-the-art approaches on the MOH-X benchmark dataset and the ZayTw metaphor dataset of tweets. The results of Shutova et al. (2016) are not directly comparable due to the difference in experimental settings.

	Tweets (test-set)				MOH-X (10-fold)			
	Prec.	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**
Shutova et al. (2016) (multimodal)	-	-	-	-	0.65	**0.87**	0.75	-
Rei et al. (2017) (SSN)	0.543	**1.0**	0.704	0.543	0.736	0.761	0.742	0.748
The proposed DME model (context indep.)	0.703	0.676	0.689	0.667	0.757	0.765	0.757	0.758
The proposed DME model (context dep.)	0.720	0.707	0.714	0.692	0.776	0.751	0.761	0.765
The proposed Concat model	**0.776**	**0.765**	**0.771**	**0.753**	**0.802**	0.772	**0.784**	**0.789**

of the model dramatically as it gives extra weight to the parts of the expression under analysis. Although linguistic features alone did not improve the model performance they performed quite well when combined with the advanced features GloVe, ELMo and the index embeddings where the highest precision is achieved. Furthermore, for the concatenation method, the best F1-score is obtained by combining the attribute-based semantic representations with the advanced features (GloVe, ELMo, BERT and the index embeddings, hereafter GEBI). For DME, combining the coarse semantic features of WordNet supersenses with the GEBI embedding set improved the model performance.

Error analysis. An error analysis is performed to determine the system flaws. I analysed the predicted classification of the best performing features under the *"Concat"* model. It was found that there were 126 incorrectly classified instances among which ∼48% were false positives and ∼51% were false negatives. Table 5.9 lists some examples of the misclassified instances from the ZayTw dataset. Some of the misclassified instances contain verbs with weak metaphoric potential where the original annotated instances had an agreement majority vote of 60%. Other examples lack enough context to fully determine the metaphoricity of the expression with some authors being deliberately vague. I also noticed that some of the misclassified tweets contain too many Twitter properties such as mentions and hashtags with little content to help in the identification.

The effect of the dataset size. The size of the employed datasets has a great effect on the model performance as discussed earlier. Although the MOH-X dataset provided a good testing ground for verb-noun metaphor identification approaches, I noticed that it might not precisely demonstrate the differences in the performance of the examined systems. Due to its relatively small size, any change in a single annotated instance drastically affects the results.

The performance of the pre-trained BERT embeddings. Interestingly, employing the pre-trained embeddings of BERT did not seem to improve the system performance under this network architecture as I expected. The BERT embeddings achieved good performance on downstream tasks by fine-tuning the pre-trained BERT architecture on the new training data of size 5K to 100K samples in order to generalise well (Devlin et al., 2018). Since the training dataset employed in this work is smaller than this range, I could not fine-tune BERT and I rather used the pre-trained embeddings out-of-the-box which did not improve the results of identifying metaphors under this neural architecture.

Concatenation versus DME. The proposed DME model scored lower results than the simple concatenation likely due to the fact that it requires more training data. Concatenation preserves the full information enclosed by each embedding type which helps when training the network on smaller datasets. However, the relatively large dimension of the embeddings after concatenation means a lot of parameters to tune for the neural network. DME helped the network to dynamically choose which embeddings are more important to the task based on the weights that the model learned and assigned to each one. In an attempt to visually interpret these weights. Figure 5.4 shows a heat-map of the attention weights for the context-dependent DME trained on the ZayTw dataset using the set of features/embeddings under study. As observed, the model learns different resolutions from each feature. In this example, ELMo and BERT embeddings emphasised the importance of the noun *"ego"* in determining the metaphoricity of the verb *"break"* while the semantic embeddings gave it a lower weight.

Limitations. One limitation of this proposed approach is that it is hard to understand and interpret the behaviour of the index embeddings. Its role is to capture where in the given sentence

the focus words (the verb and the associated noun) are. Theoretically, it could be seen as a naive attention method to emphasise the targeted verb/noun which led to a significant improvement over purely using ELMo and BERT embeddings, improving the F1-score from 69% to 76%. It is hard to deduce the reason behind this performance which might be likely due to overfitting. Experiments on a larger dataset are required to confirm this observation. Moreover, another question is whether such a naive way of highlighting the targeted words is enough to emphasise the relation between the metaphor component or a more sophisticated way is required. I will attempt to answer this question in the next section.

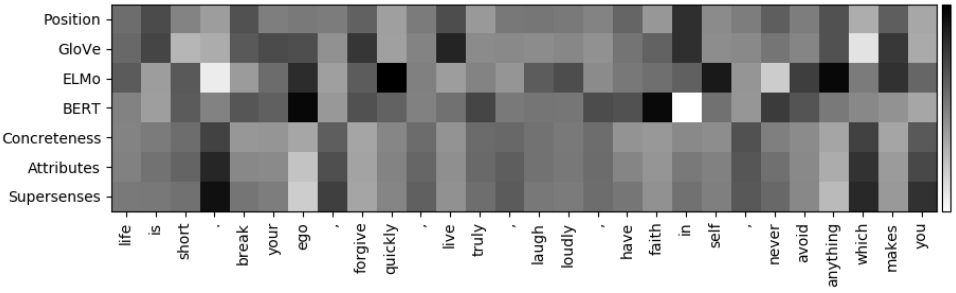


Figure 5.4: Visualisation of the attention weights for the context-dependent DME model of a tweet from the ZayTw dataset. The targeted expression is the verb-noun pair "break ego". The vertical axis shows the features/embeddings under study. A darker shade means higher weight.

Table 5.9: Examples of the misclassified tweets using the best performing features under the "Concat" model on the ZayTw dataset showing the classification probability.

	Tweet	**Probability**
Flase Negative	...#ivoted with a black pen. do not **trust pencils**. easy to rub out...	0.0
	do you **want the perfect smile makeover**? thinking about getting veneers?...	0.0
	prefer to remain but still not voted? exactly what nigel is banking on! go and do it! and **round up your friends**! #voteremain	0.001
	lost myself trying not to **lose someone** else	0.005
	important day, which will **determine our future** and our children's future #brexit	0.418
False Positive	if you do not **teach british history** in british schools, do not be surprised...	0.024
	rains **prevent my attendance** to aldeburgh. rains enable my active support of #voteremain...	0.553
	@imfactstory **change your name** to spam	0.998
	shocking that this guy would **spend so much political capital** trying to erase the first black president.	0.998
	...this was an internal tory party squabble that the country was dragged into to **placate exiters**	0.991

1 https://en.wiktionary.org/wiki/break
2 last sense: http://bit.ly/2TlWxyx
3 http://bit.ly/2x7SYUb
4 http://bit.ly/3aoa12v
5 http://bit.ly/2wtmtiI
6 http://bit.ly/32Kg7HS

5.4 CONTEXTUAL MODULATION

As discussed so far, identifying metaphors in text is very challenging and requires comprehending the underlying comparison between the metaphor component. Various levels of processing (paradigms) have been established in the literature and adopted by previous research in this area to identify metaphors in text as explained earlier. I briefly recap them here along with some related background in order to introduce the motivation behind the proposed approach in this section. Identifying metaphors on the word level could be treated as either *sequence labelling* by deciding the metaphoricity of each word in a sentence given the context or *single-word classification* by deciding the metaphoricity of a targeted word. On the other hand, relation-level identification looks at specific grammatical relations such as the *dobj* or *amod* dependencies and checks the metaphoricity of the verb or the adjective given its association with the noun. In relation-level identification, both the source and target domain words (the *tenor* and *vehicle*) are classified either as a metaphoric or literal expression, whereas in word-level identification only the source domain word(s) (the *vehicle*) are labelled. The majority of existing approaches, as well as the available datasets, pertaining to metaphor processing focus on the metaphorical usage of verbs and adjectives either on the word or relation levels. This is because these syntactic types exhibit metaphoricity more frequently than others according to corpus-based analysis (Cameron, 2003; Shutova and Teufel, 2010) (for detailed discussion see Chapter 2).

Although the main focus of both the relation-level and word-level metaphor identification is discerning the metaphoricity of the *vehicle* (source domain words), the interaction between the metaphor components is less explicit in word-level analysis either when treating the task as *sequence labelling* or *single-word classification*. Relation-level analysis could be viewed as a deeper level analysis that captures information that is not captured on the word level through modelling the influence of the *tenor* (e.g.noun) on the *vehicle* (e.g. verb/adjective). There will be reasons that some downstream tasks would prefer to have such information (i.e. explicitly marked relations), among these tasks are metaphor interpretation and cross-domain mappings (for an extensive discussion of this see Section 3.2.1). Moreover, employing the wider context around the expression is essential to improve the identification process as showed in the literature. This thesis focuses on relation-level metaphor identification represented by specific dependency relations such as verb-noun and adjective-noun.

The work presented in this section seeks an answer to **RQ2.c** in order to address the interaction between the metaphor components and to exploit wider context. I therefore propose a novel approach for context-based textual classification that utilises affine transformations. In order to integrate the interaction of the metaphor components in the identification process, I utilise affine transformation in a novel way to condition the neural network computation on the contextualised features of the given expression. The idea of affine transformations has been used in NLP-related tasks such as visual question-answering (de Vries et al., 2017), dependency parsing (Dozat and Manning, 2017), semantic role labelling (Cai et al., 2018), coreference resolution (Zhang et al., 2018), visual reasoning (Perez et al., 2018) and lexicon features integration (Margatina et al., 2019).

Inspired by the works on visual reasoning, the candidate expression of certain grammatical relations, represented by deep contextualised features, is used as an auxiliary input to modulate the proposed computational model. Affine transformations can be utilised to process one source of information in the context of another. In this case, I want to integrate: 1) the deep contextualised-features of the candidate expression (represented by ELMo sentence embeddings) with 2) the

syntactic/semantic features of a given sentence. Based on this task, affine transformations have a similar role to attention but with more parameters, which allows the proposed neural model to better exploit context. Therefore, it could be regarded as a form of a more sophisticated attention. Whereas the current "straightforward" attention models are overly simplistic, the model prioritises the contextual information of the candidate to discern its metaphoricity in a given sentence.

The proposed neural model consists of an affine transform coefficients generator that captures the meaning of the candidate to be classified, and a neural network that encodes the full text in which the candidate needs to be classified. I demonstrate that the proposed model significantly outperforms the state-of-the-art approaches on existing relation-level benchmark datasets. Since the main focus of this thesis is to identify metaphors in the user-generated content of tweets, I evaluate the proposed model on the ZayTw dataset of tweets annotated for relation-level metaphors that is developed as part of this thesis and was introduced in Section 4.2. The next subsections describe the proposed approach in detail along with the conducted experiments and comparisons on benchmark datasets.

5.4.1 Contextual Modulation-based Proposed Approach

Feature-wise transformation techniques such as feature-wise linear modulation (FiLM) have been recently employed in many applications showing improved performance. They became popular in image processing applications such as image style transfer (Dumoulin et al., 2017); then they found their way into multi-modal tasks, specifically visual question-answering (de Vries et al., 2017; Perez et al., 2018). They also have been shown to be effective approaches for relational problems as mentioned previously. The idea behind FiLM is to condition the computation carried out by a neural model on the information extracted from an auxiliary input in order to capture the relationship between multiple sources of information (Dumoulin et al., 2018).

The proposed approach presented here adopts Perez's (2018) formulation of FiLM on visual reasoning for metaphor identification. In visual reasoning, image-related questions are answered by conditioning the image-based neural network (visual pipeline) on the question context via a linguistic pipeline. In metaphor identification, it can be considered that the image in this case is the sentence that has a metaphoric candidate and the auxiliary input is the linguistic interaction between the components of the candidate itself. This will allow conditioning the computation of a sequential neural model on the contextual information of the candidate and leverage the feature-wise interactions between the conditioning representation and the conditioned network. To the best of my knowledge, this is the first work to propose such contextual modulation for textual classification in general and for metaphor identification specifically.

The proposed architecture consists of *a contextual modulation* pipeline and *a metaphor identification linguistic* pipeline as shown in Figure 5.5. The input to the contextual modulator is the deep contextualised representation of the candidate expression under study (which I will refer to as targeted expression[7]) to capture the interaction between its components. The linguistic pipeline employs an LSTM encoder which produces a contextual representation of the provided sentence where the targeted expression appeared. The model is trained end-to-end to identify relation-level metaphoric expressions focusing on verb-noun and adjective-noun grammatical

7 Targeted expressions are already annotated in the dataset and initially obtained either manually or automatically using a dependency parser as will be described in Chapter 3.

relations. The model takes as input a sentence (or a tweet) and a targeted expression of a certain syntactic construction and identifies whether the candidate in question is used metaphorically or literally by going through the following steps:

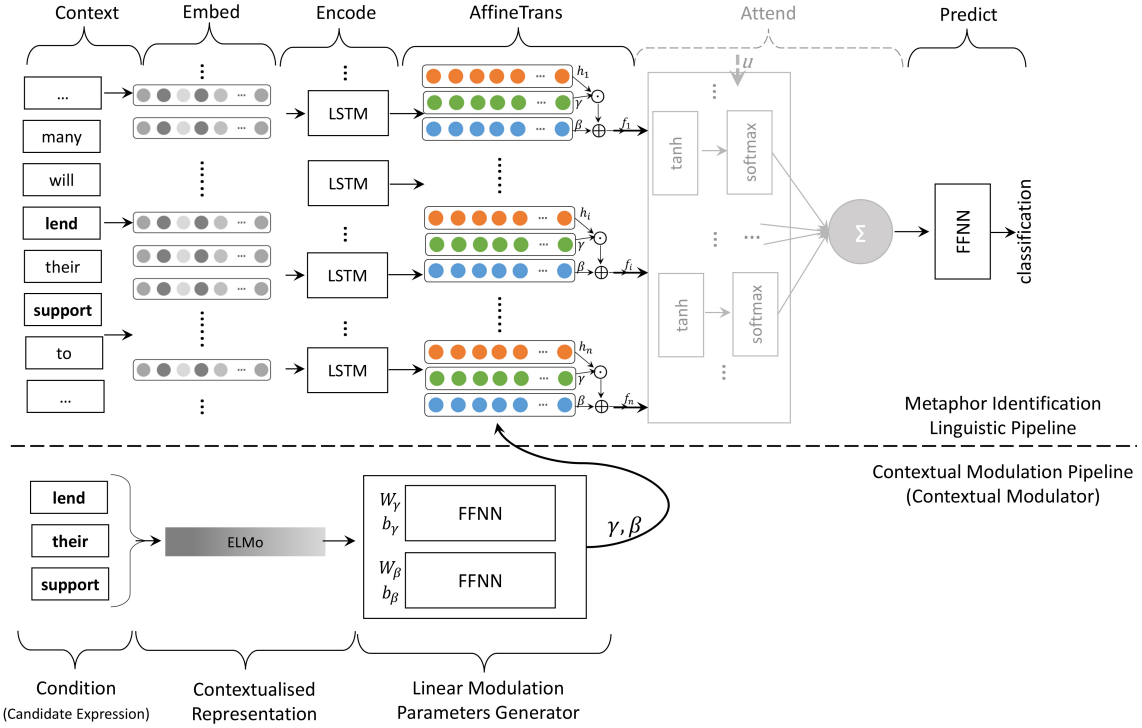


Figure 5.5: The proposed framework for relation-level metaphor identification using contextual modulation showing the system architecture in detail. The attention process is greyed out as experiments are done with and without it.

Condition: In this step the targeted expression is used as the auxiliary input to produce a conditioning representation. I first embed each candidate of verb-direct object pairs[8] (v, n) using ELMo sentence embeddings to learn context-dependent aspects of word meanings c_{vn}. I used the 1,024-dimensional ELMo embeddings pre-trained on the One Billion Word benchmark corpus (Chelba et al., 2014). The sentence embeddings of the targeted expression are then prepared by implementing an embeddings layer that loads these pre-trained ELMo embeddings from the TensorFlow Hub[9]. The layer takes in the raw text of the targeted expression and outputs a fixed mean-pooled vector representation of the input as the contextualised representation. This representation is then used as an input to the main component of this step, namely a contextual modulator. The contextual modulator consists of a fully-connected feed-forward neural network (FFNN) that produces the conditioning parameters (i.e. the shifting and scaling coefficients) that will later modulate the linguistic pipeline computations. Given that c_{vn} is the conditioning input then the contextual modulator outputs γ and β, the context-dependent scaling and shifting vectors, as follows:

$$\begin{aligned} \gamma(c_{vn}) &= W_\gamma c_{vn} + b_\gamma, \\ \beta(c_{vn}) &= W_\beta c_{vn} + b_\beta \end{aligned} \tag{5.13}$$

8 This is also done for subject-verb and adjective-noun pairs but, for simplicity, the process is demonstrated with verb-direct object pairs.

9 https://www.tensorflow.org/hub

where W_γ, W_β, b_γ, b_β are learnable parameters.

Embed: Given a labelled dataset of sentences, the model begins by embedding the tokenised sentence S of words $w_1, w_2, ..., w_n$, where n is the number of words in S, into vector representations using GloVe embeddings. I used the uncased 200-dimensional GloVe embeddings pre-trained on ∼2 billion tweets and contains 1.2 million words.

Encode: The next step is to train a neural network with the obtained embeddings. Since context is important for identifying metaphoricity, a sentence encoder is a sensible choice. I used an LSTM sequence model to obtain a contextual representation which summarises the syntactic and semantic features of the whole sentence. The output of the LSTM is a sequence of hidden states $h_1, h_2, ..., h_n$, where h_i is the hidden state at the i^{th} time-step.

Feature-wise Transformation: In this step, an affine transformation layer, hereafter *AffineTrans* layer, applies a feature-wise linear modulation to its inputs, which are: 1) the hidden states from the encoding step; 2) the scaling and shifting parameters from the conditioning step. By feature-wise, I mean that scaling and shifting are applied to each encoded vector for each word in the sentence.

$$f(h_i, c_{vn}) = \gamma(c_{vn}) \odot h_i + \beta(c_{vn}) \tag{5.14}$$

Attend: As discussed previously, attention mechanisms have recently become useful to select the most important elements in a given representation while minimising information loss. In this work, I employ an attention layer based on the mechanism presented in Lin et al. (2017). It takes the output from the *AffineTrans* layer as an input in addition to a randomly initialised weight matrix W, a bias vector b and a learnable context vector u to produce the attended output as follows:

$$e_i = tanh(Wf_i + b) \tag{5.15}$$

$$\alpha_i = softmax(ue_i) \tag{5.16}$$

$$r = \sum_{i=1}^{n} \alpha_i f_i \tag{5.17}$$

The model is trained and evaluated with and without the attention mechanism in order to differentiate between the effect of the feature modulation and the attention on the model performance.

Predict: The last step is to make the final prediction using the output from the previous step (attended output in case of using attention or the *AffineTrans* layer output in case of skipping it). I use a fully-connected feed-forward layer with a sigmoid activation that returns a single (binary) class label to identify whether the targeted expression is metaphoric or not.

5.4.2 Experiments

5.4.2.1 *Datasets*

As discussed in the previous chapters, the choice of the annotated dataset for training the model and evaluating its performance is determined by the level of metaphor identification. Given the distinction between the levels of analysis, approaches addressing the task on the word level are not fairly comparable to relation-level approaches since each task addresses metaphor identification differently. Therefore, the tradition of previous work in this area is to compare approaches addressing the task on the same level against each other on level-specific annotated benchmark datasets (see Chapter 3 for more discussion on this).

Following prior work in this area and in order to compare the performance of this proposed approach with other relation-level metaphor identification approaches, I utilise two categories of annotated datasets that support this level of processing. The first category comprises datasets that are originally prepared to directly support relation-level processing which are: 1) the benchmark TSV (Tsvetkov et al., 2014) dataset of annotated adjective-noun expressions and 2) the ZayTw dataset of tweets that is prepared as part of this thesis and annotated for verb-noun expressions (see Section 4.2 for extensive details). The second category of datasets comprises adapted datasets from other word-level benchmark datasets to suit relation-level processing. These datasets include the ones I adapted as part of this thesis as introduced in Section 4.3, namely the adaptation of the benchmark datasets TroFi (Birke and Sarkar, 2006) and VU Amsterdam metaphor corpus (VUAMC) (Steen et al., 2010). Additionally, I employ the MOH-X dataset which is an adaptation of the word-level MOH (Mohammad et al., 2016) dataset by Shutova et al. (2016). Table 5.10 revisits the statistics of these datasets including their size and percentage of metaphors. Additionally, Table 5.11 lists examples of the annotated instances from these datasets showing their format as: sentence, targeted expression and the provided label.

Table 5.10: Statistics of the employed benchmark datasets to train and evaluate the proposed models based on contextual modulation highlighting the used experimental setting.

Dataset	Syntactic structure	Text type	Size (sentences/ tweets)	% Metaphors	Average Sentence Length
The adapted TroFi Dataset	verb-direct object	50 selected verbs (News)	1,535	59.15%	48.5
The adapted VUAMC[10] (NAACL Shared Task subset)	verb-direct object	known-corpus (The BNC)	5,820	38.87%	63.5
The MOH-X Dataset (adapted from the MOH dataset[11])	verb-direct object; subject-verb	selected examples (WordNet)	647	48.8%	11
The TSV Dataset[12]	adjective–noun	selected examples (Web/Tweets)	1,964	50%	43.5
The ZayTw Dataset	verb-direct object	Tweets (general and political topics)	2,531	54.8%	34.5

5.4.2.2 *Experimental Setup*

The proposed model utilises a single-layer LSTM encoder with 512 hidden units.The word embeddings layer is initialised with the pre-trained GloVe embeddings. As mentioned earlier, I used the uncased 200-dimensional GloVe embeddings pre-trained on ∼2 billion tweets and contains 1.2 million words. I did not update the weights of these embeddings during training. The Adadelta algorithm (Zeiler, 2012) is used for optimisation during the training phase and binary cross-entropy is used as a loss function to fine tune the network. The reported results are obtained

Table 5.11: Examples of annotated instances from the employed relation-level datasets to assess the performance of the proposed contextual modulation-based approach showing their format as: sentence, targeted expression and the provided label.

Dataset	Sentence	Targeted Expression	Gold Label
TSV	Chicago is a big city, with a lot of everything to offer.	big city	0
	It 's a foggy night and there are a lot of cars on the motorway.	foggy night	0
	Their initial icy glares had turned to restless agitation.	icy glares	1
	And he died with a sweet smile on his lip.	sweet smile	1
ZayTw	insanity. ok to abuse children by locking them in closet, dark room and damage their psyche, but corporal punishment not ok? twisted!	abuse children	0
	nothing to do with your lot mate #ukip ran hate nothing else and your bloody poster upset the majority of the country regardless in or out	upset the majority	0
	nothing breaks my heart more than seeing a person looking into the mirror with anger & disappointment, blaming themselves when someone left.	breaks my heart	1
	how quickly will the warring tories patch up their differences to preserve power? #euref	patch up their differences	1
The adapted TroFi	A Middle Eastern analyst says Lebanese usually drink coffee at such occasions; Palestinians drink tea.	drink coffee	0
	In addition, the eight-warhead missiles carry guidance systems allowing them to strike Soviet targets precisely.	strike Soviet targets	0
	He now says that specialty retailing fills the bill, but he made a number of profitable forays in the meantime.	fills the bill	1
	A survey of U.K. institutional fund managers found most expect London stocks to be flat after the fiscal 1989 budget is announced, as Chancellor of the Exchequer Nigel Lawson strikes a careful balance between cutting taxes and not overstimulating the economy.	strikes a careful balance	1
The adapted VUAMC (NAACL Shared Task)	Among the rich and famous who had come to the salon to have their hair cut, tinted and set, Paula recognised Dusty Springfield, the pop singer, her eyes big and sooty , her lips pearly pink, and was unable to suppress the thrill of excitement which ran through her.	recognised Dusty Springfield	0
	But until they get any money back, the Tysons find themselves in the position of the gambler who gambled all and lost .	get any money	0
	The Labour Party Conference: Policy review throws a spanner in the Whitehall machinery	throws a spanner	1
	Otherwise Congress would have to face the consequences of automatic across-the-board cuts under the Gramm-Rudman-Hollings budget deficit reduction law.	face the consequences	1
MOH-X	commit a random act of kindness.	commit a random act	0
	The smoke clouded above the houses.	smoke clouded	0
	His political ideas color his lectures.	ideas color	1
	flood the market with tennis shoes.	flood the market	1

using batch size of 256 instances for the ZayTw dataset and 128 instances for the other employed datasets. L_2-regularisation weight of 0.01 is used to constraint the weights of the contextual modulator. In all experiments, I zero-pad the input sentences to the longest sentence length in the dataset. All the hyper-parameters were optimised on a randomly separated development set (validation set) by assessing the accuracy. I present here the best performing design choices based on experimental results but I highlight some other attempted considerations in Section 5.4.2.5. The models are implemented using Keras (Chollet, 2015) with the TensorFlow backend. The source code and best models are publicly available[13]. All experiments are done on a NVIDIA Quadro M2000M GPU of 4GB memory and the average running time for the proposed models is around one hour for maximum of 100 epochs.

5.4.2.3 *Excluding AffineTrans*

I implemented a simple LSTM model to study the effect of employing affine transformations on the system performance. The input to this model is the tokenised sentence S which is embedded as a sequence of vector representations using GloVe. These sequences of word embeddings are then encoded using the LSTM layer to compute a contextual representation. Finally, this representation is fed to a feed-forward layer with a sigmoid activation to predict the class label. I used this model with and without the attention mechanism.

5.4.2.4 *Results*

I conduct several experiments to better understand the proposed model. First, I experiment with the simple model introduced in Section 5.4.2.3. Then, I train the proposed models on the benchmark datasets discussed in Section 5.4.2.1. I experiment with and without the attention layer to assess its effect on the model performance. Furthermore, I compare the model to the current work that addresses the task on the relation level, in-line with other researchers in this area. In this work, I selected the following state-of-the-art models pertaining to relation-level metaphor identification for comparisons: the cross-lingual model by Tsvetkov et al. (2014), the multimodal system of linguistic and visual features by Shutova et al. (2016), the ATTR-EMBED model by Bulat et al. (2017) and the supervised similarity network (SSN) by Rei et al. (2017). I consider the SSN system as a baseline. For fair comparisons, I utilised their same data splits on the five employed benchmark datasets described in the previous subsection. Since the source code of Rei's (2017) system is available online[14], I trained and tested their model using the ZayTw dataset as well as the adapted VUAMC and TroFi dataset in an attempt to study the ability of their model to generalise when applied on a corpus of a different text genre with wider metaphoric coverage including less common (conventionalised) metaphors. Tables 5.12 and 5.13 show the model performance in terms of precision, recall, F1-score and accuracy. All the results presented in this section are obtained after running the experiments five times with different random seeds and taking the average.

To ensure reproducibility, Table 5.14 shows the sizes of the training, validation and test sets of each employed dataset as well as the corresponding best obtained validation accuracy by the best performing model, namely *AffineTrans* model (without attention).

13 https://github.com/OmniaZayed/affineTrans_metaphor_identification
14 https://github.com/marekrei/ssn

Table 5.12: Performance of the proposed architecture based on contextual modulation compared to the state-of-the-art approaches on the ZayTw and TSV datasets. *Statistically significant (p-value<0.01) compared to the SSN system (Rei et al., 2017).

	ZayTw (test-set)				TSV (test-set)			
	Prec.	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**
Tsvetkov et al. (2014)	-	-	-	-	-	-	**0.85**	-
Shutova et al. (2016) (multimodal)	-	-	-	-	0.67	**0.96**	0.79	-
Bulat et al. (2017) (ATTR-EMBED)	-	-	-	-	0.85	0.71	0.77	-
Rei et al. (2017) (SSN)	0.543	**1.0**	0.704	0.543	**0.903**	0.738	0.811	0.829
Simple LSTM	0.625	0.758	0.685	0.621	0.690	0.58	0.630	0.66
Simple LSTM (+ Attend)	0.614	0.866	0.718	0.631	0.655	0.55	0.598	0.63
AffineTrans	**0.804**	0.769	**0.786***	**0.773**	0.869	0.80	**0.834**	**0.84**
AffineTrans (+ Attend)	0.758	0.812	0.784*	0.757	0.875	0.77	0.819	0.83

Table 5.13: Performance of the proposed architecture based on contextual modulation compared to the state-of-the-art approaches on the MOH-X dataset, the adapted TroFi dataset and the adapted VUAMC. *Statistically significant (p-value<0.01) compared to the SSN system (Rei et al., 2017). Shutova et al. (2016) utilised different evaluation settings therefore their results on the MOH-X dataset are not strictly comparable.

	MOH-X (10-fold)				adapted TroFi (test-set)				adapted VUAMC (test-set)			
	Prec.	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**
Shutova et al. (2016) (multimodal)	0.65	**0.87**	0.75	-	-	-	-	-	-	-	-	-
Rei et al. (2017) (SSN)	0.736	0.761	0.742	0.748	0.620	0.892	0.732	0.628	0.475	0.532	0.502	0.558
Simple LSTM	0.757	0.773	0.759	0.759	0.70	0.751	0.725	0.674	0.510	0.339	0.407	0.587
Simple LSTM (+ Attend)	0.746	0.782	0.757	0.752	0.759	0.853	0.803*	0.761	0.575	0.423	0.487	0.627
AffineTrans	**0.804**	0.748	0.771	**0.780**	**0.852**	**0.909**	**0.879***	**0.858**	**0.712**	0.639	0.673*	**0.741**
AffineTrans (+ Attend)	0.753	0.813	**0.779**	0.773	0.841	0.870	0.856*	0.832	0.686	**0.679**	**0.683***	0.736

Table 5.14: Experimental information of the five benchmark datasets including the best obtained validation accuracy by the *AffineTrans* model (without attention). The splits used in the literature for the VUAMC and TSV datasets are preserved.

Dataset	**Train**	**Validation**	**Test**	**split %**	**Validation Accuracy**	**@epoch**
The adapted TroFi Dataset	1,074	150	312	70-10-20	0.914	40
The adapted VUAMC	3,535	885	1,398	-	0.748	20
The MOH-X Dataset	582 per fold	-	65 per fold	10-fold cross-validation	-	-
The TSV Dataset	1,566	200	200	-	0.905	68
The ZayTw Dataset	1,661	360	510	70-10-20	0.808	29

5.4.2.5 *Other Trials*

SENTENCE EMBEDDING: I experimented with different representations other than GLoVe in order to embed the input sentence. I tried to employ the contextualised pre-trained embeddings ELMo and BERT either instead of the GloVe embeddings or as additional features but no further improvements were observed on both validation and test sets over the best performance obtained. Furthermore, I experimented with different pre-trained GloVe embeddings including the uncased 300-dimensional pre-trained vectors on the Common Crawl dataset but I did not notice any significant improvements.

SENTENCE ENCODING: The choice of using the simple LSTM to encode the input was based on several experiments on the validation set. I tried bidirectional LSTM but observed no further improvement under this architecture. This could be due to the nature of the relation-level metaphor identification task itself as the *tenor* (e.g. noun) affects the metaphoricity of the *vehicle* (e.g. verb or adjective) so a single-direction processing was enough.

5.4.3 Discussion

Overall performance. I analysed the model performance by inspecting the classified instances. I noticed that it did a good job identifying conventionalised metaphors as well as uncommon ones. Tables 5.15 and 5.16 show examples of classified instances by the system from the employed benchmark datasets. The proposed model achieves significantly better F1-score over the state-of-the-art SSN system (Rei et al., 2017) under the one-tailed paired *t-test* (Yeh, 2000) at p-value$<$0.01 on three of the five employed benchmark datasets. Moreover, the architecture showed improved performance over the state-of-the-art approaches on the TSV and MOH-X datasets. It is worth mentioning that the size of their test sets is relatively smaller; therefore any change in a single annotated instance drastically affects the results. Moreover, the approach proposed by Tsvetkov et al. (2014) relies on hand-coded lexical features which justifies its high F1-score.

Table 5.15: Examples of correctly classified instances by the *AffineTrans* model (without attention) from the ZayTw and TSV datasets showing the classification probability.

Model	ZayTw		TSV	
Classification	**Expression**	**Prob.**	**Expression**	**Prob.**
Metaphor	poisoning our democracy	0.999	rich history	0.999
	binding the country	0.942	rocky beginning	0.928
	see greater diversity	0.892	foggy brain	0.873
	patch up their differences	0.738	steep discounts	0.723
	seeking information	0.629	smooth operation	0.624
	retain eu protection	0.515	dumb luck	0.512
Not Metaphor	shake your baby	0.420	filthy garments	0.393
	enjoy a better climate	0.375	clear day	0.283
	improve our cultural relations	0.292	slimy slugs	0.188
	placate exiters	0.225	sour cherries	0.102
	betrayed the people	0.001	short walk	0.014
	washing my car	0.000	hot chocolate	0.000

The effect of contextual modulation. When excluding the *AffineTrans* layer and only using the simple LSTM model, I observe a significant performance drop that shows the effectiveness of

Table 5.16: Examples of correctly classified instances by the *AffineTrans* model (without attention) from the MOH-X dataset, the adapted TroFi dataset and the adapted VUAMC showing the classification probability.

Model	MOH-X		adapted TroFi		adapted VUAMC	
Classification	Expression	Prob.	Expression	Prob.	Expression	Prob.
Metaphor	absorbed the knowledge	0.987	grasped the concept	0.985	bury their reservations	0.999
	steamed the young man	0.899	strike fear	0.852	reinforce emotional reticence	0.871
	twist my words	0.770	ate the rule	0.781	possess few doubts	0.797
	color my judgment	0.701	planted a sign	0.700	suppress the thrill	0.647
	poses an interesting question	0.543	examined the legacy	0.599	considers the overall effect	0.568
	wears a smile	0.522	pumping money	0.529	made no attempt	0.517
Not Metaphor	shed a lot of tears	0.484	pumping power	0.427	send the tape	0.482
	abused the policeman	0.361	poured acid	0.314	asking pupils	0.389
	tack the notice	0.274	ride his donkey	0.268	removes her hat	0.276
	stagnate the waters	0.148	fixed the dish	0.144	enjoying the reflected glory	0.188
	paste the sign	0.002	lending the credit	0.069	predict the future	0.088
	heap the platter	0.000	destroy coral reefs	0.000	want anything	0.000

leveraging linear modulation. This layer adaptively influences the output of the model by conditioning the identification process on the contextual information of the targeted expression itself which significantly improved the system performance, as observed from the results. Moreover, employing the contextualised representation of the targeted expression, through ELMo sentence embeddings, was essential to explicitly capture the interaction between the verb/adjective and its accompanying noun. Then, the *AffineTrans* layer was able to modulate the network based on this interaction.

The effect of attention. It is worth noting that the attention mechanism did not help much in the *AffineTrans* model because affine transformation itself could be seen as playing a similar role to attention, as discussed at the beginning of this section. In attention mechanisms important elements are given higher weight based on weight scaling whereas in linear affine transformation scaling is done in addition to shifting which gives prior importance (probability) to particular features.

Error analysis. An error analysis is performed to determine the model flaws by analysing the predicted classification. I examined the false positives and false negatives obtained by the best performing model, namely *AffineTrans* (without attention). Interestingly, the majority of false negatives are from the political tweets in ZayTw dataset. Table 5.17 lists some examples of misclassified instances in the TSV and ZayTw datasets. Some instances could be argued as being correctly classified by the model. For instance, *"spend capital"* could be seen as a metaphor in that the noun is an abstract concept referring to actual money. Examples of misclassified instances from the other employed datasets are presented in Table 5.18. Interestingly, I noticed that the model was able to spot mistakenly annotated instances. Although the adapted VUMAC subset contains various expressions which should help the model perform better, I noticed annotation inconsistency in some of them. For example, the verb *"choose"* associated with the noun *"science"* is annotated once as metaphor and twice as literal in very similar contexts. This aligns well with the findings that I discussed in Section 4.3.2.2 while adapting this dataset where I questioned the annotation of around 5% of the instances in this subset mainly due to annotation inconsistency.

Analysis of some misclassified verbs. I noticed that sometimes the model got confused while identifying the metaphoricity of expressions where the verb is related to emotion and cognition such as: *"accept, believe, discuss, explain, experience, need, recognise,* and *want"*. The model tends to

Table 5.17: Misclassified examples by the *AffineTrans* model (without attention) from ZayTw and TSV test sets. Sentences are truncated for better representation. *The *AffineTrans* model was able to spot some mistakenly annotated instances.

	ZayTw		TSV	
	Tweet	**Prob.**	**Sentence**	**Prob.**
False Negative	hard to **resist the feeling** that remain is further [...]	0.46	You have a **shiny goal** in mind that is distracting you with its awesomeness.	0.49
	@abpi uk: **need #euref final facts**? read why if [...]	0.08	The first hours of a **shaky ceasefire** are not "the best of times".	0.14
	#ivoted with a black pen. do not **trust pencils**. [...]	0.003	The French bourgeoisie has rushed into a **blind alley**.	0.00
False Positive	[...] this guy would **spend so much political capital** trying to erase the [...]	0.96	I could hear the **shrill voices** of his sisters as they dash about their store helping customers.	0.98
	#pencilgate to **justify vitriolic backlash** if #remain wins [...]	0.94	[...] flavoring used in cheese, meat and fish to give it a **smoky flavor** could in fact be toxic.	0.82
	@anubhuti921 @prasannas it **adds technology** to worst of old police state practices, [...]	0.76*	Usually an overly dry nose is a precursor to a **bloody nose**.	0.64

classify them as not metaphors. I include different examples from the ZayTw dataset of the verbs *"experience"* and *"explain"* with different associated nouns along with their gold and predicted classifications in Table 5.19. The prediction of the model seems reasonable given that the instances in the training set were labelled as not metaphors. It is not clear why the gold label for *"explain this mess"* is not a metaphor while it is metaphor for *"explain implications"*; similarly, the nouns *"insprirations"* and *"emotions"* with the verb *"experience"*.

5.5 SUMMARY

This chapter covered the work done in this thesis under the second research theme to identify linguistic metaphors in tweets under the relational paradigm. I explored three main ideas to achieve this which are distributional semantics, meta-embedding learning and contextual modulation. The chapter recapped the difference between the levels of metaphor identification focusing on the word and relation levels. As discussed, the majority of previous works adopted the word-level paradigm to identify metaphors in text. The main distinction between the relation-level and the word-level paradigms is that the former makes the context more explicit than the latter through highlighting the relationship between the metaphor components (the *tenor* and the *vehicle*). These explicitly marked relations is important to support other downstream metaphor processing tasks such as interpretation and cross-domain mappings. From a cognitive perspective, it is necessary to determine such relation between the metaphor components (i.e. the *ground*) in order to comprehend a metaphor.

At the beginning of this work I investigated the feasibility of introducing a minimally supervised approach based on distributional semantics to identify linguistic metaphors with the aim of aiding in the creation and annotation process of a metaphor dataset of tweets. I explored the use of different pre-trained word embedding models to identify relation-level metaphors focusing on verb-noun expressions. The proposed approach, namely *DistSemant*, employs a predefined seed set of metaphoric expressions to classify new unseen ones that are highlighted in a given sentence based on semantic similarities between verbs and nouns. As discussed, in contrast to other related approaches, the proposed approach employs fewer lexical resources and does

Table 5.18: Misclassified examples by the *AffineTrans* model (without attention) from the adapted TroFi and VUAMC test sets as well as the relation-level datasets MOH-X, TSV and ZayTw datasets. *The *AffineTrans* model was able to spot some mistakenly annotated instances in the dataset.

	Dataset	Sentence	Prob.
False Negative	TroFi	Unself-consciously , the littlest cast member with the big voice steps into the audience in one number to open her wide cat-eyes and throat to **melt the heart** of one lucky patron each night.	0.295
		Lillian Vernon Corp., a mail-order company, said it is experiencing delays in **filling orders** at its new national distribution center in Virginia Beach,Va.	0.006
	VUAMC	It is a curiously paradoxical foundation uponupon which to **build a theory** of autonomy.	0.410
		It has turned up in Canberra with Japan to develop Asia Pacific Economic Co-operation (APEC) and a new 12-nation organisation which will **mimic the role** of the Organisation for Economic Co-operation and Development in Europe.	0.000
	MOH-X	When does the **court of law sit**?	0.499
		The **rooms communicated**.	0.000
	TSV	It was great to see a **warm reception** for it on twitter.	0.488
		An **honest meal** at a reasonable price is a rarity in Milan.	0.000
	ZayTw	#brexit? we **explain likely implications** for business insurances on topic of #eureferendum	0.2863
		@abpi uk: **need #euref final facts**? read why if you care about uk life sciences we're #strongerin.	0.0797
False Positive	TroFi	As the struggle enters its final weekend , any one of the top contenders could **grasp his way** to the top of the greasy pole.	0.998*
		Southeastern poultry producers fear **withering soybean supplies** will force up prices on other commodities.	0.507
	VUAMC	Or after we **followed the duff advice** of a legal journalist in a newspaper?	0.999*
		Aristotle said something very interesting in that extract from the Politics which I quoted earlier; he said that women have a deliberative faculty but that it **lacks full authority**.	0.525
	MOH-X	All our **planets condensed** out of the same material.	0.999
		He bowed before the King.	0.868
	TSV	Bags two and three will only have **straight edges** along the top and the bottom.	0.846
		Mountain climbers at **high altitudes** quickly acquire a tan from the sun.	0.986
	ZayTw	delayed flight in fueturventura due to french strikes **restricting access** across french airspace =/ hopefully get back in time to #voteleave	0.9589
		in manchester more young people are expected to **seek help** in the coming months and years #cypiapt #mentalhealth	0.7055*

Table 5.19: Examples of classified instances of the verbs *"experience"* and *"explain"* in the ZayTw test set.

	Expression	tweet	Predicted	Prob.	Gold
experience	the inspiration	relive the show , re - listen to her messages, re - experience the inspiration, refuel your motivation	0	0.220	1
	your emotions	do not be afraid to experience your emotions; they are the path to your soul. trust yourself enough to feel what you feel.	0	0.355	0
	this shocking behaviour	a friend voted this morning & experienced this shocking behaviour. voting is everyone 's right. #voteremain	0	0.009	0
explain	likely implications	#brexit? we explain likely implications for business insurances on topic of #eureferendum	0	0.2866	1
	this mess	@b_hanbin28 ikr same here :D imagine hansol & shua trynna explain this mess to other members :D	0	0.109	0
	the rise	loss aversion partly explains the rise of trump and ukip	1	0.618	1

not require annotated datasets or highly-engineered features. This gives it a flexibility to be easily adapted to new domains or text types. Moreover, it generalises better when compared to related minimally supervised approaches. Several experiments have been performed to assess the performance of this approach on benchmark datasets.

In this work, I also studied the various features that have been introduced to identify metaphors in text by previous approaches under a unified neural architecture. I investigated meta-embedding learning methods to study the effectiveness of semantic and psycholinguistic features in conjunction with deep contextualised features to identify relation-level metaphors in tweets. The effect of each of the proposed features is studied individually and collectively using two strategies of creating feature ensembles which are concatenation (Concat) and dynamic meta-embeddings (DME). The former strategy is straightforward and involves concatenating all embeddings along the sequence dimension whereas the latter depends on a linear projection of the original embeddings. The analysis of the proposed models revealed that employing well-established linguistic features for metaphor identification in a neural architecture along with the advanced deep contextualised features led to a significant improvement over the current work on the ZayTw metaphor dataset of tweets. The proposed DME model allowed the network to automatically select different embeddings based on the training data. However, one limitation of this model is that it requires more training data which is one of the challenges that faces the metaphor processing research.

Finally, I looked at contextual modulation, specifically through affine transformations, which have been shown to be a powerful device for relational problems in both the field of NLP and visual reasoning. The motivation behind this idea is to exploit the interaction between the metaphor components as well as the context (e.g. a sentence/tweet) to inform the decision regarding a specific relation in the text. Based on this idea, I therefore introduced a novel architecture to identify metaphors by utilising feature-wise affine transformation and deep contextual modulation. The proposed approach employs a contextual modulation pipeline to capture the interaction between the metaphor components. This interaction is then used as an auxiliary input to modulate a metaphor identification linguistic pipeline. I showed that such modulation allowed the computational model to dynamically highlight the key contextual features to identify the metaphoricity of a given expression. The approach is applied to relation-level metaphor identification to classify expressions of certain syntactic constructions for metaphoricity as they occur in context. It significantly outperformed the state-of-the-art approaches for this level of analysis on benchmark datasets. The experiments also showed that the proposed contextual modulation-based model, namely *AffineTrans*, can generalise well to identify the metaphoricity of unseen instances in different text types including the noisy user-generated text of tweets. The *AffineTrans* model was able to identify both conventionalised common metaphoric expressions as well as less common ones. To the best of my knowledge, this is the first attempt to computationally identify metaphors in tweets and the first approach to study the employment of feature-wise linear modulation on metaphor identification in general. The proposed methodology is generic and can be applied to a wide variety of text classification approaches including sentiment analysis or term extraction.

6 METAPHOR INTERPRETATION IN TWEETS

"Metaphor is the dream work of language and, like all dream work, its interpretation reflects as much on the interpreter as on the originator. [...] and the act of interpretation is itself a work of the imagination. So too understanding a metaphor is as much a creative endeavor as making a metaphor, and as little guided by rules."

(Davidson, 1978)

This chapter addresses the third, and last, research theme in this thesis, which is *Metaphor Interpretation in Tweets*. The main aim of the work presented here is to investigate the feasibility of employing advanced neural models to automatically interpret and explain the intended meaning of a metaphor. This thesis defines metaphor interpretation as a definition generation task with the aim of aidi language learners and non-native speakers to understand metaphors as well as enrich the process of developing lexical resources. I therefore investigate the feasibility of such task formulation by studying definition modelling. I then propose a neural approach based on sequence-to-sequence language modelling and applied it to interpret linguistic metaphors of the predicate type in tweets.

The work presented in this chapter, under the aforementioned research theme, seeks an answer to research question RQ3 that was discussed in detail in Chapter 1, and is formulated as follows:

RQ3 Can an advanced neural architecture be implemented to generate reliable definitions (interpretations) of metaphoric expressions that aid people in understanding them?

The chapter first recaps the idea of metaphor interpretation and its importance. Then, it highlights the motivation behind the choice of defining this task as definition generation (modelling) in this thesis. Section 6.2 presents a detailed overview of definition modelling and the various approaches proposed to study it in the literature. Then, Section 6.3 discusses the proposed approach and the experiments explored in this work in order to interpret verb-noun metaphoric expressions.

6.1 INTRODUCTION

Metaphor understanding and interpretation is a crucial element of human cognition and communication. Understanding the intended meaning of a metaphor is highly subjective and depends on various linguistic, cultural and psychological aspects. Metaphors rely on the imaginative and creative employment of words by the metaphor creator and, in turn, their interpretation relies on

the imagination of the receiver; therefore there could be several interpretations to some metaphors based on what they call to the attention (Davidson, 1978). However, in this thesis, I am only interested in obtaining the most general interpretation that comes to the mind of the majority of people. As part of the work done in this thesis, an experiment was conducted to prepare a dataset for metaphor interpretation, as discussed in Section 4.4, in which six native speakers were asked to select the most suitable definition for a given metaphoric expression among various choices. As discussed earlier, the main goal of this thesis is to process linguistic metaphors in tweets. And since Twitter is a social media platform for informal communications that gathers people from different backgrounds, obtaining all the possibilities of metaphoric interpretations might be challenging and is beyond the scope of this work.

This chapter presents the proposed approach to automatically interpret a given metaphor focusing on linguistic metaphors of the predicate type. As discussed in Section 3.3.1, metaphor interpretation can be viewed as lexical substitution, paraphrase generation or definition generation. Although the majority of previous works adopt the first approach to tackle the interpretation task, I adopt the last one, in this thesis, having the language learning and the enrichment of lexical resources as the end applications in mind.

The first part of the work presented in this chapter focuses on studying the previous research pertaining to definition generation in general. This work further investigates the feasibility of casting the task of metaphor interpretation as definition generation. I propose a sequence-to-sequence neural model based on a dual-encoder architecture for context-aware definition modelling in order to interpret metaphors in tweets. I investigate the effectiveness of the proposed approach to obtain sense specific definitions for metaphoric expressions. I conduct experiments on benchmark datasets of definitions as well as the metaphor interpretation dataset introduced in Chapter 4.

The following contributions are made in this chapter, which cover the last research theme of this thesis and seek an answer to **RQ3**, as follows:

- Approaching the metaphor interpretation task as definition generation and investigating definition modelling of metaphoric expressions.
- Employing an attention-based sequence-to-sequence neural model that utilises a dual encoder architecture and contextualised sentence embeddings to interpret metaphors as they occur in text.

6.2 DEFINITION GENERATION

Definition generation is important for various applications such as language learning and lexical resource preparation. This task focuses on automatically generating a meaning (definition) of a targeted word (or expression) in a given context in a way similar to the manual process of looking up dictionaries or lexicons. Basically, it is a natural language generation problem where a sequence of words should be generated using a target word. As discussed in Section 3.3.3, Martin (1990) was the first to introduce this task in the context of metaphor processing to interpret metaphoric expressions using previously stored knowledge about conventional metaphors. Recently, Noraset et al. (2017) introduced the task of *definition modelling* to generate a dictionary definition of a given word using word embeddings. Various works took this idea forward including Ni and Wang

(2017); Gadetsky et al. (2018); Chang et al. (2018); Ishiwatari et al. (2019); Mickus et al. (2019); Washio et al. (2019) and Li et al. (2020). The next paragraphs explain definition modelling as well as these approaches in detail.

While previous approaches, such as Wang et al. (2015) and Hill et al. (2016), have considered using dictionary definitions to learn word embeddings, Noraset et al. (2017) employed word embeddings to generate definitions and coined the term *definition modelling* to describe the task. The goal of definition modelling is to predict the probability of a definition $D = w_1, \ldots, w_T$ given the word being defined w^* and assuming that the probability of generating the t^{th} word of the definition text depends on both the previous words and w^*. This process can be modelled with a conditional language model, hence the name definition modelling, as follows:

$$p(D|w^*) = \prod_{t=1}^{T} p(w_t|w_{i<t}, w^*) \tag{6.1}$$

The idea of Noraset et al. is to employ a language model (LM) based on recurrent neural networks (RNN) (Mikolov et al., 2010), henceforth RNN-LM, that takes the word being defined as a seed to condition the modelling process. Their model represents w^* and w_t in terms of their corresponding word embeddings v^* and v_t, respectively. And then outputs a hidden representation h_t at each time step as follows:

$$h_t = g(v_{t-1}, h_{t-1}, v^*) \tag{6.2}$$

where g is a recurrent nonlinear function.
The authors introduced three models that vary in the method of incorporating w^* in their computations, namely *Seed, Input* and *Gate*. The best performing one is the *Gate* model which utilises a gated update function that dynamically controls the influence of the word being defined on the recurrent model at each time step. The authors explored morphological features including character embeddings to represent affixes of the word being defined in addition to hypernym embeddings. This approach was evaluated on a dataset of definitions compiled from WordNet (Fellbaum, 1998) and the Collaborative International Dictionary of English (GCIDE). One drawback of this approach is that it ignores the context where the word being defined occurs which means that the model cannot handle polysemous words. Therefore, the prepared dataset as part of this work is context-agnostic which means that there is no example sentence provided with each word and its associated definition.

Ni and Wang (2017) took a step towards overcoming the main limitation of Noraset et al.'s approach by considering the context around the word being defined. Unlike Noraset et al. (2017), they formulated the task of definition modelling as a sequence-to-sequence language modelling task. The authors introduced a sequence-to-sequence model to explain non-standard (slang) English expressions as they occur in text. The proposed approach utilises both the context words and the target expression to generate the explanation. The proposed sequence-to-sequence model conditions the probability of the generated definition D on both the word being defined w^* and

the context words $C = w_1, \ldots, w_c$ (where the word w^* is part of the words in of the context C). Therefore, Equation 6.1 can be extended as follows:

$$p(D|w^*, C) = \prod_{t=1}^{T} p(w_t|w_{i<t}, w^*, C) \tag{6.3}$$

The network architecture is based on a dual encoder using LSTMs (Hochreiter and Schmidhuber, 1997) that comprises a word-level encoder to encode the context words and a character-level one to encode the non-standard expression. The LSTM-based decoder is responsible for generating the possible explanation of the target expression. The authors introduced the first non-standard English corpus from the Urban Dictionary to train and evaluate their model. The dataset includes example sentences for the slang terms and their definitions.

Gadetsky et al. (2018) also considered the context around the word being defined to generate its definition. The proposed model conditions the probability of the generated definition D on both the word being defined w^* and the context words C by employing Equation 6.3. The model concatenates the embeddings of the word being defined with the embeddings of the generated word at each time step, similar to the *Input* model of Noraset et al. (2017). The context words are used to disambiguate the word being defined using a sigmoid-based gating mechanism. The authors prepared a context-aware dataset from the Oxford Dictionary[1] that comprises the word being defined, its definition and an example of the use of this word under the given sense (meaning).

The main aim of Chang et al. (2018) is to interpret word embeddings and then employ these interpretations for definition generation. The authors introduced an explainable word sense networks (xSense) that comprises four main modules in an encoder-decoder neural architecture. The network first encodes the context around the word being defined to obtain word embeddings then a sparse vector representation of the word being defined is extracted. A mask generator is then used to combine the encoded context and the representation of the word being defined in an attempt to encode the sense information. Finally, the encoded information is passed to a decoder to generate the corresponding definition. This approach was evaluated on a compiled dataset from the Oxford Dictionary that the authors introduced to include many senses and example sentences for a given word. This dataset differs from the one introduced by Gadetsky et al. (2018) in that it contains more examples for each word sense.

Another prominent work on definition modelling is Ishiwatari et al. (2019) who formulated the task as sequence-to-sequence language modelling. The authors followed the task formulation of Ni and Wang (2017). Therefore, unlike the majority of previous approaches who employed a variant of the encoder-decoder architecture by replacing the RNN-based encoder with an embedding layer, this work employed the traditional encoder-decoder architecture of sequence-to-sequence networks (Sutskever et al., 2014; Bahdanau et al., 2015). The authors introduced local and global context-aware description generator (LOG-CaD) that employs an attention-based encode-decoder architecture to define a given target word (or an expression). The model comprises two encoders and a definition decoder similar to Ni and Wang's dual encoder architecture. The first encoder is an LSTM-based one to encode the local context that represents the sentence (context) around the target word or expression to be defined. The second encoder is used to obtain the embeddings of

1 The authors specified this source for the Oxford Dictionary which was accessed in 2018 `oxforddictionaries.com`. Although not specified, I assume that they used the Oxford English Dictionary since their research main focus was the English language.

the target word which the authors considered as the global context. The authors utilised a gated update method similar to Noraset et al. (2017) to control the influence of the context and the word to be defined on the generated definition. As additional features the authors incorporated the character embeddings of the target word. It is worth mentioning that this work also considers defining multi-word expressions (MWEs) by simply summing up all the embedding vectors of the words in the expression. Following Noraset et al. (2017), the authors prepared two datasets, one from WordNet and the other from Wikipedia, to train and test the proposed model. The proposed context-aware datasets include the word (expression) being defined, its definition and an example sentence that describes the usage of the word's meaning. Additionally, the proposed model was evaluated on the context-aware datasets proposed by Gadetsky et al. (2018) from the Oxford Dictionary and Ni and Wang (2017) from the Urban Dictionary for comparisons purposes.

Mickus et al. (2019) stressed that definition modelling should be treated as a sequence-to-sequence task (similar to what Ni and Wang (2017); Gadetsky et al. (2018); Chang et al. (2018) and Ishiwatari et al. (2019) have introduced) rather than a word-to-sequence one (following the formulation of Noraset et al. (2017)). They too incorporate the context around the word being defined, but unlike previous approaches which incorporated it separately, the authors proposed encoding it directly with its context. The proposed architecture which is based on transformers (Vaswani et al., 2017) and the OpenNMT library (Klein et al., 2017) takes as an input a sentence and a highlighted word to be defined. The proposed approach was evaluated on the definition context-agnostic and context-aware datasets of Noraset et al. (2017) and Gadetsky et al. (2018), respectively.

The work introduced by Washio et al. (2019) focused on utilising the implicit lexical semantic relations between the words being defined and the words of their corresponding definition. The proposed approach represents semantic relations of word pairs through learning word-pair embeddings (Washio and Kato, 2018a,b). These embeddings are prepared in an unsupervised fashion based on the co-occurrences between word-pairs and lexico-syntactic patterns in a corpus. The authors proposed an encoder-decoder neural architecture to generate the definitions. The proposed model was evaluated on both the context-agnostic dataset of Noraset et al. (2017) and context-aware dataset of Gadetsky et al. (2018).

Chang and Chen (2019) further developed the previous work Chang et al. (2018) and reformulated the task of definition modelling to be definition selection. Therefore, they view the task as a classification task, instead of a natural language generation one. The authors introduced an approach to obtain word definitions using contextualised word embeddings, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), by learning a mapping between two semantically continuous spaces. The main idea behind this proposed approach is to learn a non-linear mapping to transform the contextualised word embeddings of the word being defined (and its context) into a corresponding sense-specific definition. This can be formulated as follows:

$$f^* = \underset{f}{\operatorname{argmin}} \|f(x) - y\|_2 \tag{6.4}$$

where x comprises the context-independent embeddings of the target word concatenated with either its context-dependent embeddings or the embeddings of its context; and y is the embeddings of the corresponding definition. The authors introduced a neural model that encodes these embeddings using a pre-trained transformer-based universal sentence encoder (Cer et al., 2018).

The model was trained and evaluated on the dataset of definitions from the Oxford Dictionary that was prepared by Chang et al. (2018).

More recently, Li et al. (2020) proposed the idea of explicit semantic decomposition (ESD) for definition generation to explicitly model the semantic components of the word being defined using discrete latent variables. Similar to Ishiwatari et al. (2019), the proposed architecture comprises two encoder; one to obtain the embeddings of the word being defined and the other to encode its surrounding context. The model also includes a semantic component predictor to model the semantic components of the word being defined and a definition decoder that generates the corresponding definition. The approach was evaluated on the datasets introduced in Ishiwatari et al. (2019) from the Oxford Dictionary and WordNet.

Bevilacqua et al. (2020) introduced a generative dictionary approach to lexical semantics, namely Generationary, to address definition modelling. The authors employ a span-based encoding scheme to fine-tune a pre-trained sequence-to-sequence model based on transformers, namely BART (Lewis et al., 2020), to generate context-aware definitions. The proposed approach was evaluated using the context-aware dataset of definitions from the Oxford Dictionary (Chang et al., 2018). The authors also introduced a dataset of uncommon adjective-noun phrases that are not found in traditional dictionaries (e.g. *"exotic cuisine"*). The dataset, referred to as Hei++, comprises around 713 adjective-noun phrases with their corresponding definitions that were written by a lexicographer.

6.3 METAPHOR INTERPRETATION AS DEFINITION GENERATION

As discussed earlier, metaphor interpretation can be defined as a definition generation task. The work presented in this section seeks an answer to **RQ3** which focuses on exploring advanced neural approaches to metaphor interpretation in order to aid language learners in understanding metaphoric expressions. Inspired by the previous works on definition generation, my goal is to explore a definition modelling formulation of metaphor interpretation with the aim of providing reliable explanations (definitions) of given metaphoric expressions as they occur in text. Definition modelling has been viewed as a conditional language modelling task, which generates a sequence of words given the word being defined (and/or its surrounding context). Therefore, the majority of previous approaches on definition modelling employed either RNN-based or sequence-to-sequence-based language models, as discussed in the previous section.

In this section, I explore utilising a sequence-to-sequence language model trained on a dataset of dictionary definitions to interpret metaphors. I follow the formulation of Ni and Wang (2017) and Ishiwatari et al. (2019) for context-aware definition modelling. The proposed neural model follows the dual encoder architecture introduced in the literature in order to encode the word being defined as well as the context around it. I performed experiments to train the proposed model on context-aware datasets of definitions which are the WordNet dataset (Ishiwatari et al., 2019) and the Oxford Dictionary dataset (Gadetsky et al., 2018). I applied the model on the metaphor interpretation dataset that is developed as part of this thesis, and was introduced in Section 4.4, in order to test the performance of the proposed model on interpreting metaphors. The next subsections explains the proposed approach in detail along with the conducted experiments and findings.

6.3.1 Proposed Sequence-to-Sequence Model

The proposed architecture consists of a dual encoder and a definition generation decoder as shown in Figure 6.1. The system takes as an input a word (or expression) to be defined and an example sentence, which represents the usage of the target word in a context, and then generates a possible definition based on the word sense in the given context. The main components of the system can be described as follows:

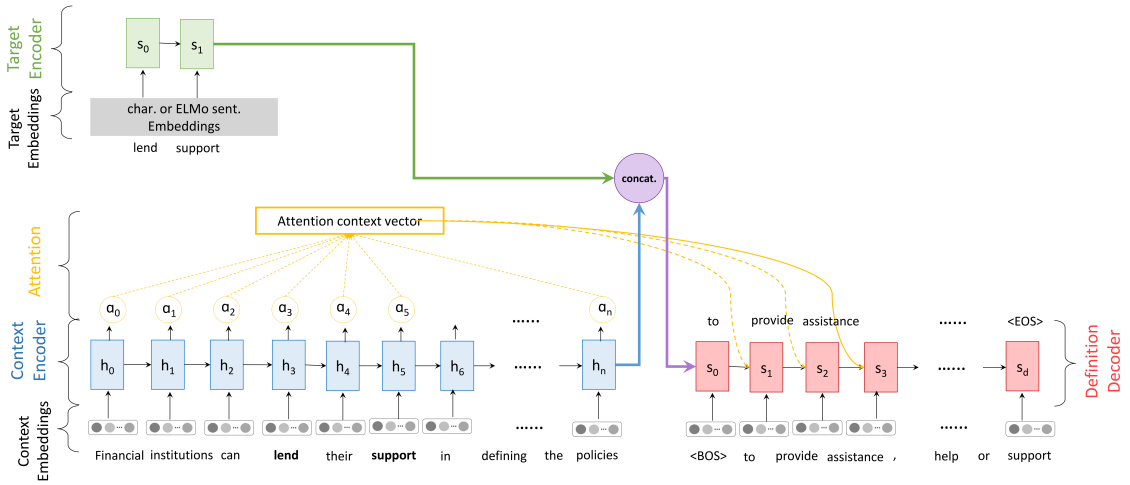


Figure 6.1: The proposed attention-based sequence-to-sequence architecture for definition modelling utilising a dual encoder.

Context Encoder: Given an example sentence that represents the context C around the target word w^*, the model begins by embedding the sentence to its vector representation V depending on the chosen embedding method. I define the context as $C = w_1, \ldots, w_n$ (where w^* could be either a single word or a multi-word expression and is part of the words of the context) and the embeddings of the context words as $V = v_1, \ldots, v_n$. I experimented with pre-trained context-independent embeddings. More specifically, I employed the 300-dimensional GloVe embeddings (Pennington et al., 2014) pre-trained on the Common Crawl dataset. The embedded context is then passed to an LSTM encoder to produce a sequence of hidden states $H = h_1, \ldots, h_n$ that represent the input as follows:

$$h_t = LSTM(h_{t-1}, v_{t-1}) \tag{6.5}$$

where h_t is the hidden state at each time step.

Target Encoder: The second encoder in the proposed dual encoder architecture focuses on the target word or expression to be defined which is denoted as w^*. I experiment with embedding the target using either character embeddings similar to Ni and Wang (2017) or using contextualised sentence embeddings, namely ELMo sentence embeddings (Peters et al., 2018). For the ELMo embeddings, I used the 1,024-dimensional embeddings pre-trained on the One Billion Word benchmark corpus (Chelba et al., 2014). The idea of using sentence embeddings here is to capture the interaction between the words in the case of defining a multi-word target expression such as a metaphor of verb-noun pair. For the character-level embeddings, I utilise an LSTM encoder of

256 hidden units to obtain H^* continuous hidden states which will be linearly concatenated with the hidden states of the context encoder. This concatenation can be defined as $H' = [H; H^*]$.

Attention: To allow the decoder to attend to important information in the example sentence, I employ the attention mechanism proposed by Luong et al. (2015). This mechanism gives the decoder a weighted average of the encoder states for each element of the decoder sequence. This weighted average is called the attention context vector and is obtained as follows:

$$\alpha_i = \text{softmax}_i(s_{t-1} \cdot h_i) \tag{6.6}$$

$$d_t = \sum_{i=1}^{n} \alpha_i h_i \tag{6.7}$$

where s_t is the hidden state of the decoder at time step t, h_i is the encoder hidden state and d_t is the attention context vector which will be concatenated to the decoder state during the generation step. This concatenation can be denoted as $s'_t = [s_t; d_t]$.

Definition Decoder: In order to generate the definition, the concatenated hidden states from both encoders H' are used to initialise an LSTM-based decoder. The decoder computes the conditional probability of generating a definition by producing a word at each time step given its state which is concatenated to the attention context vector and the previously generated word. The final prediction can be obtained by Equation 6.3. Therefore, the decision y_t at each time step can be formulated as:

$$y_t = \underset{w_t \in D}{\text{argmin}}\, p(w_t | w_{i<t}, w^*, C) \tag{6.8}$$

$$p(w_t | w_{i<t}, w^*, C) = \text{softmax}(W s'_t + b) \tag{6.9}$$

where W and b are learnable and denote a weight matrix and a bias vector, respectively.

6.3.2 Experiments

6.3.2.1 *Datasets*

The proposed model is trained on two benchmark datasets for definition modelling which are the WordNet dataset (Ishiwatari et al., 2019) and the Oxford Dictionary dataset (Gadetsky et al., 2018). I discuss the properties of each dataset as follows:

The WordNet Dataset: As part of the work presented in Ishiwatari et al. (2019), a dataset was compiled to evaluate the proposed approach. The authors followed the same method that Noraset et al. (2017) utilised to prepare their context-agnostic dataset of definitions but to prepare a context-aware dataset. Around 20k words from the 50k most frequent words in the Google Web 1T corpus (Brants and Franz, 2006) were sampled to query WordNet and obtain a definition for each word and an associated example sentence to represent its usage in a context. The dataset comprises around 10K single words to be defined along with their definitions and example sentences. A word might have multiple senses based on the context, therefore the total number of entries in the dataset is around 17K.

The Oxford Dictionary Dataset: Gadetsky et al. (2018) prepared a context-aware definitions dataset from the Oxford Dictionary focusing on English text. The authors employed the dictionary

API[2] to retrieve the definitions of a given word and it example sentence. This dataset comprises around 51K words and around 122K entries as some words might have multiple senses. As noted, this dataset is much larger than the one from WordNet (Ishiwatari et al., 2019).

Both datasets are already split into train, validation and test splits. The same splits are preserved in the conducted experiments in this work. Table 6.1 presents the statistics of these datasets including their size and context length. Additionally, Table 6.2 lists examples from the datasets showing the targeted word to be defined, its example sentence (context) and the provided definition. As shown in the table, some words might have different senses according to their usage in the context of the example sentence.

Table 6.1: Statistics of the context-aware datasets of definitions from the Oxford Dictionary (Gadetsky et al., 2018) and WordNet (Ishiwatari et al., 2019).

Dataset	Split	#Target Words	#Entries	Target Length	Context Length	Definition Length
WordNet	Train	7,938	13,883	1.00	5.81	6.61
	Valid	998	1,752	1.00	5.64	6.61
	Test	1,001	1,775	1.00	5.77	6.85
The Oxford dictionary	Train	33,128	97,855	1.00	17.74	11.02
	Valid	8,867	12,232	1.00	17.80	10.99
	Test	8,850	12,232	1.00	17.56	10.95

Table 6.2: Examples of instances from the context-aware datasets of definitions from the Oxford Dictionary (Gadetsky et al., 2018) and WordNet (Ishiwatari et al., 2019) showing the targeted word to be defined, the example sentence and the corresponding definition.

Dataset	Target Word	Example Sentence	Definition
WordNet	vigilance	vigilance is especially susceptible to fatigue	the process of paying close and continuous attention
	wonder	I wonder whether this was the right thing to do	to place in doubt or express doubtful speculation
	read	read the advertisement	to interpret something that is written or printed
	bright	the sun was bright and hot	emitting or reflecting light readily or in large amounts
	easy	an easy job	posing no difficulty
	easy	an easy job	requiring little effort
	easy	knowing that I had done my best, my mind was easy	free from worry or anxiety
The Oxford Dictionary	contain	she shouted at him, barely containing herself.	control or restrain (oneself or a feeling)
	break	one table had an older family, taking a break from cooking at home.	an interruption of continuity or uniformity
	break	then on monday schools in the paris region returned from their easter break, and young students marched out of classes in their tens of thousands.	a short holiday
	break	she was going to prove he hadn't broken her spirit.	crush the emotional strength, spirit, or resistance of
	renew	downslope, a patch of creeping red fescue grows naturally, requiring mowing just once or twice a year to renew growth.	give fresh life or strength to

2 https://developer.oxforddictionaries.com

6.3.2.2 *Experimental Setup*

HYPER-PARAMETERS: The proposed architecture utilises a dual encoder for the context and the target word. Both encoders utilise a single-layer LSTM with 256 hidden units. The word embeddings layer is initialised with the pre-trained GloVe embeddings. As mentioned earlier, I used the 300-dimensional GloVe embeddings pre-trained on the Common Crawl dataset. For the sentence embeddings, I employed ELMo 1,024-dimensional pre-trained embeddings from the TensorFlow Hub[3]. I did not update the weights of the GLoVe or ELMo embeddings during training. Table 6.3 summarises the employed hyper-parameters.

Table 6.3: Specification of the proposed attention-based sequence-to-sequence models for definition modelling based on a dual encoder architecture.

Model	Layer	Type and #layers	Hidden Units
Word-Character Dual Encoder	Context Embeddings	GLoVe	300
	Context Encoder	Single-Layer LSTM	256
	Target Embeddings	Character (One-Hot)	45
	Target Encoder	Single-Layer LSTM	256
	Definition Embeddings	GLoVe	300
	Definition Decoder	Single-Layer LSTM	256
Word-Contextual Dual Encoder	Context Embeddings	GLoVe	300
	Context Encoder	Single-Layer LSTM	256
	Target Embeddings	ELMo (Sentence Embeddings)	1,024
	Target Encoder	-	-
	Definition Embeddings	GLoVe	300
	Definition Decoder	Single-Layer LSTM	256

The Adam algorithm (Kingma and Ba, 2015) is used for optimisation during the training phase and sparse cross-entropy as a loss function to fine tune the network. The reported results are obtained using batch size of 128 instances for both the WordNet and the Oxford Dictionary datasets. In all experiments, I zero-pad the input sentences and the definitions to the longest sequence length in the dataset. All the hyper-parameters were optimised on the corresponding validation set of each dataset by assessing the accuracy. The models are implemented using Keras (Chollet, 2015) with the TensorFlow backend. The source code and best models will be publicly available[4]. All experiments are done on a NVIDIA Quadro M2000M GPU of 4GB memory. The average running time for the proposed models is around two hours on the WordNet dataset and 10 hours on the Oxford Dictionary dataset, for maximum of 50 epochs.

EVALUATION METRICS: I employ the bilingual evaluation understudy (BLEU) (Papineni et al., 2002) metric to assess the performance of the proposed approach on the metaphor interpretation dataset. This metric is used widely to assess language generation tasks such as machine translation. The idea behind it is that it automatically compares the generated text against a set of gold reference text. **BLEU** focuses on calculating the geometric mean of the overlapping n-gram

3 https://www.tensorflow.org/hub
4 https://github.com/OmniaZayed/

precision. A weight is assigned for each n-gram precision, which could be either set to 1 or equally distributed for higher n-grams. The *BLEU* score can be calculated as follows:

$$BLEU{-}n = \min\left(1, \frac{\text{generated text length}}{\text{reference length}}\right) \sum_{1}^{n} precision_i \tag{6.10}$$

where n is the number of overlapping tokens and it can range from 1 to 4. The score value can range from 0 to 1 but usually it is given as a percentage.

6.3.2.3 *Results*

Since the main aim of the work presented in this chapter is to interpret metaphors, I applied the proposed approach on the metaphor interpretation dataset that was prepared as part of this thesis and was introduced in Section 4.4. The dataset comprises around 1,500 metaphoric verb-direct object expressions and their contextual usage in tweets along with a manually annotated full explanation of each given metaphoric expression.

I conducted several experiments to better study the proposed approach and to asses its performance in interpreting metaphors in the informal context of tweets. First, I experiment with the proposed dual encoder architecture using character-level encoder to capture the surface information of the target word. This model, namely *Word-Character Dual Encoder*, is considered as a baseline. I then evaluate the model by employing context-dependent embeddings to encapsulate the target word before passing it to the encoder as discussed in Section 6.3.1. Since this model employs contextualised sentence embeddings, it is denoted as *Word-Contextual Dual Encoder*. I trained both models on the two context-aware datasets that were discussed earlier. Table 6.4 shows the performance of the trained models on the metaphor interpretation dataset. I also performed an in-domain evaluation by testing the models on the corresponding test split of the training dataset. I report *BLEU-1* and *BLEU-2* scores for unigrams and bigrams overlap, respectively. For that, I employed the *corpus-BLEU* from the NLTK library[5].

Table 6.4: Evaluation of the proposed attention-based sequence-to-sequence definition modelling approach that utilise a dual encoder. The models are trained on the Oxford Dictionary and WordNet datasets of definitions and tested on the corresponding test splits of the same data as well as the metaphor interpretation dataset.

Model	**Training**	**Test-sets**			
		in-domain		**metaphors**	
		BLEU-1	**BLEU-2**	**BLEU-1**	**BLEU-2**
Word-Character Dual Encoder	WordNet	14.81	4.33	16.52	5.67
	Oxford	10.21	2.96	11.45	3.36
Word-Contextual Dual Encoder	WordNet	15.63	5.84	24.18	6.99
	Oxford	11.81	1.13	18.71	4.45

5 https://www.nltk.org/_modules/nltk/translate/bleu_score.html

6.3.3 Discussion

Overall performance. As seen in Table 6.4, the model achieves quite low BLEU scores in the in-domain experiments on the corresponding test sets to the training data. However, these scores align with the ones obtained by similar approaches such as Ni and Wang (2017) and Ishiwatari et al. (2019) who approached definition modelling as a natural language generation task. As discussed earlier, the goal of this work is to interpret metaphors as they occur in text. The results on the metaphor interpretation dataset are promising and indicate that this approach can provide reliable interpretations of metaphoric expressions in a given context.

Error analysis. I did a manual analysis of the results to better understand the behaviour of the model. The inspection of a random sample suggests that some errors are due to predicting a definition for the wrong part-of-speech of the word. A few of the predicted definitions were semantically similar to the gold reference (as will be discussed in the next paragraph). Other errors were caused by redundancy, incomplete sentences and using the most common words in the dataset. The rest of the errors are due to a totally incorrect definition or predicting the definition of the word's antonym. Some of these findings align with the ones highlighted by Noraset et al. (2017). Another interesting issue is that some erroneously generated definitions make sense from a human (sarcastic) point of view. For example, the model defined the word *"healthy"* in the sentence *"a healthy diet"* as *"lacking in flavor"* instead of *"promoting health"*.

The limitations of the evaluation scheme. Despite the overall low performance under the adopted evaluation scheme, the proposed model was able to generate semantically similar definitions to the gold references. Table 6.5 shows examples of the predicted definitions by the *Word-Contextual Dual Encoder* model. In the first example, the predicted word *"stark"* is a synonym to the word *"resolute"* in the gold definition and both words can be considered definitions of the target word *"stoutly"*. The rest of the examples could be considered as possible explanations of the corresponding target words. This manual analysis of the predicted definitions highlighted the limitations of the adopted evaluation metrics which consider semantically similar definitions as a mismatch. An evaluation scheme that considers matching synonyms and semantically similar generated text will better suit assessing this task. One idea could be to represent all gold definitions using their contextualised sentence representations and then assess the semantic similarity between the predicted definition and the gold references of the target word (sense) in this mapped semantic space. This is certainly an issue that I am going to explore in future work.

Training data type. As noted from the results, although these models are trained on a non-figurative dataset, they were able to perform relatively well on interpreting metaphor expressions despite the subjectivity of the task. Table 6.6 shows examples of the predicted definitions of some metaphoric expressions from the metaphor interpretations datasets. Some of the predicted definitions are reliable while the others are erroneous. As a future direction, I am planning to train the proposed model on a figurative dataset of idiomatic definitions which can be obtained from Wiktionary. As discussed in 4.4.1, Wiktionary[6] has a large set of idioms along with their corresponding definitions under the *English Idioms Category*[7]. Furthermore, it will be interesting to evaluate the performance of the proposed approach on defining slang and uncommon expressions by utilising the Urban Dictionary dataset (Ni and Wang, 2017) as well as the Hei++ dataset of uncommon adjective-noun phrases (Bevilacqua et al., 2020).

6 https://www.wiktionary.org
7 https://en.wiktionary.org/w/index.php?title=Category:English_idioms

Table 6.5: Examples of the generated definitions by the proposed model from the WordNet dataset showing the target word, the example sentence and the gold reference definition.

Target	Context	Gold/Predicted Definitions
stoutly	he was stoutly replying to his critics	**Gold:** in a resolute manner **Predicted:** in a stark manner
morbid	morbid interest in death	**Gold:** suggesting an unhealthy mental state **Predicted:** the state of being free from emotion
stir	stir the soup	**Gold:** to move an implement through **Predicted:** to cook or cook
decided	The case was decided	**Gold:** to bring to an end **Predicted:** to go in a certain way
wonder	I wonder whether this was the right thing to do	**Gold:** to place in doubt or express doubtful speculation **Predicted:** to make a certain position or doubt of
swallow	I can not swallow these lies any more	**Gold:** believe or accept without questioning or challenge **Predicted:** to distinguish the meaning of something
kernel	a kernel of corn	**Gold:** a single whole grain of a cereal **Predicted:** a small indefinite number or amount or extent
mysterious	mysterious symbols	**Gold:** beyond ordinary understanding **Predicted:** not clearly

Table 6.6: Examples of the generated definitions by the proposed model from the metaphor interpretation dataset showing the target metaphoric expression, the example tweet and the gold reference definition.

Target	Context	Gold/Predicted Definitions
speak volumes	his actions here will speak volumes to how safe he wants us to be [...]	**Gold:** To make something very obvious **Predicted:** to express or express
breaks my heart	nothing breaks my heart more than seeing a person looking into the mirror with anger & disappointment, blaming themselves when someone left.	**Gold:** to make somebody feel so sad, lonely, etc. that they cannot live a normal life **Predicted:** to make a certain condition
puts fear	no bitch or nigga puts fear in me dawg & idgaf if we bleed the same blood	**Gold:** to make somebody/something feel something or be affected by something **Predicted:** to be frightened
pump out tax payer	both sides were allowed to pump out tax payer - funded propaganda, you idiot. #euref	**Gold:** to use the funds of the people **Predicted:** to make a place
resist the temptation	#ivoted just about managed to resist the temptation to vote for a #brexit	**Gold:** to stop yourself from having something you like or doing something you very much want to do **Predicted:** to destroy or manage to manage
framed this debate	@jasoncowleyns #euref is about the winners and losers of globalisation . ukip's great success was how they framed this debate around eu.	**Gold:** to move or put an argument in a certain context **Predicted:** in a manner or miserable manner

Aspects of meaning. Although the proposed approach was able to generate interpretations of metaphors in a given context, as shown in Table 6.6, it focused on capturing the literal meaning of the expression and into some respect ignored the pragmatic aspect of meaning. The usage, and hence the interpretation, of metaphors includes various social, emotional and pragmatic aspects that contribute to the process of understanding their meaning which goes beyond the literal meaning (Gibbs et al., 2011). These aspects are referred to as "utterance meaning" by Searle (1993). It is important to consider both the literal (word) meaning and the speaker's utterance meaning (as well as the relation between them) while interpreting metaphors. It is quite challenging for humans, let alone computational models, to capture these aspects which help in the overall understanding and interpretation of a given metaphoric expression. I am planning to perform a deeper investigation of these aspects focusing on the tweets dataset in future work.

6.4 SUMMARY

This chapter discussed the work done in this thesis to address the last research theme that focuses on interpreting linguistic metaphors as they occur in text. The main goal of this work was to explore an neural approach to generate reliable interpretations of metaphoric expressions with the aim of aiding language learners and enriching lexical resources. The chapter discussed the task formulation as a definition modelling (generation) task after recapping how previous works addressed metaphor interpretation and the motivation behind my choice of the task formulation.

I started this work by studying definition modelling and how it was approached in the literature. As discussed at the beginning of this chapter, the majority of previous approaches addressed definition modelling as conditional language modelling. In this formulation the goal is to generate a contextually appropriate definition given a target word that needs to be defined. Therefore, the majority of the developed systems utilised either RNN-based or sequence-to-sequence-based language models. Several approaches have proposed methods to incorporate the surrounding context around the word being defined that represents its usage (sense). More recently, a line of research emerged that treats the task as a definition classification task.

In this work, I proposed a neural approach based on sequence-to-sequence language modelling that utilises a dual encoder architecture and contextualised sentence embeddings to represent the target word or expression to be defined. I applied the proposed model to interpret metaphors in tweets. The conducted experiments showed that the utilisation of the contextualised sentence embeddings to encapsulate the target expression was effective in dealing with multi-word expressions as it allows the model to capture the interaction between the words of the target expression. An analysis was conducted to better understand the system limitations and to identify the common errors and flaws of the proposed models. This was done by inspecting a random sample of the generated data. This analysis also highlighted the limitations of the adopted evaluation scheme which calls for a more semantically driven approach to assess the performance of such models. This work opens interesting avenues for future work either for definition modelling in general or for metaphor interpretation defined as definition generation in specific. This includes improving the proposed approach and its ability to generate more reliable definitions either for figurative or non-figurative language. There is also room for further investigation into improving the evaluation scheme to better suit the definition modelling task in general.

7 CONCLUSIONS

"In literature and in life we ultimately pursue, not conclusions, but beginnings."

Sam Tanenhaus, Literature Unbound

This thesis explored metaphor processing in tweets focusing on two main tasks which are metaphor identification and interpretation. Three research themes have guided the work presented in this thesis which are: *Resource Preparation*, *Metaphor Identification in Tweets* and *Metaphor Interpretation in Tweets*. These research themes and the research questions that they covered were presented in detail in Chapter 1. The following sections conclude the work done under each research theme and highlight the open questions as well as the future directions of this thesis.

7.1 GENERAL CONCLUSIONS AND CONTRIBUTIONS

This thesis began by studying and reviewing the previous research pertaining to processing linguistic metaphors in text with the aim to provide a better understanding of the task as well as identifying possible research opportunities and highlighting previous limitations. Chapter 3 explained the different tasks of processing metaphors in text and gave a detailed explanation of the main tasks of interest in this thesis, namely metaphor identification and interpretation. The chapter then traced the development of metaphor processing during the past few decades highlighting the change in concerns regarding the adopted processing paradigms and theories of metaphor and how this affected the choice of the employed approaches and the selection of features. A detailed overview of the state-of-the-art approaches and techniques which support 1) the automatic identification and interpretation of linguistic metaphors, 2) the development of metaphor annotation schemes and the preparation of resources and datasets for both tasks was presented. More specifically, this overview provided extensive details about existing datasets for metaphor identification focusing on English text and discussed the annotation scheme, type of metaphor and level of processing for each dataset. An extensive literature review of the various approaches pertaining to metaphor identification over the past few decades was done, from a chronological perspective, highlighting the various adopted paradigms to process metaphors on the sentence, relation, and word levels in order to investigate how these paradigms affected the choice of approaches, developed architectures and selected features. Regarding metaphor interpretation, the literature review also discussed how previous approaches categorised the task as either lexical substitution, paraphrase generation or definition generation and how this task categorisation affected the preparation of the datasets.

7.1.1 Resource Preparation

A major part of the work done in this thesis was devoted to developing metaphor datasets with two main goals 1) creating new gold-standard datasets for metaphor identification and its interpretation focusing on tweets; 2) adapting and improving existing benchmark datasets annotated for linguistic metaphors. The main scope of this thesis is on identifying and thus annotating linguistic metaphors in English text on the relation level focusing on verb-noun and adjective-noun grammar relations. Once the identification of metaphors is done, the work proceeded with obtaining their interpretation by formulating the task as definition generation. Chapter 4 discussed the preparation of the aforementioned metaphor datasets of tweets. The chapter began by discussing the main challenges that face research into metaphor processing in terms of corpora preparation. Then, it went on to present the proposed approaches to address these challenges by providing details on the proposed annotation methodologies to identify and interpret metaphors.

The chapter first discussed the proposed annotation methodology to create a dataset of tweets annotated to identify linguistic metaphors. The methodology was developed with the aim of reducing the cognitive load on the annotators and maintaining consistency. Although the methodology is employed to annotate linguistic metaphors of the predicate type (i.e. verb-direct object pairs) in tweets, it can be applied to any text type, metaphor type or level of analysis. The tweets selection process was driven by achieving high accuracy, sense coverage and verbs representativeness. The resulting metaphor dataset consists of various topic genres focusing on tweets of general topics and political tweets related to Brexit. A substantial inter-annotator agreement was achieved among five annotators, who are native speakers of English, despite the difficulty of defining metaphor, the conventionality of metaphors, and the noisy nature of the user-generated text of tweets.

The chapter then discussed the work done to adapt word-level benchmark datasets to suit relation-level metaphor identification. This step was essential towards filling the gap of the availability of large benchmark datasets for relation-level metaphor processing in English. A semi-automatic approach was employed to adapt the existing benchmark datasets including the most well-known and widely used corpus for metaphor identification, namely the VU Amsterdam Metaphor Corpus (VUAMC), to better suit identifying metaphors on the relation level avoiding the need for extensive manual annotation.

Finally, the chapter concluded by presenting the work done on creating the first gold-standard dataset for metaphor interpretation along the more complex "definition generation" approach which provides full explanation of a given metaphoric expression. The methodology of preparing the dataset was demonstrated which combines an automatic retrieval approach with manual annotation to ensure wide coverage, accuracy and consistency. As a result, around 1,500 metaphoric verb-direct object expressions in tweets were annotated. The methodology and annotation scheme can be generalised to annotate metaphors of any syntactic structure in any text genre/type. The proposed resources from this thesis will be published[1] in order to facilitate research on metaphor processing in general and in tweets specifically.

1 This will be done by following the copyrights and licensing types of each of them.

7.1.2 Metaphor Identification in Tweets

Chapter 5 turned to the second research theme covered in this thesis that is pertaining to metaphor identification with specific interest in the computational modelling of metaphors in tweets. The chapter discussed three main NLP and deep learning approaches to identify metaphors on the relation level which are distributional semantics, meta-embeddings learning and contextual modulation. The work done under this theme started by investigating distributional approaches to metaphor identification with the aim to design and employ a minimally supervised one to aid in the data annotation process. After that it was important to study the various features employed in the literature to identify metaphors in text. Therefore, the second part of Chapter 5 investigated meta-embedding learning methods in order to study the effectiveness of an ensemble of features to identify metaphoric expressions on the relation level. Finally, inspired by works in visual reasoning, I proposed a novel approach for context-based textual classification that utilises affine transformations. I applied this approach that is based on contextual modulation to identify metaphoric expressions focusing on verb-noun and adjective-noun dependency relations in tweets.

In this work, I first investigated the feasibility of introducing a minimally supervised approach based on distributional semantics to identify linguistic metaphors with the aim of aiding in the creation and annotation process of a metaphor dataset of tweets. I explored the use of different pre-trained word embedding models to identify relation-level metaphors focusing on verb-noun expressions. The proposed approach, namely *DistSemant*, employs a predefined seed set of metaphoric expressions to classify new unseen ones that are highlighted in a given sentence based on semantic similarities between verbs and nouns. The proposed approach is flexible and can be easily adapted to new domains or text types. Experiments showed that it also generalises better when compared to related minimally supervised approaches since it employs fewer lexical resources and does not require annotated datasets or highly-engineered features.

I then studied the various features that have been introduced to identify metaphors in text by previous approaches under a unified neural architecture. I investigated meta-embedding learning methods to study the effectiveness of semantic and psycholinguistic features in conjunction with deep contextualised features to identify relation-level metaphors in tweets. The effect of each of the proposed features is studied individually and collectively using two strategies of creating feature ensembles which are concatenation (Concat) and dynamic meta-embeddings (DME). The analysis of the proposed models revealed that employing well-established linguistic features for metaphor identification in a neural architecture along with the advanced deep contextualised features led to a significant improvement over the current work on the metaphor dataset of tweets. The proposed DME model allowed the network to automatically select different embeddings based on the training data. However, one limitation of this model is that it requires more training data which is one of the challenges that faces the metaphor processing research.

Finally, I looked at contextual modulation, specifically through affine transformations, which have been shown to be a powerful device for relational problems in both the field of NLP and visual reasoning. The motivation behind this idea is to exploit the interaction between the metaphor components as well as the context (e.g. a sentence/tweet) to inform the decision regarding a specific relation in the text. Based on this idea, I therefore introduced a novel architecture to identify metaphors by utilising feature-wise affine transformation and deep contextual modulation. The approach is applied to relation-level metaphor identification to

classify expressions of certain syntactic constructions for metaphoricity as they occur in context. It significantly outperformed the state-of-the-art approaches for this level of analysis on benchmark datasets. To the best of my knowledge, this is the first approach to study the employment of feature-wise linear modulation on metaphor identification in general. The proposed methodology is generic and can be applied to a wide variety of text classification approaches including sentiment analysis or term extraction.

7.1.3 Metaphor Interpretation in Tweets

Finally, Chapter 6 explored the feasibility of employing advanced neural models to automatically interpret and explain the intended meaning of a metaphor. In the context of this thesis, metaphor interpretation is viewed as a definition generation task with the aim to aid language learners and non-native speakers to understand metaphors as well as enrich the process of developing lexical resources. I therefore investigated the feasibility of such task formulation by studying definition modelling. I then proposed a neural approach based on sequence-to-sequence language modelling and applied it to interpret linguistic metaphors of the predicate type in tweets.

The work under this research theme began by studying the previous research pertaining to definition generation in general. I then proposed a neural approach based on sequence-to-sequence language modelling that utilises a dual encoder architecture and contextualised sentence embeddings to represent the target word or expression to be defined. The proposed model was applied to interpret metaphors in tweets. Experiments were conducted to investigate the effectiveness of the proposed approach to obtain sense specific definitions for metaphoric expressions. The conducted experiments showed that the utilisation of the contextualised sentence embeddings to encapsulate the target expression was effective in dealing with multi-word expressions as it allows the model to capture the interaction between the words of the target expression. An analysis was conducted to better understand the system limitations and to identify the common errors and flaws of the proposed models.

7.2 LIMITATIONS AND LESSONS LEARNT

This section reviews some of the identified limitations of the proposed work in this thesis as well as lessons learnt.

7.2.1 Metaphor Dataset Annotation

The effect of the cultural background on the annotation of metaphor. The choice of the annotators as well as the data source have an effect on annotation process from a cultural perspective. The pilot study, introduced in Section 4.2.1.2, revealed that the cultural background of the annotators affected the annotation process that was based on the annotators' intuition of what a metaphor is. In order to avoid this effect, I controlled the main annotation experiment by hiring annotators with the same nationality. Furthermore, the data source could have an implicit cultural bias. For instance, the majority of tweets about a certain topic are expected to be drawn from a specific country such as the tweets about Brexit.

Annotation guidelines. The strict following of the guidelines by the annotators led them to focus on identifying the metaphoricity of a given expression based on its abstractness/concreteness nature. This lead to some disagreements between the annotators around some verbs such as *"enjoy, imagine,* and *remember"*.

Dataset balance. One question could be raised is the effect of creating a balanced dataset on the development of the computational models. In this work, one of the considered factors during the creation of the ZayTw dataset was to ensure balance in order to avoid bias towards a certain class while training the identification model which might lead to poor minority class classification performance. However, the real distribution of metaphors is not balanced, with metaphors being the minority class of interest. The distribution of the data have a big impact on the computational task and should be considered when creating a dataset.

7.2.2 Metaphor Identification

The application of the proposed approaches on larger contexts. One of the open questions about the proposed approaches for metaphor identification in this thesis is their applicability to longer texts. The main focus of the work presented here is to process metaphor in short texts (specially in the context of tweets). The adaptation of the proposed approaches will be required to include broader context beyond the single sentence or tweet when applying the proposed metaphor identification approaches on longer documents.

7.2.3 Metaphor Interpretation

Modelling the different aspects of meaning when interpreting a metaphor The proposed approach to interpret metaphor as definition generation does not take into account the pragmatic aspect when inferring the intended meaning of the metaphoric expression. Since metaphor is a contextual and situational phenomenon it conveys the intention of the author as well as the effect on the listener. Computational models should consider these aspects of meaning as well as the literal aspect that contribute to the process of understanding and comprehending metaphors while interpreting or explaining a given metaphoric expression in a given context.

7.3 POSSIBLE APPLICATIONS AND USE CASES

This thesis focused on introducing computational models for metaphor processing in social media text (particularly tweets). The motivation behind the work introduced in this thesis is to improve real-world applications that can benefit from metaphor processing. Examples of practical applications that automatic metaphor identification enables include analysing healthcare communication and political discourse in social media. People tend to employ metaphoric expressions when talking about certain political topics such as Brexit. Metaphoric language can also be employed by social media users when talking about health-related issues such as cancer. Identifying the metaphoric expressions in such discourse will allow sociologists, psychologists and political scientists to scientifically study and explore metaphoric usage on social media which will allow a deeper understanding of the language on many computational and linguistic

levels. On the other hand, the automatic interpretation of metaphors can be of benefit to non-native speakers and language learners by providing them with explanations (interpretations) of metaphoric expressions that might not be found in traditional dictionaries. Furthermore, it could help in developing and enriching lexical resources.

7.4 OPEN QUESTIONS AND FUTURE DIRECTIONS

This section summarises the open questions that this thesis revealed and the future directions and recommendations to address them.

7.4.1 Metaphor Datasets and Resources

A lot of effort is still required to develop resources for metaphor processing as well as improving and enriching the existing ones. As discussed in Chapter 4, annotating datasets for metaphors requires extensive annotation efforts, time and money. I am planning to extend the current work on preparing metaphor datasets of social media content. To this end, I am planning to explore semi-supervised approaches to extend the metaphor dataset of tweets proposed in this work. Another issue that this thesis highlighted is the different levels of metaphor annotation either on the sentence, relation or word level with the majority of approaches focusing on the last level. I will build upon the proposed effort to adapt word-level benchmark datasets to suit the relational approaches. However, I will shift my focus to find a common ground between the two paradigms with the aim to introduce a unified evaluation platform.

The work done to adapt the well-known VU Amsterdam Metaphor corpus (VUAMC) revealed some annotation inconsistencies and questioned the quality of the existing annotation which was done on the word-level and is based on the annotator's intuition to obtain the sense of the word from dictionaries. Since this is a very useful resource for metaphor identification it is better to address these issues and refine it. One idea is that instead of employing the dataset to evaluate the proposed approach it could be interesting to look at how can the current state-of-the-art approaches help in improving the annotation of existing datasets. For example, having a hybrid model where the metaphoricity classification is taken based on the outputs from the multiple approaches. This will allow the utilisation of the strengths of existing approaches to verify the quality of the manual annotations.

A more long-term future work regarding metaphor datasets that requires the collaboration of the scientific community in this area stems from the fact that the majority of researchers carried out their annotations by leveraging their in-house team of annotators, with various backgrounds and expertise, and only few employed crowd-sourcing platforms. Additionally, many researchers developed their own annotation procedures depending on the type of metaphor and the level of processing. Further, the majority relied on the annotator's intuition to define metaphor. These variations in metaphor corpora design considerations pose a limitation on cross-systems comparisons and the possibility of a unified performance evaluation and interpretation. This calls for a large-scale annotation effort that could be inspired by the PARSEME shared-task (Walsh et al., 2018) on automatic identification of verbal multi-word expressions (MWEs).

7.4.2 Metaphor Identification

The main issue that I am interested to explore as a future direction under the theme of metaphor identification is to bridge the gap between the different levels of analysis. Having a unified evaluation scheme that can fairly evaluate and compare the performance of the proposed approaches under both the relational and word-level paradigms is necessary and will facilitate future research in this area.

7.4.3 Metaphor Interpretation

As discussed in Chapter 6, there is room for investigation into the formulation of metaphor interpretation as definition generation. An open question is how to adapt current definition generation models to suit the task of metaphor interpretation. I am planning to build upon the proposed approach to include other features such as the part-of-speech tag of the word being defined or the head of the expression in the case of multi-word expressions. Also, I am interested in investigating other features such as employing the hypernym relation between the metaphor components in the definition generation process. An interesting idea, inspired by the work done on contextual modulation to identify metaphors in this thesis, is to employ affine transformation to modulate the interaction between the metaphor expression to be defined and the definition generation process.

Another direction of future work that could be investigated under this research theme is related to the generated definition itself. One of the issues that the investigation of this theme revealed is how to deal with the generated definitions of a figurative nature. An approach is required to first identify the expressions that have figurative interpretations and then redefine them using more literal language. Finally, another avenue to explore is the possibility of improving the evaluation scheme to address semantically similar definitions. The current evaluation scheme considers synonymous and semantically similar definitions as a mismatch. One idea could be to represent all gold definitions using their contextualised sentence representations and then assess the semantic similarity between the predicted definition and the gold references of the target word (sense) in this mapped semantic space.

REFERENCES

James F. Allen, Mary Swift, and Will de Beaumont. 2008. Deep semantic analysis of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, page 343–354, Venice, Italy. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR 2015.

Cong Bao and Danushka Bollegala. 2018. Learning word meta-embeddings by autoencoding. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 1650–1661, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Eduard Barbu, Maria Martín-Valdivia, Eugenio Martínez-Cámara, and L. López. 2015. Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42:5076–5086.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Cambridge, MA, USA. Association for Computational Linguistics.

Yosef Ben Shlomo and Mark Last. 2015. MIL: automatic metaphor identification by statistical learning. In *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing*, volume 1410 of *DMNLP '15*, pages 19–29, Porto, Portugal.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 7207–7221, Online. Association for Computational Linguistics.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, NM, USA. Association for Computational Linguistics.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '06, pages 329–336, Trento, Italy.

Julia Birke and Anoop Sarkar. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28, Rochester, NY, USA.

Yuri Bizzoni, Stergios Chatzikyriakidis, and Mehdi Ghanimifard. 2017. "deep" learning : Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52, Copenhagen, Denmark. Association for Computational Linguistics.

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and BiLSTMs two neural networks for sequential metaphor detection. In *Proceedings of the First Workshop on Figurative Language Processing*, pages 91–101, New Orleans, LA, USA.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the First Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Max Black. 1962. *Models and metaphors: Studies in language and philosophy*. Cornell University Press, Ithaca, NY, USA.

Max Black. 1993. More about metaphor. In Andrew Ortony, editor, *Metaphor and thought*, 2nd edition, pages 19–41. Cambridge University Press, Cambridge, UK.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, page 89–97, Beijing, China. Association for Computational Linguistics.

Danushka Bollegala, Kohei Hayashi, and Ken-Ichi Kawarabayashi. 2018. Think globally, embed locally: locally linear meta-embedding of words. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI '18, pages 3970–3976.

Danushka Bollegala and Ekaterina Shutova. 2013. Metaphor interpretation using paraphrases extracted from the web. *PLoS ONE*, 8(9):1–10.

Marianna Bolognesi, Romy van den Heerik, and Esther van den Berg. 2018. VisMet 1.0: An online corpus of visual metaphors. In *Visual metaphor: Structure and Process*, pages 89–114. John Benjamins, Amsterdam.

David B. Bracewell, Marc T. Tomlinson, Michael Mohler, and Bryan Rink. 2014. A tiered approach to the recognition of metaphor. *Computational Linguistics and Intelligent Text Processing*, 8403:403–414.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1. *Linguistic Data Consortium*.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, volume Interactive Presentation Sessions of *COLING-ACL '06*, pages 77–80, Sydney, Australia. Association for Computational Linguistics.

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 102–110. Springer.

George Aaron Broadwell, Jennifer Stromer-Galley, Tomek Strzalkowski, Samira Shaikh, Sarah Taylor, Ting Liu, Umit Boz, Alana Elia, Laura Jiao, and Nick Webb. 2012. Modeling sociocultural phenomena in discourse. *Natural Language Engineering*, 19(2):213–257.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pages 523–528, Valencia, Spain.

Lou Burnard. 2007. Reference guide for the British National Corpus (XML edition).

Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 2753–2765, Santa Fe, NM, USA. Association for Computational Linguistics.

Lynne Cameron. 2003. *Metaphor in educational discourse*. Advances in Applied Linguistics. Continuum, London, UK.

Lynne Cameron and Graham Low. 1999. *Researching and Applying Metaphor*. Cambridge Applied Linguistics. Cambridge University Press, Cambridge, UK.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *arXiv preprint*, arXiv:1803.11175.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Ting-Yun Chang, Ta-Chung Chi, Shang-Chi Tsai, and Yun-Nung Chen. 2018. xSense: Learning sense-separated sparse representations and textual definitions for explainable word sense networks. *arXiv preprint*, arXiv:1809.03348.

Jonathan Charteris-Black. 2011. Metaphor in Political Discourse. In *Politicians and Rhetoric: The Persuasive Power of Metaphor*, pages 28–51. Palgrave Macmillan UK, London.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *The 15th Annual Conference of the International Speech Communication Association*, INTERSPEECH '14, pages 2635–2639, Singapore.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 740–750, Doha, Qatar.

Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, Online. Association for Computational Linguistics.

François Chollet. 2015. Keras. https://github.com/keras-team/keras.

Yulia Clausen and Vivi Nastase. 2019. Metaphors in text simplification: To change or not to change, that is the question. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 423–434, Florence, Italy.

Joshua Coates and Danushka Bollegala. 2018. Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2: Short Papers of *NAACL '18*, pages 194–198, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33A(4):497–505.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Seana Coulson. 2009. Metaphor and conceptual blending. In J. L. Mey, editor, *Concise Encyclopedia of Pragmatics*, 2nd edition, Encyclopedia of Language and Linguistics, pages 615–621. Oxford: Elsevier.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

David Crystal. 2008. *A Dictionary of Linguistics and Phonetics*. Blackwell Publishing.

Verna Dankers, Karan Malhotra, Gaurav Kudva, Volodymyr Medentsiy, and Ekaterina Shutova. 2020. Being neighbourly: Neural metaphor identification in discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 227–234, Online. Association for Computational Linguistics.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

Donald Davidson. 1978. What metaphors mean. *Critical Inquiry*, 5(1):31–47.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, volume 6 of *LREC '06*, pages 449–454, Genoa, Italy.

Alice Deignan. 2005. *Metaphor and Corpus Linguistics*. John Benjamins.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805.

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018a. Killing four birds with two stones: Multi-task learning for non-literal language detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLing '18, pages 1558–1569, Santa Fe, NM, USA. Association for Computational Linguistics.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, CA, USA.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018b. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 1412–1424, Brussels, Belgium.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. 2018. Feature-wise transformations. *Distill*. Https://distill.pub/2018/feature-wise-transformations.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Jonathan Dunn. 2013a. Evaluating the premises and results of four metaphor identification systems. In *Proceedings of the 14th International conference on Computational Linguistics and Intelligent Text Processing*, volume 7816 of *CICLing '13*, pages 471–486, Samos, Greece. Springer Berlin Heidelberg.

Jonathan Dunn. 2013b. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, Atlanta, GA, USA. Association for Computational Linguistics.

Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Laurel J. End. 1986. Grounds for metaphor comprehension. In I. Kurcz, G. W. Shugar, and J. H. Danks, editors, *Knowledge and language*, volume 39 of *Advances in Psychology*, pages 327–345. Elsevier Science (North-Holland).

David Evans. 2007. Compiling a corpus. [online - accessed December 23, 2018].

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, 16:235–250.

John R. Firth. 1957. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Winthrop Nelson Francis and Henry Kucera. 1979. The Brown Corpus: A standard corpus of present-day edited American English. Technical report, Brown University Liguistics Department.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2: Short Papers of *ACL '18*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Lisa Gandy, Nadji Allan, Mark Atallah, Ophir Frieder, Newton Howard, Sergey Kanareykin, Moshe Koppel, Mark Last, Yair Neuman, and Shlomo Argamon. 2013. Automatic identification of conceptual metaphors with limited knowledge. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, AAAI '13, page 328–334, Bellevue, WA, USA.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 1412–1424, Brussels, Belgium.

Andrew Gargett and John Barnden. 2015. Modeling the interaction between sensory and affective meanings for detecting metaphor. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 21–30, Denver, CO, USA. Association for Computational Linguistics.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, ScaNaLU '06, pages 41–48, New York City, NY, USA.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155 – 170.

Dedre Gentner, Brian Bowdle, Phillip Wolff, and Consuelo Boronat. 2001. Metaphor is like analogy. In D. Gentner, K. J. Holyoak, and B. N. Kokinov, editors, *The analogical mind: Perspectives from cognitive science*, pages 199–253. The MIT Press, Cambridge, MA, USA.

Dedre Gentner and Catherine Clement. 1988. Evidence for relational selectivity in the interpretation of analogy and metaphor. In Gordon H. Bower, editor, *Psychology of Learning and Motivation: Advances in research and theory*, volume 22, pages 307–358. Academic Press, New York, NY, USA.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015a. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 470–478, Denver, CO, USA.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015b. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Lisbon, Portugal. Association for Computational Linguistics.

Raymond Gibbs, Markus Tendahl, and Lacey Okonski. 2011. Inferring pragmatic messages from metaphor. *Lodz Papers in Pragmatics*, 7(1):3–28.

Raymond W. Gibbs. 1992. When is metaphor? the idea of understanding in theories of metaphor. *Poetics Today*, 13(4):575–606.

Raymond W. Gibbs. 1999. Researching metaphor. In Lynne Cameron and Graham Low, editors, *Researching and Applying Metaphor*, Cambridge Applied Linguistics, page 29–47. Cambridge University Press, Cambridge, UK.

Andrew Goatly. 1997. *The Language of Metaphors*. Routledge, London, UK.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. IlliniMet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153, Online. Association for Computational Linguistics.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English Gigaword. *Linguistic Data Consortium*.

Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. 2017. Stance and influence of twitter users regarding the Brexit referendum. *Computational Social Networks*, 4(6):1–25.

Elkin Darío Gutiérrez, Guillermo A. Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2923–2930, Copenhagen, Denmark.

Elkin Darío Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 183–193, Berlin, Germany.

Patrick Hanks. 2006. Metaphoricity is gradable. In Anatol Stefanowitsch and Stefan Th. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*, volume 171 of *Trends in Linguistics. Studies and Monographs [TiLSM]*, pages 17 – 35. De Gruyter Mouton, Berlin, Boston.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. The MIT Press.

Patrick Hanks. 2016. Three kinds of semantic resonance. In *Proceedings of the 17th EURALEX International Congress*, pages 37–48, Tbilisi, Georgia.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphors with lda topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, GA, USA.

Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with LCC's groundhog system. In *Proceedings of the 2nd PASCAL Challenges Workshop*, volume 18.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Janet Ho and Winnie Cheng. 2016. Metaphors in financial analysis reports: How are emotions expressed? *English for Specific Purposes*, 43:37 – 48.

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

Keith Holyoak and Dusan Stamenković. 2018. Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, 144:641—-671.

Matthew Honnibal. 2016. Embed, encode, attend, predict: The new deep learning formula for state-of-the-art nlp models. [online - accessed April 16, 2019].

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–56, Atlanta, GA, USA.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers) of *NAACL-HLT '19*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The tenten corpus family. In *Proceedings of the 7th International Corpus Linguistics Conference*, CL '13, pages 125–127, Lancaster, UK.

Anja Jamrozik, Eyal Sagi, Micah Goldwater, and Dedre Gentner. 2013. Relational words have high metaphoric potential. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 21–26, Atlanta, GA, USA. Association for Computational Linguistics.

Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rosé. 2016. Metaphor detection with topic transition, emotion and cognition in context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1 of *ACL '16*, pages 216–225, Berlin, Germany. Association for Computational Linguistics.

Hyeju Jang, Keith Maki, Eduard Hovy, and Carolyn Rosé. 2017. Finding structure in figurative language: Metaphor detection with topic-based frames. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 320–330, Saarbrücken, Germany. Association for Computational Linguistics.

Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rose. 2015a. Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '15, pages 384–392, Prague, Czech Republic.

Hyeju Jang, Mario Piergallini, Miaomiao Wen, and Carolyn Rosé. 2014. Conversational metaphors in use: Exploring the contrast between technical and everyday notions of metaphor. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 1–10, Baltimore, MD, USA.

Hyeju Jang, Miaomiao Wen, and Carolyn Rosé. 2015b. Effects of situational factors on metaphor detection in an online discussion forum. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 1–10, Denver, CO, USA. Association for Computational Linguistics.

Yael Karov and Shimon Edelman. 1998. Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–59.

Douwe Kiela, Changhan Wang, and Kyunghyun Cho. 2018. Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 1466–1477, Brussels, Belgium.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, pages 7–36.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, CA, USA.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, LREC '02, pages 1989–1993, Las Palmas, Canary Islands, Spain.

Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic bulletin and review*, 7(2):257–266.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, NIPS '15, pages 3294–3302, Montreal, Canada.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, GA, USA. Association for Computational Linguistics.

Beata Beigman Klebanov, Ben Leong, Michael Heilman, and Michael Flor. 2014a. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, Baltimore, MD, USA. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20, Denver, CO, USA. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutiérrez, Ekaterina Shutova, and Michael Flor. 2016a. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, Berlin, Germany.

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018a. A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers) of *NAACL-HLT '18*, pages 86–91, New Orleans, LA, USA. Association for Computational Linguistics.

Beata Beigman Klebanov, Ekaterina Shutova, and Patricia Lichtenstein, editors. 2014b. *Proceedings of the Second Workshop on Metaphor in NLP*. Association for Computational Linguistics, Baltimore, MD, USA.

Beata Beigman Klebanov, Ekaterina Shutova, and Patricia Lichtenstein, editors. 2016b. *Proceedings of the Fourth Workshop on Metaphor in NLP*. Association for Computational Linguistics, San Diego, CA, USA.

Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee Leong, editors. 2018b. *Proceedings of the First Workshop on Figurative Language Processing*. Association for Computational Linguistics, New Orleans, LA, USA.

Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee Leong, Anna Feldman, and Debanjan Ghosh, editors. 2020. *Proceedings of the Second Workshop on Figurative Language Processing*. Association for Computational Linguistics, Online.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume System Demonstrations of *ACL '17*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Arlene Koglin and Rossana Cunha. 2019. Investigating the post-editing effort associated with machine-translated metaphors: a process-driven analysis. *The Journal of Specialised Translation*, 31(01):38–59.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, SENSE '18, pages 24–30, Valencia, Spain.

Simon Krek, Iztok Kosem, John P McCrae, Roberto Navigli, Bolette S Pedersen, Carole Tiberius, and Tanja Wissik. 2018. European lexicographic infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress*, Ljubljana, Slovenia.

Klaus Krippendorff. 2004. Reliability in content analysis. *Human Communication Research*, 30:411–433.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Rochester, NY, USA.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *ICML'15*, pages 957–966, Lille, France.

George Lakoff, Jane Espenson, and Alan Schwartz. 1991. The master metaphor list. Technical report, University of California at Berkeley.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago, USA.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Duong Minh Le, My Thai, and Thien Huu Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, AAAI '20, New York, NY, USA.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, volume 32 of *ICML'14*, pages 1188—-1196, Beijing, China.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS '00, page 535–541, Denver, CO, USA.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the First Workshop on Figurative Language Processing*, pages 56–66, New Orleans, LA, USA.

Samuel R. Levin. 1977. *The semantics of metaphor*. Johns Hopkins University Press, Baltimore, MD, USA.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5:361–397.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7871–7880, Online. Association for Computational Linguistics.

Jiahuan Li, Yu Bao, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2020. Explicit semantic decomposition for definition generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 708–717, Online. Association for Computational Linguistics.

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 315–323, Singapore. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Hugo Liu and Push Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '14, pages 55–60, Baltimore, MD, USA.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 1222–1231, Melbourne, Australia.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3888–3898, Florence, Italy.

Katerina Margatina, Christos Baziotis, and Alexandros Potamianos. 2019. Attention-based conditioning methods for external knowledge integration. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3944–3951, Florence, Italy.

James H. Martin. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc., San Diego, CA, USA.

André Martins, Noah Smith, Eric Xing, Pedro Aguiar, and Mário Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 34–44, Cambridge, MA, USA.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Marina Meila and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*, AISTATS '01, Key West, FL, USA.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford: Oxford University Press.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*, ICLR '13, Scottsdale, AZ, USA.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, INTERSPEECH '10, pages 1045–1048, Makuhari, Chiba, Japan.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2 of *NIPS '13*, pages 3111–3119, Lake Tahoe, NV, USA.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, Plainsboro, NJ, USA.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, SemEval '18, pages 1–17, New Orleans, LA, USA.

Saif M. Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, *Sem '16, pages 23–33, Berlin, Germany.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013a. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35, Atlanta, GA, USA.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC '16, pages 4221–4227, Portorož, Slovenia.

Michael Mohler, Marc Tomlinson, and David Bracewell. 2013b. Applying textual entailment to the interpretation of metaphor. In *Proceedings of the 2013 IEEE 7th International Conference on Semantic Computing*, pages 118–125, Irvine, CA, USA. IEEE.

Alessandro Moschitti, Daniele Pighin, and Roberto Basili. 2006. Tree kernel engineering for proposition re-ranking. In *Proceedings of the International Workshop on Mining and Learning with Graphs*, MLG '06, pages 165–172, Berlin, Germany.

Jesse Mu, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Learning outside the box: Discourse-level features improve metaphor identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, Minneapolis, MN, USA.

Srinivas Narayanan. 1999. Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of the 16th AAAI Conference on Artificial Intelligence*, AAAI '99, Orlando, FL, USA.

Srinivas Sankara Narayanan. 1997. *Knowledge-based Action Representations for Metaphor and Aspect (KARMA)*. Ph.D. thesis, UNIVERSITY of CALIFORNIA at BERKELEY, Berkeley, CA, USA.

Roberto Navigli and Federico Martelli. 2019. An overview of word and sense similarity. *Natural Language Engineering*, 25(6):693–714.

Brigitte Nerlich. 2009. Metonymy. In J. L. Mey, editor, *Concise Encyclopedia of Pragmatics*, 2nd edition, Encyclopedia of Language and Linguistics, pages 631–634. Oxford: Elsevier.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor identification in large texts corpora. *PLOS ONE*, 8(4):1–9.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, volume 2: Short Papers, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems*, FOIS '01, page 2–9, Ogunquit, ME, USA. Association for Computing Machinery.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the international conference on Information and Knowledge Engineering*, IKE '03, pages 412–416, Las Vegas, NV, USA.

Thanapon Noraset, Chen Liang, Lawrence A Birnbaum, and Douglas C Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3259–3266, San Francisco, CA, USA.

Geoffrey Nunberg. 1987. Poetic and prosaic metaphors. In *Proceedings of the 1987 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '87, page 198–201, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Ekaterina Ovchinnikova, Ross Israel, Suzanne Wertheim, Vladimir Zaytsev, Niloofar Montazeri, and Jerry Hobbs. 2014. Abductive inference for interpretation of metaphors. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 33–41, Baltimore, MD, USA. Association for Computational Linguistics.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016a. PROMETHEUS: A corpus of proverbs annotated with metaphors. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC '16, pages 3787–3793, Portorož, Slovenia. European Language Resources Association (ELRA).

Gözde Özbal, Carlo Strapparava, Serra Sinem Tekiroğlu, and Daniele Pighin. 2016b. Learning to identify metaphors from a corpus of proverbs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 2060–2065, Austin, TX, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Natalie Parde and Rodney Nielsen. 2018. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC '18, pages 1535–1540, Miyazaki, Japan.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543, Doha, Qatar.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI '18, New Orleans, LA, USA.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, New Orleans, LA, USA.

Wim Peters and Ivonne Peters. 2000. Lexicalised systematic polysemy in WordNet. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, LREC '00, Athens, Greece.

The Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Sunny Rai and Shampa Chakraverty. 2017. Metaphor detection using fuzzy rough sets. In *Proceedings of the International Joint Conference on Rough Sets*, IJCRS '17, pages 271–279, Olsztyn, Poland. Springer.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys*, 53(2):1–37.

Sunny Rai, Shampa Chakraverty, and Devendra K. Tayal. 2016. Supervised metaphor detection using conditional random fields. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 18–27, San Diego, CA, USA.

Sunny Rai, Shampa Chakraverty, Devendra K. Tayal, and Yash Kukreti. 2017. Soft metaphor detection using fuzzy c-means. In *Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration*, MIKE '17, pages 402–411, Hyderabad, India. Springer.

Sunny Rai, Shampa Chakraverty, Devendra K Tayal, and Yash Kukreti. 2018. A Study on Impact of Context on Metaphor Detection. *The Computer Journal*, 61(11):1667–1682.

Sunny Rai, Shampa Chakraverty, Devendra K Tayal, Divyanshu Sharma, and Ayush Garg. 2019. Understanding metaphors using emotions. *New Generation Computing*, 37(1):5–27.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In

Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions, LAW-MWE-CxG '18, pages 222–240, Santa Fe, NM, USA. Association for Computational Linguistics.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 1537–1546, Copenhagen, Denmark.

Vassiliki Rentoumi, George A. Vouros, Vangelis Karkaletsis, and Amalia Moser. 2012. Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective. *ACM Transactions on Speech and Language Processing*, 9(3):1–31.

Philip S. Resnik. 1993. *Selection and Information: A Class-based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Ivor A. Richards. 1936. *The Philosophy of Rhetoric*. A Galaxy Book. Oxford University Press.

Lala Septem Riza, Andrzej Janusz, Christoph Bergmeir, Chris Cornelis, Francisco Herrera, Dominik Ślęzak, and José Manuel Benítez. 2014. Implementing algorithms of rough set theory and fuzzy rough set theory in the r package "roughsets". *Information Sciences*, 287:68 – 89.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 2890–2895, Online. Association for Computational Linguistics.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *NAACL-HLT '19*, pages 2692–2698, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2017. Sluice networks: Learning what to share between loosely related tasks. *arXiv preprint*, arXiv:1705.08142.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, volume 33 of *AAAI '19*, pages 4822–4829, Honolulu, HI, USA.

Evan Sandhaus. 2008. The New York Times annotated corpus. *Linguistic Data Consortium*, 6(12).

Marc Schulder and Eduard Hovy. 2014. Metaphor detection through term relevance. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 18–26, Baltimore, MD, USA. Association for Computational Linguistics.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Karin Kipper Schuler. 2006. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

John R. Searle. 1993. Metaphor. In Andrew Ortony, editor, *Metaphor and Thought*, 2nd edition, pages 83–111. Cambridge University Press, Cambridge, UK.

Elena Semino, Zsofia Demjen, Andrew Hardie, Sheila Alison Payne, and Paul Edward Rayson. 2018. *Metaphor, Cancer and the End of Life: A Corpus-based Study*. Routledge, London, UK.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '10, pages 1029–1037, Los Angeles, CA, USA.

Ekaterina Shutova. 2011. Computational approaches to figurative language. Technical Report UCAM-CL-TR-803, University of Cambridge, Computer Laboratory.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Ekaterina Shutova, Tim Van de Cruys, and Anna Korhonen. 2012. Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING '12, pages 1121–1130, Mumbai, India.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, pages 160–170, San Diego, CA, USA.

Ekaterina Shutova, Beata Beigman Klebanov, and Patricia Lichtenstein, editors. 2015. *Proceedings of the Third Workshop on Metaphor in NLP*. Association for Computational Linguistics, Denver, CO, USA.

Ekaterina Shutova, Beata Beigman Klebanov, Joel Tetreault, and Zornitsa Kozareva, editors. 2013a. *Proceedings of the First Workshop on Metaphor in NLP*. Association for Computational Linguistics, Atlanta, GA, USA.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 978–988, Atlanta, GA, USA.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1002–1010, Beijing, China.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC '10, pages 255–261, Malta.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013b. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. Mc Graw-Hill, Hill, Boston, USA.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

Egon Stemle and Alexander Onysko. 2018. Using language learner data for metaphor detection. In *Proceedings of the First Workshop on Figurative Language Processing*, pages 133–138, New Orleans, LA, USA.

Kevin Stowe, Sarah Moeller, Laura Michaelis, and Martha Palmer. 2019. Linguistic analysis improves neural metaphor detection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, CoNLL '19, pages 362–371, Hong Kong, China. Association for Computational Linguistics.

Kevin Stowe and Martha Palmer. 2018. Leveraging syntactic constructions for metaphor identification. In *Proceedings of the First Workshop on Figurative Language Processing*, pages 17–26, New Orleans, LA, USA.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC '04, pages 1083–1087, Lisbon, Portugal. European Language Resources Association (ELRA).

Tomek Strzalkowski, George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Samira Shaikh, Ting Liu, Boris Yamrom, Kit Cho, Umit Boz, Ignacio Cases, and Kyle Elliot. 2013. Robust extraction of metaphor from novel data. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 67–76, Atlanta, GA, USA.

Tomek Strzalkowski, Samira Shaikh, Kit Cho, George Aaron Broadwell, Laurie Feldman, Sarah Taylor, Boris Yamrom, Ting Liu, Ignacio Cases, Yuliya Peshkova, and Kyle Elliot. 2014. Computing affect in metaphors. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 42–51, Baltimore, MD, USA. Association for Computational Linguistics.

Chang Su, Shuman Huang, and Yijiang Chen. 2017. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.

Chang Su, Jia Tian, and Yijiang Chen. 2016. Latent semantic similarity based interpretation of chinese metaphors. *Engineering Applications of Artificial Intelligence*, 48:188–203.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39, Denver, CO, USA. Association for Computational Linguistics.

Paul H. Thibodeau and Frank H. Durgin. 2011. Metaphor aptness and conventionality: A processing fluency account. *Metaphor and Symbol*, 26(3):206–226.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pages 263–287. Springer International Publishing.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 248–258, Baltimore, MD, USA.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, GA, USA.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Edinburgh, Scotland, UK.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31th International Conference on Neural Information Processing Systems*, NIPS '17, pages 5998–6008, Long Beach, California, USA.

Tony Veale and Yanfen Hao. 2008. A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of the 22nd International Conference on Computational Linguistics*, COLING '08, pages 945–952, Manchester, UK.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor: A Computational Perspective*, volume 31 of *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool.

Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron Courville. 2017. Modulating early visual processing by language. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS '17, pages 6597–6607, Long Beach, California, USA.

Abigail Walsh, Claire Bonial, Kristina Geeraert, John P. McCrae, Nathan Schneider, and Clarissa Somers. 2018. Constructing an annotated corpus of verbal MWEs for english. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, LAW-MWE-CxG-2018, pages 193–200, Santa Fe, NM, USA.

Tong Wang, Abdelrahman Mohamed, and Graeme Hirst. 2015. Learning lexical embeddings with syntactic and lexicographic knowledge. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2: (Short Papers) of *ACL-IJCNLP '15*, pages 458–463, Beijing, China. Association for Computational Linguistics.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Koki Washio and Tsuneaki Kato. 2018a. Filling missing paths: Modeling co-occurrences of word pairs and dependency paths for recognizing lexical semantic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long Papers) of *NAACL-HLT '18*, pages 1123–1133, New Orleans, Louisiana. Association for Computational Linguistics.

Koki Washio and Tsuneaki Kato. 2018b. Neural latent relational analysis to capture lexical semantic relations in a vector space. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 594–600, Brussels, Belgium. Association for Computational Linguistics.

Koki Washio, Satoshi Sekine, and Tsuneaki Kato. 2019. Bridging the defined and the defining: Exploiting implicit lexical semantic relations in definition modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 3521–3527, Hong Kong, China. Association for Computational Linguistics.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44, Atlanta, GA, USA.

Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, and computers*, 20(1):6–10.

Magdalena Wolska and Yulia Clausen. 2017a. Simplifying metaphorical language for young readers: A corpus study on news text. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 313–318, Copenhagen, Denmark. Association for Computational Linguistics.

Magdalena Wolska and Yulia Clausen. 2017b. Simplifying metaphorical language for young readers: A corpus study on news text. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 313–318, Copenhagen, Denmark. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the First Workshop on Figurative Language Processing*, pages 110–114, New Orleans, LA, USA.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, COLING '00, pages 947–953, Saarbruecken, Germany.

Wenpeng Yin and Hinrich Schütze. 2016. Learning word meta-embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers of *ACL '16*, pages 1351–1360, Berlin, Germany.

Kai Yu, Shipeng Yu, and Volker Tresp. 2005. Soft clustering on graphs. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS '05, page 1553–1560, Vancouver, British Columbia, Canada. MIT Press.

Omnia Zayed, John P. McCrae, and Paul Buitelaar. 2018. Phrase-level metaphor identification using distributed representations of word meaning. In *Proceedings of the First Workshop on Figurative Language Processing*, pages 81–90, New Orleans, LA, USA.

Omnia Zayed, John P. McCrae, and Paul Buitelaar. 2019. Crowd-sourcing a high-quality dataset for metaphor identification in tweets. In *Proceedings of the 2nd Conference on Language, Data and Knowledge*, LDK '19, pages 10:1–10:17, Leipzig, Germany.

Omnia Zayed, John P. McCrae, and Paul Buitelaar. 2020a. Adaptation of word-level benchmark datasets for relation-level metaphor identification. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 154–164, Online.

Omnia Zayed, John P. McCrae, and Paul Buitelaar. 2020b. Contextual modulation for relation-level metaphor identification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Findings of the Association for Computational Linguistics: EMNLP 2020, pages 388–406, Online. Association for Computational Linguistics.

Omnia Zayed, John P. McCrae, and Paul Buitelaar. 2020c. Figure Me Out: A gold standard dataset for metaphor interpretation. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, LREC '20, pages 5810–5819, Marseille, France. European Language Resources Association.

Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint*, arXiv:1212.5701.

Rui Zhang, Cícero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. 2018. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2: Short Papers) of *ACL '18*, pages 102–107, Melbourne, Australia. Association for Computational Linguistics.

Chang-Le Zhou, Yun Yang, and Xiao-Xi Huang. 2007. Computational mechanisms for metaphor in languages: A survey. *Journal of Computer Science and Technology*, 22(2):308–319.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.