Automatic identification and translation of Multiword Expressions

Shiva Taslimipoor

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

2018

Signature: . . . . . .

Date: . . . . . .

Abstract

Multiword Expressions (MWEs) belong to a class of phraseological phenomena that is ubiquitous in the study of language. They are heterogeneous lexical items consisting of more than one word and feature lexical, syntactic, semantic and pragmatic idiosyncrasies. Scholarly research on MWEs benefits both natural language processing (NLP) applications and end users.

This thesis involves designing new methodologies to identify and translate MWEs. In order to deal with MWE identification, we first develop datasets of annotated verb-noun MWEs in context. We then propose a method which employs word embeddings to disambiguate between literal and idiomatic usages of the verb-noun expressions. Existence of expression types with various idiomatic and literal distributions leads us to re-examine their modelling and evaluation. We propose a type-aware train and test splitting approach to prevent models from overfitting and avoid misleading evaluation results.

Identification of MWEs in context can be modelled with sequence tagging methodologies. To this end, we devise a new neural network architecture, which is a combination of convolutional neural networks and long-short term memories with an optional conditional random field layer on top. We conduct extensive evaluations on several languages demonstrating a better performance compared to the state-of-the-art systems. Experiments show

that the generalisation power of the model in predicting unseen MWEs is significantly better than previous systems.

In order to find translations for verb-noun MWEs, we propose a bilingual distributional similarity approach derived from a word embedding model that supports arbitrary contexts. The technique is devised to extract translation equivalents from comparable corpora which are an alternative resource to costly parallel corpora. We finally conduct a series of experiments to investigate the effects of size and quality of comparable corpora on automatic extraction of translation equivalents.

Contents

List of Tables

List of Figures

List of Acronyms

ACL: Association for Computational Linguistics

AM: Association Measure

BNC: British National Corpus

CBOW: Continuous Bag Of Word

CC: Comparable Corpora

CL: Computational Linguistics

CNN: Convolutional Neural Network

ConvNet: Convolutional Neural Network

CRF: Conditional Random Field

DT: Decision Tree

IOB: Inside-Outside-Beginning

LR: Logistic Regression

LSA: Latent Semantic Analysis

LSTM: Long-Short Term Memory

LVC: Light Verb Construction

MI: Mutual Information

ML: Machine Learning

MLP: Multi Layer Perceptron

MVC: Multi-Verb Construction

MWE: Multiword Expression

NBC: Naïve Bayes Classifier

NER: Named Entity Recognition

NLP: Natural Language Processing

NN: Neural Network

PMI: Point-wise Mutual Information

POS: Part of Speech

RF: Random Forest

RNN: Recurrent Neural Network

SMT: Statistical Machine Translation

SVD: Singular Value Decomposition

SVM: Support Vector Machine

VNIC: Verb+Noun Idiomatic Combination

VPC: Verb-Particle Construction

VSM: Vector Space Model

WSD: Word Sense Disambiguation

Acknowledgements

There are many colleagues and friends whom I would like to thank for their kind help and support since the very beginning of this PhD. First and foremost, I would like to offer my sincerest gratitude to my supervisor, Prof. Ruslan Mitkov for all his support and guidance and the time he spent on my project despite all his commitments and academic duties. His valuable feedback truly helped my development over the years. I would also like to thank my second supervisor, Prof. Gloria Corpas Pastor for her always useful comments due to her deep linguistics knowledge.

A massive thank you goes to my external supervisor, Dr. Afsaneh Fazly, who actually started me on this path when I was a Master's student. I was lucky to meet her and am so grateful for being privileged to receive her insightful ideas and observations also for my PhD.

I would like to thank the examiners of my PhD thesis, Dr Aline Villavicencio and Prof, Mike Thelwall for their time spent on evaluating my work, their interest in the thesis and their invaluable comments.

My sincere gratitude goes to Dr. Sara Moze, Dr. Carla Parra, Dr. Irene Renu and Emma Franklin who helped me prepare the annotation guidelines. I appreciate the precious time they spent to shape the data preparation for this project. Special thanks to Sara and Carla for all the fruitful discussions and friendly advice. I would like to acknowledge lovely young students who delicately annotated the data: Manuela Cherchi, Anna Desantis, Andrea Silvestra, Lorena

Gomez, Martina Cotella and Alondra Nava Zea. I extend my gratitude to lovely Rut Gutiérrez Florido, Pilar Castillo Bernal and María Luisa Rodríguez Muñoz for their great help for pilot annotations.

I wish to express my deep thanks to current and past friendly members of the group, especially, Amanda Bloore, Kate Wilson, April Harper, Emma Franklin and Yvonne Skalban who proof-read different versions of my thesis and papers, and also for always kindly offering their technical and administrative help. As for the latter, I wish to name and thank Helen Williams and Iain Mansell. During the years I was a graduate student, I had the pleasure of meeting many bright people. I thank all my colleagues, fellow research scholars and visitors at RGCL: Victoria Yaneva and Richard Evans for our side collaborations and Miguel Rios, Rohit Gupta, Hernani Costa and Liling Tan for constructive discussions during their stay in RGCL. Many thanks to Dr. Le An Ha for the thoughtful comments during collaborations that we had in the final year of my study and his help with providing us with the GPU hardware required for running programs.

Moreover, I owe a word of special thanks to Omid Rohanian, who is the most inspiring person that I have met in my life, for his criticisms and encouraging me to try new things and compete. We had one of the most productive types of collaborations and we learned a lot together.

I am thankful for the financial and travel supports I received from RGCL.

Finally, it is a pleasant occasion to express my deepest appreciation to my parents and also to my gorgeous sister, Sahba for their love, support and trust in my abilities. I love them so much and I could not imagine any happy moment without their blessing.

List of Publications

Parts of this thesis have appeared in the following peer-reviewed publications:

- Taslimipoor, S., Rohanian, O., Mitkov, R. and Fazly, A. (2018). Identification of multiword expressions: A fresh look at modelling and evaluation. In S. Markantonatou, C. Ramisch, A. Savary, and V. Vincze (Eds.) *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop.* Language Science Press.
- Mitkov, R. and Taslimipoor, S. What matters more: the size of the corpora or their quality? The case of automatic translation of multiword expressions using comparable corpora. In Corpas, G and J.P. Colson (Eds). *Computational Phraseology.* John Benjamins. (To appear).
- Taslimipoor, S., Rohanian, O., Mitkov, R. and Fazly, A. (2017). Investigating the Opacity of Verb-Noun Multiword Expression Usages in Context. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE@EACL2017).* Association for Computational Linguistics, Valencia, Spain, pp. 133–8.
- Taslimipoor, S., Desantis, A., Cherchi, M., Mitkov, R. and Monti, J. (2016). Language resources for Italian: towards the development of a corpus of annotated Italian multiword expressions. In *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016.* Napoli, Italy, pp. 285–90.
- Taslimipoor, S., Mitkov, R., Corpas Pastor, G. and Fazly, A. (2016). Bilingual Contexts from Comparable Corpora to Mine for Translations of Collocations. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, Volume 9624 LNCS, 2018, pp 115–126, Springer.

- Taslimipoor, S. (2015). Cross-lingual Extraction of Multiword Expressions. In *Proceedings of Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives (Europhras2015)*, Malaga, Spain, pp. 232–7.
- Taslimipoor, S., Mitkov, R. and Corpas Pastor, G. (2015). Using Cross-lingual Contexts to Extract Translation Equivalents for Multiword Expressions from Parallel Corpora. In *Proceedings of Nuevos horizontes en los Estudios de Traduccin e Interpretacin (Comunicaciones completas)/New Horizons in Translation and Interpreting Studies*, Malaga, Spain, pp. 174–80.

Chapter 1

Introduction

1.1 Background and Motivation

A multiword expression (MWE) is a combination of two or more words that together exhibit idiosyncratic behaviour. Examples are *in someone's shoes*, *loose lips* and *give up*. MWEs are a pervasive phenomenon in language with their computational treatment being important for users and NLP applications (Ramisch and Villavicencio, 2018; Baldwin and Kim, 2010; Granger and Meunier, 2008).

The idiosyncratic properties of MWEs create difficulties in any language processing that involves them, such as machine translation, summarisation, sense disambiguation and question answering. In particular, Baldwin et al. (2005) have reported a notable amount of parsing failures caused by missing MWEs. Some of those challenging properties of MWEs are as follows:

Fixedness, the degree to which an expression is immutable, is one of the characteristics that has been used to recognise MWEs (Fazly and Stevenson, 2008). Whilst some expressions such as *by and large* are fully fixed and lexicalised, others undergo either morphosyntactic (e.g. *giving up*, *attorneys general*) or internal (e.g. *make sense*/*make perfect sense*) modifications.

Non-compositionality or **idiomaticity** relates to cases where the overall meaning of an expression cannot be derived from the meaning of its constituent words. An expression like *kick the bucket* is fully non-compositional; while the expression *spill the beans* is semi-compositional as the word *spill* can be interpreted as *reveal* and the word *beans* signifies *secrets*. On the other hand, the expression *climb up* is fully compositional since the whole meaning of the expression is easy to perceive by knowing the meanings of the components.

Semi-productivity refers to how open a constituent of an expression is, to substitute the other constituent with semantically similar words in order to construct new valid expressions. For instance, in the case of *fast food*, the substitution of *fast* with *quick* is not acceptable. *Eat up* is productive to some extent in that *gobble up* or *drink up* are valid. However, *swallow up* does not follow this productivity and is an idiom that suggests several meanings such as, *to take control*, *to consume* or *to destroy something completely*.

These heterogeneous properties make word-for-word processing of multiword expressions futile. MWEs don't fit well in the traditional grammar descriptions where there is a distinct line between lexicon and grammar (Green et al., 2013). It has been widely discussed that simply listing these expressions in lexica is not a feasible solution (Hanks, 2013) and obtaining wide-coverage language resources for general MWEs is currently a bottleneck for NLP systems (Ramisch et al., 2010). Listing MWEs as strings is only adequate for expressions which allow absolutely no variability (i.e. truly fixed

expressions). Even listing some MWEs in lexica requires special treatment (Corpas Pastor, 2017).

Language change is a constant process and just like new words enter the dictionary, new MWEs are coined every day, unexpectedly and without any underlying rules. In particular, the widespread use of social media has made it easier to contribute to the evolution of language. Many MWEs occur spontaneously, in a similar fashion to slang expressions. One example is *binge watching*, which is defined by the Oxford Dictionary as 'watching multiple episodes (of a television programme) in rapid succession', typically by means of DVDs or digital streaming. Some MWEs are used less frequently, until they eventually disappear, as they are being replaced by new ones.

Over the years, MWEs have been the topic of increasing interest for NLP researchers. Ramisch et al. (2013b) have gathered relevant statistics from the papers in the anthology of Association for Computational Linguistics (ACL) for the years 1965 to 2006. They have plotted the ratio of papers which mention "multiword","collocation" or "idiom" with respect to the total number of papers in that time span. Their plot shows a general increase in the proportion of papers conscious of the phenomenon. We have conducted a similar experiment in which the occurrences of the word "multiword" are counted in the papers in the newer version of ACL anthology[1] from the year 1980 to 2015. We plot the ratio of the number of papers in which the word

[1]ACL anthology reference corpus version 20160301 accessed from: `http://acl-arc.comp.nus.edu.sg`

"multiword" appears with a frequency of more than 1 with respect to the total number of papers in those years. We also plot ratios of the number of papers that have the word in their title. The plots are depicted in Figure 1.1. The increasing use of the above terms reflects the growing importance of multiword units in Natural Language Processing.

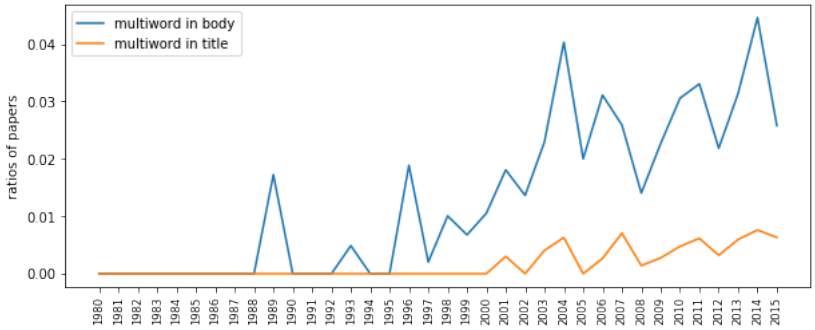


Figure 1.1: Ratios of papers in ACL anthology mentioning the word "multiword".

Although the computational treatment of multiword units has been extensively explored in Computational Linguistics (Baldwin and Kim, 2010; Granger and Meunier, 2008), we believe that there is a lot of ground to be covered and new promising results to be achieved in this field. Gold standards and tagged corpora are scarce in many languages and the computational methodologies and results are often incomparable with each other. We compile new corpora and revisit supervised methods to devise new and more robust approaches for MWE identification and evaluation of its modelling. [2]

There are various patterns of MWEs with regards to specific roles that

[2] In this context, modelling refers to developing a machine learning model in order to identify Multiword Expressions.

they play or parts of speech that they receive in a sentence and different lexical properties of their components. Examples include:

> *take place* which is a combination of a verb and a noun or *give up* which is a combination of a verb and a particle, and both MWEs function as verbs,
>
> *flood water* which is a combination of two nouns that together make one noun entity,
>
> *by and large* whose adverbial function has nothing to do with the common roles of its constituent words (the preposition *by* and the adjective *large*).

The most thorough approach to analyse MWEs is to examine corpora and look into the contexts in which they appear. However, it is quite challenging to find common features to train a computational model that can treat MWEs as special meaningful lexical units made up of several words.

Among different types of MWEs, we focus on 'Verb + Noun' combinations, which are challenging for automatic language processing, because they are not truly fixed (e.g. while *take place* is a fixed expression, *make decisions* is not and can be altered to *make a good decision*). The word components may or may not be inflected and the whole meaning may or may not be derived from the meaning of the components. Ramisch et al. (2013a) have reported that current Statistical Machine Translation (SMT) systems can only correctly translate 27% of phrasal verbs.

In the case of identification of verb-noun MWEs, we mainly focus on their idiomatic behaviour rather than their habitual juxtaposition. Specifically, our goal is to discriminate between idiomatic and literal expressions. We are aware that not all MWEs are idiomatic, and different MWEs exhibit a cline of fixedness and idiomaticity. However, a scale-wise modelling of MWEs is out of the scope of this study.

One might find most of our focused expressions to be light verb constructions (LVCs) as extensively discussed in Fazly (2007). A light verb construction is a combination of a verb and a noun in which the verb has little semantic content of its own and the noun carries the main meaning of the expression (e.g. *make a decision*, *have a speech*). We do not investigate LVC properties; rather, following Fazly (2007)'s categorisation of verb-noun combinations into 1) literal combinations, 2) abstract combinations, 3) light verb constructions and 4) idiomatic combinations, we differentiate between compositional verb phrases (literal combinations and abstract combinations) and multiword predicates (light verb constructions and idiomatic combinations). In Chapter 2, we discuss previous computational works on the whole group of MWEs, including idioms, light verb constructions and collocations. In this thesis, we mostly refer to them with the general term, MWE.

Recently, automatic tagging of corpora for MWEs (Schneider et al., 2014a; Constant and Tellier, 2012) has received significant attention and new datasets of running texts, tagged for MWEs, are devised (Savary et al., 2017). Most of the works in this area model the task as structured prediction. Accord-

ing to Constant et al. (2017), broad-coverage MWE identification is still an open issue and the use of end-to-end sequence taggers based on recurrent and deep neural networks remains to be explored. Chapter 5 proposes and evaluates a novel approach for neural network based structured prediction in order to solve the problem which is defined in the shared task on automatic identification of verbal multiword expressions.

Whilst there are many studies on the automatic extraction of MWEs from monolingual text, there are only a few studies that draw on bilingual resources for the automatic treatment of such expressions (Bouamor et al., 2012; Corpas Pastor, 2017; Mendoza Rivera et al., 2013; Morin and Daille, 2010; Daille et al., 2004). Corpas Pastor (2017) has extensively discussed the need for such expressions to be represented in bilingual dictionaries, focussing on collocations as one type of MWEs. For example, collocations such as *pay attention* and *pay homage* require different translations of the collocative verb to Spanish depending on the base noun: *prestar/poner atención*, *rendir homenaje*. In addition, automatic extraction and translation of multiword units from comparable corpora, which is a rich and plentiful resource, is an under-researched topic.

Dealing with MWEs bilingually is very interesting for two reasons: firstly, finding translation equivalents for these expressions remains an unresolved issue in NLP; secondly, using bilingual corpora we can improve their identification, especially for resource-poor languages (Salehi and Cook, 2013; Tsvetkov and Wintner, 2014).

Due to the limited availability of parallel data in many languages, we propose a methodology that benefits from comparable corpora to find translation equivalents for collocations (as a specific type of difficult-to-translate multi-word expressions). Our novel approach is based on bilingual context extraction and build a word (distributional) representation model drawing on these bilingual contexts. We show that the bilingual context construction is effective for the task of translation equivalent learning.

1.2 Aim and Scope

We focus on a cross-lingually prevalent class of MWEs, namely verb-noun constructions, which are commonly and productively formed from a frequent verb followed by a noun. We investigate both identification and finding translation equivalents for these expressions. In this thesis, we mostly develop language independent systems for both identification and translation of verb-noun MWEs.

1.2.1 Identification of Verb-Noun MWEs

As the first step we explore methods to identify verb-noun idiomatic expressions. We perform our experiments with three languages: Italian, Spanish and English. Due to the scarcity of resources to evaluate the proposed approaches, our first task is to compile gold-standard datasets. Then after some analysis on discovering potential MWEs in the datasets, we look closer on the ambiguous usages of candidate expressions. The fact that some can-

didate expressions have both literal and idiomatic usages (e.g. *make face* which is literal in *they make faces on heads with pieces of cheese olive or egg* and idiomatic in *she made a face and turned away*) prompts us to deeply investigate the identification of usages of MWEs in context.

1.2.2 Tagging Verbal MWEs in Running Text

Inspired by the recent interest towards tagging corpora for MWEs and structured prediction in general, we also devote a chapter of this thesis to our proposed system for identification of verbal MWEs in running texts for which we experiment with existing datasets.

1.2.3 Translation Equivalents for Verb-Noun MWEs

In this thesis, we also focus on automatically finding translation equivalents for verb-noun MWEs specifically from bilingual comparable corpora. Bilingual comparable corpora (McEnery and Xiao, 2007) are promising resources which are available in far greater amount compared to parallel corpora, specially for resource-poor languages. Translation equivalents of MWEs are useful in order to help improve the performance of machine translation systems. For the purpose of this study, we compile English-Spanish comparable corpora and perform the experiments on the English-Spanish language pair, while the proposed approach is applicable to any two languages.

1.3 Research Questions and Contributions

The following research questions are considered in this thesis.

RQ 1) To what extent can the challenges with the availability and appropriateness of gold-standards for computational treatment of MWEs be resolved?

After years of studies on MWEs and their importance, there are insufficient standard resources (gold-standards, tagged corpora, lexica, etc) for MWEs of different languages and there is no consensus among researchers on how to model MWEs. In order to answer the first research question, we focus on a specific group of expressions namely verb-noun expressions. We put forward the hypothesis that if we deeply investigate the question with this group of MWEs, the approach can be adapted to any kind of expressions. While there are resources available for English verb-noun MWEs (Cook et al., 2008), currently, there is no available gold-standard data for languages like Italian and Spanish. We perform several sets of experiments in two different settings: out-of-context and in-context:

1) We extract lists of verb-noun expressions for three languages: English, Spanish and Italian. We prepare a set of guidelines for native speakers to annotate the expressions as idiomatic or not. As a result, we provide sizeable MWE datasets, validated with computed inter-annotator agreements. These resources can be used for any future research. Furthermore, we report extensive results on the performance of different statistical association measures in the task of distinguishing between literal and idiomatic expressions. Limitations in dealing with this scenario are illustrated, which led us to the second sets of experiments.

2) We compile large corpora of verb-noun expressions augmented with context using the SketchEngine tool for Italian and Spanish (there is a benchmark data available for English). We design guidelines for annotation, and with the collaboration of several native speakers we construct gold-standards for identifying verb-noun MWEs in context. The information about the developed resource is published in Taslimipoor et al. (2016a). These datasets are a by-product of this study. The first part of Chapter 3, details how RQ 1 has been addressed.

RQ 2) How can we disambiguate between different occurrences of the same expression type which are idiomatic in some contexts but literal in others?

An expression like *have a word* is literal in *does Spanish have a word for it?*, and idiomatic in *he is going to have a word with his daughter*. NLP systems should discriminate between the idiomatic and literal usages of these expressions. We propose a new approach to deal with such expressions that are ambiguous in their usages as MWE or not. Previous studies (as discussed in Chapter 2) have used syntactical and lexical features to disambiguate these cases. However, we focus on challenging cases where tokens have the exact same lexical and syntactic features. In order to distinguish idiomatic from literal occurrences, we utilise contextual features derived from state-of-the-art distributional similarity methods in a supervised scenario. The experiments and results of this study are covered in the second part of Chapter 3. This work was published in Taslimipoor et al. (2017).

RQ 3) What is the reason behind the significant variation in reported results for MWE identification and is there a better way of modelling MWEs and more reliable evaluation methodology?

We first analyse our data in order to find the best modelling approach (standard classification or structured prediction) to identify MWEs for this data. On the basis of the conducted experiments, we conclude that MWEs in this data could be better modelled by using classification rather than tagging methods. The results of this investigation will be published as part of the research in Taslimipoor et al. (2018).

Based on the experiments we have done on identifying MWEs, we believe that some of the seemingly high evaluation scores reported in the literature are misleading, and stem from the fact that the same expression types frequently occur in both the training and test data. This phenomenon leads to a form of overfitting which can be overlooked by standard evaluation methods.

We propose what we call 'type-aware' splitting of data into train and test in order to make the learning process more generalised and the evaluation more rigorous. The innovative train and test splitting approach is proposed as a new benchmark for modelling and evaluating the task of MWE identification. Extensive experiments and results using different classification algorithms are reported in Chapter 4 and published as part of the upcoming book chapter by Taslimipoor et al. (2018).

RQ 4) In the case of automatic identification of MWEs in running text or tagging corpora for MWEs which is a recent direction

in NLP studies on MWEs, to what extent can the state-of-the-art be improved?

The recent interest in the methodologies to tag corpora for MWEs leads us to investigate more in this direction. To address the research question, we target the shared task on 'automatic identification of verbal multiword expressions'. To tag verbal MWEs, we propose a new hybrid deep learning approach which is a combination of convolutional neural networks and long short term memories with an optional conditional random field layer on top. To the best of our knowledge, this is the first attempt at using such a hybrid model for MWE tagging. We report the promising results of our system in Chapter 5. The proposed model outperforms other systems applied to the datasets provided by the shared task, without using any task-specific domain knowledge beyond generic POS tags and pretrained embeddings.

RQ 5) Since parallel data is limited, can we determine a new approach to extract translation equivalents for verb-noun multiword expressions that works better than methods used in previous studies on translation equivalent extraction?

In order to answer this question, we perform several sets of experiments. We first compile English-Spanish comparable corpora from news sources on the web. The resulting compiled dataset is another outcome of this research. Furthermore, we propose a novel approach based on distributional similarity, in order to find translation equivalents for these expressions from comparable corpora. This method is especially relevant for resource-poor languages where

translation resources are scarce. Distributional similarity-based approach has proven successful for finding similarities in monolingual data, however, we propose a new bilingual similarity extraction methodology inspired by state-of-the-art word embedding approaches. The findings of this work are published in Taslimipoor et al. (2016b) and reported in Chapter 6.

Furthermore, we investigate the effect of quality and quantity of comparable corpora in the task of translation equivalents extraction. To this end, we conduct extensive experiments using our corpora and also the freely available Wikipedia comparable corpora. Our results, which correspond to the size and quality of comparable corpora in certain ranges and how they influence the quality of translation equivalents, are reported in Chapter 6 and published as a book chapter (Mitkov and Taslimipoor, nd).

1.4 Thesis Outline

Following Chapter 1, the remaining chapters of this thesis are organised as follows.

Chapter 2 provides the definition of MWEs and discusses the related work in computational treatment of these expressions. This chapter also includes description of previously proposed approaches for identifying MWEs and finding translations for them. The statistical approaches and machine learning techniques employed in this study are further detailed in this chapter.

The subsequent chapters each include description of methodologies, their

evaluation and results.

Chapter 3 first describes the data preparation stage of the project, annotation, analysis and details of the data, both for in-context and out-of-context expressions. The chapter then details the proposed methodology to disambiguate verb-noun MWEs in context. The idea is to use vector representations of the component words and their context in a supervised scenario. The experiments and results of the approach are detailed in this chapter.

Chapter 4 focuses on the limitations of modelling MWE in context and presents the first attempt to integrate a more generalised way of training and evaluation using the new type-aware train and test splitting. In this chapter, we also propose classification over tagging as the better approach for modelling MWEs in context for our data.

Chapter 5 is devoted to the presentation of the system that we propose for identifying MWEs in running text. The results of the system on the datasets provided by the shared task on automatic identification of verbal multiword expressions are reported and discussed in this chapter.

Chapter 6 describes the task of finding translation equivalents for MWEs using comparable corpora. The compilation of comparable corpora for English and Spanish is outlined in this chapter. We present a new distributional similarity based methodology for finding translation equivalents for verb-noun MWEs. The evaluation and results of the approach are also covered. This chapter also provides insights into a detailed study on the effects of quality and quantity of comparable corpora on finding translation equiv-

alents.

Chapter 7 revisits the main research questions and original contributions of this thesis, commenting on their strengths and limitations, and discusses suggestions for future research.

Chapter 2

Related Work

This chapter reviews various previous studies on MWEs. First we provide an overview of the definitions and characteristics of MWEs. Subsequently, in section 2.2 we review popular available resources for computational treatment of these expressions. Then, in section 2.3 various computational studies on MWE identification are discussed in detail. Next, we explore more deeply the studies on ambiguity of MWEs. The materials and methods used subsequently in this thesis are all introduced and discussed in this chapter. Finally, we move to studies on finding translation equivalents in general and for MWEs. The chapter concludes with a brief summary of the literature reviewed.

2.1 Multiword Expressions Definitions and Properties

Almost all researchers in natural language processing agree with the definition by Sag et al. (2002) that MWEs are idiosyncratic interpretations that cross word boundaries (or spaces). In this sense, MWEs are combinations of words (not necessarily continuous) for which the whole unit has idiosyncratic behaviour. Idiosyncrasy of MWEs can be observed in different ways:

- **Lexical**, when there is no entry for one or more components in lexica of the language (Baldwin and Kim, 2010). An examples is: *ad hoc*

- **Syntactic**, as in *by and large* with special combination of a preposition and adjective (Baldwin and Kim, 2010), or *shoot the breeze* which is not flexible enough to be passivised (i.e., **the breeze was shot*) (Fazly et al., 2009).

- **Semantic**, when the meaning of the whole expression cannot be predicted from a simple composition of the meanings of its component words. *spill the beans* is an example made in Sag et al. (2002), for which *spill* has the sense of *reveal* and *the beans* means *the secrets* in the idiomatic usage of this expression.

- **Statistical**, because their frequencies are very high. For instance, we find the expression *take place* to be the most frequent light verb + noun expression in the BNC.

MWEs have been referred to with several different terms (Ramisch et al., 2013b; Schneider et al., 2014b). Phraseology is the discipline which studies MWEs or their related concepts referred to by scholars as, for example, multiword units, multiword expressions, fixed expressions, set expressions, phraseological units, formulaic language, phrasemes, idiomatic expressions, idioms, collocations, and polylexical expressions (Monti et al., 2018). The term 'multiword expression' is the most accepted in NLP and is used in a

series of annual workshops in major conferences in computational linguistics, namely Multiword Expression Workshops, since 2001.[1]

MWEs are a recurring theme in any language with some sources estimating their number to be in the same range as single words (Jackendoff, 1997) or even beyond (Sag et al., 2002). Identification of MWEs has been shown to be effective in different NLP tasks, such as machine translation (Pal et al., 2011; Mitkov et al., 2018) and automatic parsing (Constant et al., 2012). In part-of-speech tagging, parsing and machine translation, these expressions should be treated either before the task (Nivre and Nilsson, 2004) or combined with the process (Constant and Tellier, 2012; Kordoni et al., 2011; Nasr et al., 2015).

Among different types of MWEs, verbal MWEs are one of the most challenging due to their complex characteristics including discontinuity, non-compositionality, heterogeneity and syntactic variability. The PARSEME COST Action[2] regarded verbal MWEs in different languages and resulted in two editions of shared tasks on verbal MWE identification (Savary et al., 2017). Various researchers focus on specific categoreis of verbal MWEs such as verb particle constructions (VPCs) (Villavicencio, 2005), light verb constructions (Stevenson et al., 2004), verb-verb compounds (Uchiyama et al., 2005) and, verb+noun idiomatic combinations (VNIC) (Fazly et al., 2009; Salton et al., 2016).

[1] `http://multiword.sourceforge.net/PHITE.php?sitesig=CONF`
[2] `http://www.parseme.eu`

Verbal MWEs and especially verb-noun idiomatic combinations are of interest in this study. Fazly (2007) drew a figurativeness continuum line for verb-noun combinations as depicted in Figure 2.1. We mainly focus on differentiating between more literal (compositional verb phrases) and more idiomatic (multiword predicates) expressions.

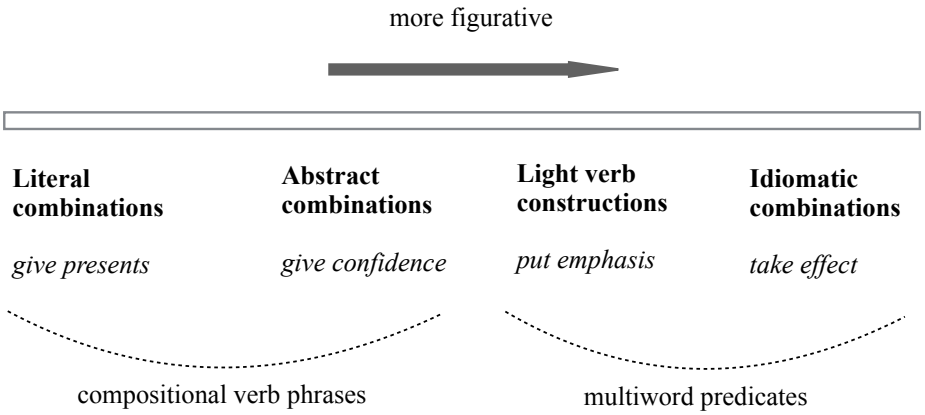


Figure 2.1: Classes of verb+noun combinations on the figurativeness continuum.

Depending on the motivation and the purpose of their studies, researchers have analysed MWEs from different angles. A fundamental categorisation in MWE studies regards the question whether one aims to extract potential MWEs from a corpus in a language or to tag a corpus with actual occurrences of such expressions. The answer to this question leads to quite different directions for investigation of MWEs and their modelling. The former, which is referred to as type-based extraction of MWEs or MWE discovery, is a traditional approach which is of use to lexicographers as pointed in Ramisch (2014); the latter considers studies on idiomatic usages of expressions, namely

MWE tagging or token-based identification of MWEs and is more practical for NLP applications (Schneider et al., 2016). MWE tagging is also very related to the task of chunking which is a basis for parsing (Tjong Kim Sang, 2000) and super sense tagging (Schneider et al., 2016).

It is important to differentiate between these two directions for two main reasons. First, there are many expressions that are well-known for their idiomatic behaviour such as *kick the bucket*, however they can also be found in their usages with literal interpretations. There are also lexical units whose co-occurrence is naturally literal, such as *play games*, that can act as an idiom in some contexts. Second, the reference resources and gold standard datasets to be used for these directions are a bit different. In the next section, we provide a detailed overview of available datasets for computational modelling of MWEs for both directions.

2.2 Resources

In order to set up the computational treatment of MWEs, first, suitable resources should be compiled. After more than two decades of computational studies on MWEs, the lack of proper gold standards is still an issue. More traditional studies (Lin, 1999; Baldwin et al., 2003) used as their gold standard either idiom dictionaries or WordNet (Fellbaum, 1998). Lexical resources like dictionaries have limited coverage for these expressions (Losnegaard et al., 2016) and properly tagged corpora of MWEs in different languages are scarce (Schneider et al., 2014b).

Datasets for MWE types extraction usually include lists of expressions which are annotated with properties related to their acceptability as being MWE or not. Such datasets should feature positive and negative cases or different categories of MWEs in order to be useful for computational studies. Most available resources have targeted certain kinds of MWEs and excluded the others. They looked at, for example, subclasses of nominal compounds (Kim and Baldwin, 2006; Farahmand et al., 2015), light verb constructions (Fazly, 2007), collocations (Gelbukh and Kolesnikova, 2010), verb particle constructions (Baldwin, 2005) or all kinds of verbal MWEs (Savary et al., 2017).

For example, Farahmand et al. (2015) presented a set of 1,048 noun-noun compounds annotated as non-compositional, compositional, conventionalised and not conventionalised. Fazly and Stevenson (2007) compiled a list of 563 verb-noun phrases which are labelled as one of the four categories: literal, abstract, LVC or idiom.

Quite a number of such resources have been maintained by SIGLEX-MWE, the Special Interest Group on the Lexicon of the Association for Computational Linguistics. To list a few, Tu and Roth (2011) have collected a balanced benchmark dataset with 2,162 sentences from BNC for English LVCs, constructed with 6 most frequently used light verbs: *do*, *get*, *give*, *have*, *make* and *take*. Cook et al. (2008) have also compiled a list of 2,984 English verb-noun tokens from BNC annotated for being idiomatic or not. Two English native speakers selected the expression types based on whether

they have the potential for occurring in both idiomatic or literal senses. Then all (around 3,000) sentences from BNC, in which those selected expressions occurred, have been collected and annotated. The dataset which is called VNC-Tokens is a benchmark for English verb-noun idiomatic expressions and has been used for identifying MWE tokens in a number of previous studies such as Fazly et al. (2009) and Salton et al. (2016). Hashimoto and Kawahara (2008) have developed a corpus of Japanese idioms which can be considered as the Japanese idiom counterpart of VNC-Tokens and is one of the largest of this kind with as many as 102, 846 example sentences. However, there is no such dataset for other languages.

In general, evaluating all occurrences of expressions in the whole corpus of large size is not feasible and there are comparatively few small datasets available for token-based identification of MWEs. The most commonly used datasets of this kind are the ones provided by shared tasks related to MWEs. Recently, two successful shared tasks were held in the field. One is the SemEval (2016) shared task on Detecting Minimal Semantic Units and their Meanings (DiMSUM) in which Schneider et al. (2016) provide a corpus of not very large size but comprehensively tagged with all kinds of MWEs. The PARSEME network has also initiated a comprehensive annotation of MWEs in several languages, focusing on verbal expressions only. The first phase resulted in the PARSEME shared task on verbal MWE identification (Savary et al., 2017) which released MWE-annotated corpora for 18 languages and the second phase led to the second edition of the shared task on verbal MWE

identification to be organised in COLING 2018.

To summarise, building datasets for MWEs is challenging, since:

1) annotating large corpora for all kinds of MWEs is not feasible,

2) linguistic features for different categories of MWEs are different and researchers have to focus on specific categories which might not be the focus for others,

3) finding human annotators, with linguistic knowledge to annotate corpora is difficult. Even when experienced linguists annotate the expressions, as explained in our experiments in Chapter 3, the agreement between annotators is not sufficient most of the times.

For this thesis, we compile our datasets for two languages: Italian and Spanish. We have them annotated for verb-noun MWEs as explained in Chapter 3. We also use the aforementioned recent datasets for further studies as reported in Chapter 5.

2.3 Extraction and Identification

Studies on MWEs can be divided into two main categories. One includes works regarding the canonical forms [3] of expressions, their lexical properties and their potential to be considered as MWEs, namely type-based extraction of MWEs or MWE discovery; the other comprises studies on tagging texts

[3]Canonical form of an expression is the non-inflected form of the expression that is listed in dictionaries.

for the idiomatic usages of expressions, namely MWE tagging or token-based identification of MWEs. The former is a traditional approach which is of use to lexicographers (as pointed in Ramisch (2014)); the latter though, is more practical for NLP applications (Schneider et al., 2016). In any case, the most common approach to treat MWEs computationally is by examining corpora in any language (Evert and Krenn, 2005; Ramisch et al., 2010; Villavicencio, 2005).

2.3.1 Type-based Extraction of MWEs

Previous work on processing MWEs in the direction of type-based extraction mostly used statistical association measures (Manning and Schütze, 1999; Smadja, 1993). There is a lot of work which focused on semantic investigations of certain kinds of MWEs (Baldwin et al., 2003; Bannard et al., 2003; Fazly et al., 2009; McCarthy et al., 2003). Therefore, it has become standardised in NLP to deal with MWE types by ranking them according to either the degrees of association between their components (Ramisch et al., 2010) or degree of compositionality of their meanings (Baldwin and Kim, 2010).

2.3.1.1 Association Measures

Statistical association measures are widely used in MWE identification/aquisition due to the collocational behaviour of these expressions. There is no consensus, however, about which measure is best suited for identifying MWEs in general. Evert and Krenn (2005) argued that the results of an evaluation experiment on extracting collocation cannot easily be generalised to a differ-

ent setting; and only an empirical evaluation can identify the best association measure under a given set of conditions.

An association measure is a statistical quantity used to indicate the strength of the relationship between two variables, which are words in this context. Different association measures work on contingencies between words in different ways. We use the association measures, as they are identified in the widely used SketchEngine (Kilgarriff et al., 2004), to compute the association between the components of the expressions.[4]

SketchEngine is a corpus manager and text analysis tool which provides numerous corpora in various languages. It features collocation search and generates statistics related to the co-occurrences of collocations' constituents.

To describe the statistics used in SketchEngine, the following conventions apply unless specified otherwise:

N – corpus size,

f_A – number of occurrences of the keyword A, which is a target verb for our purpose, in the whole corpus,

f_B – number of occurrences of the collocate B, which is a noun-tagged word for our purpose, in the whole corpus,

f_{AB} – number of co-occurrences of the verb A and the noun B, which in this study we consider them when they are adjacent.

[4]There are different estimations for some association measures (e.g. log-likelihood) in the literature. We adopt the ones used in SketchEngine which is widely used among lexicographers.

Mutual Information is one of the oldest association measures used as a statistical approach to score the co-occurrence of two words, initially defined in Church and Hanks (1990). The mutual information score for the bigram $(A; B)$ is computed as the logarithm of the probability of seeing two words A and B together, divided by the product of the words' individual probabilities (whether they occur together or in isolation). There exist different variations of this measure in the literature and it has been widely used as a basis for collocation extraction (Fazly, 2007). A variant of this measure which is implemented in SketchEngine is defined in equation 2.1. This definition is widely known as point-wise mutual information (PMI).

$$\text{PMI-Score} = \log_2 \frac{f_{AB} \times N}{f_A \times f_B} \tag{2.1}$$

The probabilities of words A and B are calculated directly using relative frequency. A large mutual information score signifies that a given bigram is found more often than chance. However, in practice, mutual information misclassifies some low-frequency data as collocations.

Log-likelihood ratio is another very common association measure which is well-known for being effective in extracting collocations. It allows a direct comparison of the significance of common and rare phenomena and works well with both large and small sizes of corpora (Dunning, 1993). Again, large Log-likelihood scores are interpreted as an association between the words in a bigram. Log-likelihood is computed as in equation 2.2.

$$\begin{aligned}\text{Log-likelihood} = 2 \times (&xlx(f_{AB}) + xlx(f_A - f_{AB}) + xlx(f_B - f_{AB}) + xlx(N) \\ &+ xlx(N + f_{AB} - f_A - f_B) - xlx(f_A) - xlx(f_B) \\ &- xlx(N - f_A) - xlx(N - f_B))\end{aligned} \tag{2.2}$$

where $xlx(f)$ is $f \times ln(f)$. This computation of Likelihood is according to Dunning (1993).

T-Score measure or t-test as described in Krenn and Evert (2001) is an association measure defined to alleviate the low-frequency bias in point-wise mutual information. It is defined as follows:

$$\text{T\_Score} = \frac{(f_{AB} - \frac{(f_A . f_B)}{N})}{\sqrt{f_{AB}}} \tag{2.3}$$

Log_Dice, which is the logarithmic form of the Dice formula (Dice, 1945), was reported as one of the most effective association measures in MWE induction in Schone and Jurafsky (2001).

$$Log\_Dice = \ln \frac{2.f_{AB}}{f_A + f_B} \tag{2.4}$$

Salience is a recently proposed association measure which is a combination of several statistical measures for association strength. Specifically, this is an adjustment to point-wise mutual information and is estimated as the product of mutual information and log frequency. According to Kilgarriff et al. (2004) this measure is called $MI.log - f$ and is computed as follows:

$$Salience = MI\_Score \times \ln (f_{AB} + 1) \tag{2.5}$$

Salience is also known as lexicographer's PMI or LMI (Biemann and Riedl, 2013).

These equations are just a small selection of the many association measures that have been suggested and used over the years. Evert (2005) discussed more than 30 different measures, Pecina (2005) listed over 80 measures, and new measures and variants are constantly being invented (Evert, 2008). Evert (2008) mentioned that while some measures have been established as de-facto standards (e.g. log-likelihood in computational linguistics, t-score and MI in computational lexicography), there is no ideal association measure for all purposes. Each measure is focused on a certain aspect of collocation strength and its suitability differs depending on the task.

Ramisch et al. (2010) have developed a toolkit to compute association measures for expressions based on user-defined grammatical rules. The program uses these as features to extract MWEs. They view the task of MWE identification as a classification problem and feed the computed association measures as features to the WEKA[5] machine learning toolbox. They evaluated their approach in identifying domain-specific multiword terms by comparing the performance with that of Xtract (Smadja, 1993).

Using the MWEToolkit developed by Ramisch et al. (2010), Rondon et al. (2015) built a system based on supervised machine learning approaches that continuously learns new expressions from the web. They used MWEToolkit to extract MWE expressions at the first stage and they mainly focused on

[5]http://www.cs.waikato.ac.nz/ml/weka/

the improvement of this extraction over time. In terms of using supervised machine learning methods for MWE identification, many alternatives have been tested among which Pecina (2008) proposed a logistic regression classifier which uses as features a set of different lexical association measures.

While having a long history, association measures are still of interest for MWE extraction and identification, specially due to the increasing availability of large corpora.

2.3.1.2 Other Computational Approaches

There exist other statistical measures in the literature (usually derived from association measures) that reflect various properties of MWEs. For example Stevenson et al. (2004) devised a formula to account for semi-productivity of LVCs, and Fazly et al. (2009) proposed similar metrics for fixedness of idiomatic expressions.

Some studies focus on other aspects of idiosyncrasies in MWEs beside their statistical properties. They use various syntactic or semantic features and computational approaches (supervised or unsupervised) for canonical extraction of MWEs.

Baldwin et al. (2003) estimated the degrees of decomposability of MWEs by computing the semantic similarity between the expressions and their constituent words. They hypothesise that the higher similarities indicate the greater decomposability of expressions. They used latent semantic analysis (LSA) for the similarity method. Salehi et al. (2015) have a similar hypothe-

sis to predict the compositionality of MWEs and proposed a word embedding method to find similarities between expressions and their component words. Kim and Baldwin (2005) used WordNet similarity to find semantic relations between components of noun compounds.

Villavicencio (2005) used the productive pattern of verb-particle constructions to determine new constructions for semantically similar words. A relevant work which considers the substitution of expression constituents with other words is Lin (1999). He compared mutual information of MWEs with the mutual information of similar expressions constructed from substituting one of their constituents with a similar word. The hypothesis is that greater difference between the mutual information measure of an expression and that of its similar expression (obtained by substitution) implies a higher non-compositionality for that expression.

There are also some previous work which leverage parallel corpora and word-alignment strategies in order to identify MWEs (Moirón and Tiedemann, 2006; de Caseli et al., 2010).

Although discovering canonical forms of multiword expressions is still an active research area (Salehi and Cook, 2013; Farahmand and Martins, 2014), recently classifying tokens of MWEs (Fazly et al., 2009; Gharbieh et al., 2016) or automatic tagging of corpora for MWEs (Schneider et al., 2014a; Constant and Tellier, 2012) have gained more traction. Related work on these is outlined in the following section.

2.3.2 Token-based Identification of MWEs

Researchers have modelled token-based identification of MWEs in different ways. Some of them first identify MWE types and then locate each expression type in the corpus and assign a proper label to it, either identical to the idiomatic interpretation of the type (Brooke et al., 2014; Cordeiro et al., 2016) or based on some lexical, syntactic or semantic features extracted and compared with its canonical form (Fazly et al., 2009).

Brooke et al. (2014) proposed an unsupervised and language-independent approach to identify MWE types, by extracting common n-grams and then segmenting the corpus based on maximising word prediction. They then refine the resulting lexicon of MWE types (again based on their prediction-based decomposition method) and accordingly tag all their corresponding token occurrences in a corpus. They evaluated their approach by searching for exact matches of the identified segments with WordNet entries. This methodology might be more useful in the case of longer idiomatic expressions that is the focus of that study. Nevertheless for expressions with fewer words, the opacity of tokens limit the efficacy of such techniques. Cordeiro et al. (2016) employed a similar approach by extracting and filtering MWEs using MWEToolkit (Ramisch et al., 2010) and then located them in the test data of the SemEval 2016 shared task, DIMSUM (Schneider et al., 2016).

Fazly et al. (2009) proposed statistical measures of fixedness to discover verb-noun idiomatic construction types. In order to identify potential id-

iomatic expressions, they designed an unsupervised method drawing on the statistical measures for automatic acquisition of canonical forms. The idea is that a verb-noun idiomatic construction occurs in its canonical form with higher frequency than in any other syntactic form. Finally, to distinguish between the idiomatic and literal usages of expressions, they compared lexical and syntactic patterns of expression occurrences with the type-based knowledge derived from their corresponding canonical forms. Specifically, relying on the fixedness property ascribed to idiomatic expressions, an expression occurrence is idiomatic when it is more similar to an idiom's canonical form (i.e. either it occurs in one of the canonical syntactic forms or is distributionally similar to the representation of the canonical form). Otherwise it is literal. This approach is more related to disambiguation between different usages of MWEs which is further explained in Section 2.5.

Other studies model the task either as classification or tagging in running text. In order to train a supervised approach it is a requirement to have large enough data annotated for MWEs. Most of the work that model the task as classification (Katz and Giesbrecht, 2006; Hashimoto and Kawahara, 2008; Salton et al., 2016) mainly focus on disambiguating between literal and idiomatic usages of expressions which is detailed in Section 2.5. With the increasing use of sequence labelling and structured prediction approaches in NLP, recent work on MWEs is also moving towards using them in tagging or segmenting corpora (Constant and Tellier, 2012; Schneider et al., 2014a).

2.4 Classification and Tagging Models

In this section, we describe the machine learning models, algorithms, and resources used in this thesis.

2.4.1 Word Representation

The first step in most NLP models (supervised or unsupervised, classification or tagging, type-based or token-based) is to represent words with numerical features. Distributional representation is one of the pioneering ideas based on Firth (1957). In the models based on this idea, words are represented as vectors in a high dimensional semantic space, where each dimension corresponds to a (context) word and the values are based on statistical analysis of the co-occurrences of target words with context words. These models have the following parameters.

Context type: The context of a token can simply be the neighbouring words of its token-level occurrences. However, to enrich the contextual information and reduce the effect of polysemy, we can also consider other information from co-occurring words such as part-of-speech tags or dependency relations. We can also decide whether to ignore some of context words such as stopwords that are frequent and carry little or no semantic content, such as determiners.

Context window size: The number of neighbouring words around a target can also be tuned. The context scope can be a sentence, a paragraph or even a document. It can also be based on a context window of specific

number of words either on the left side, or on the right side or on both sides of the target word.

Context vector values: The values of a distributional vector represent the degree of association between the target and context words. This association can be measured using raw co-occurrence frequency, binary co-occurrence value, or any of the association measures as defined in Section 2.3.1.1.

2.4.1.1 Vector Space Models

Vector Space Models (VSMs) are promising approaches in distributional semantics (Turney and Pantel, 2010). Since the dimension size and therefore the vector size in these models usually end up being very large, other methodologies are devised for dimensionality reduction while preserving the necessary information. These include singular value decomposition (SVD) used by Schütze (1998) or latent semantic analysis (LSA) proposed by Deerwester et al. (1990). The studies that use these methodologies in MWE identification are discussed both in Section 2.3.1.2 and Section 2.5.

2.4.1.2 Word Embeddings

Word embedding is the name for language modelling based methodologies that still follow the principles of distributional semantics, but aim at learning low-dimension vectors of real numbers. In practice they can be derived by feeding one-hot [6] or randomly initialised vectors into a neural network and

[6]One-hot vector encoding of a target word is a vector of dimension size equal to the number of all possible words (in the dictionary). All entries of the vector are zero except

updating the weights in subsequent iterations. Dense vectors of this kind have only a few hundred dimensions which makes them more practical due to the decrease in the amount of memory required to train them.

These models are originally derived from neural network language modelling techniques (Bengio et al., 2003; Collobert et al., 2011). Various methods are proposed to learn these mappings (Pennington et al., 2014; Lebret and Collobert, 2014), however, introduction of the efficient approach, `word2vec`, proposed by Mikolov et al. (2013c), brought this particular variant into widespread use. According to Mikolov et al. (2013a) these models perform significantly better than LSA and are also computationally less expensive. One standard implementation of `word2vec` is provided by Gensim (Řehůřek and Sojka, 2010).

`word2vec` uses two new neural network architectures to accomplish word representation learning, namely, the Skip-gram model and the Continuous Bag Of Words (CBOW) model. In both cases, a feed forward neural network is used where the standard non-linear hidden layer in neural network language models is removed and a projection layer is shared for all words. The Skip-gram model receives the target word type as input to predict the context. On the other hand, the CBOW model receives the context as input to predict the target word type. Skip-gram model is further improved by computing hierarchical softmax probability, and using negative sampling

for the single entry corresponding to the target word itself which is assigned the value of one.

and subsampling (Mikolov et al., 2013c).

Softmax (a generalisation of the logistic function) is a normalised exponential function which is used to model probability distributions. In the case of Skip-gram word embedding, given the words w_O and w_I, it is defined as $p(w_O|w_I) = \frac{exp(v'^T_{w_O} v_{w_I})}{\Sigma exp(v'^T_w v_{w_I})}$, where v'_w and v_w are the "input" and "output" vector representations of w. In negative sampling, the model is given noise words from a noise distribution to distinguish the target word from a noise word. In the case of subsampling, the intuition is that frequent words usually provide less information than rare words and therefore the probability of dropping a training example was proportional to the frequency of the occurrence of the target word.

In `word2vec` the vector size and the context window size are again parameters that should be defined beforehand. In terms of context type there is a recent work by Levy and Goldberg (2014) that generalises `word2vec` Skip-gram by replacing the word contexts with arbitrary contexts. In that specific study, they used dependency structures as arbitrary contexts to train the model which is called `word2vecf`. This approach will be further explained in Section 5.3.1.

2.4.2 Classification Methodologies

In this subsection, we briefly describe some of the classification methodologies that we use in this thesis.

Support Vector Machine (SVM) is one of the most commonly used

standard ML algorithms with competitive results in most NLP tasks. Like any supervised model, it is trained to recognise patterns in training data; it then uses the patterns to predict labels for test instances. As an example in our task in Chapter 4, the labels are `idiomatic` or `non-idiomatic` and the features (representing the patterns) are context word vectors. The features and the label for each data instance are used in an optimisation formula in finding the optimal hyperplane (with slope, w, and intercept, b) that provides the largest separation (margin) between instances of the two classes (See Figure 2.2 for the case of linear SVM optimisation). Given training vectors $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, in two classes, and a vector $y \in \{1, -1\}^n$, SVM solves the optimisation function as follows:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{n=1}^{n} \zeta_i \tag{2.6}$$

subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i$, $\zeta_i \geq 0$, $i = 1, ..., n$, where ϕ is a function that maps training instances into a higher (maybe infinite) dimensional space.

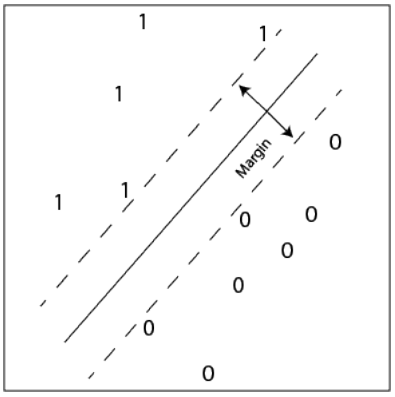


Figure 2.2: SVM optimisation

Logistic Regression (LR) is an algorithm that is based on a sigmoid (a.k.a logistic) function and predicts the probability of an instance belonging to the default class in a binary classification task.

Input values (x) are combined linearly using weights or coefficient values (b) to predict an output value (y). Equation 2.7 defines the logistic function,

$$y = \frac{e^{(b_0+b_1x)}}{(1+e^{(b_0+b_1x)})} \tag{2.7}$$

where y is the predicted output, b_0 is the bias or intercept term and b_1 is the coefficient for the single input value (x). Each column in the input data has an associated b coefficient (a constant real value) that must be learned from the training data. Learning coefficients is done using maximum-likelihood estimation (e.g. gradient descent). For the details of optimisation algorithms used for LR refer to Friedman et al. (2001).

Naïve Bayes Classifier (NBC) is a simple probabilistic supervised classifier based on applying Bayes' theorem with strong (naive) independence assumption between features. The probability function for the Gaussian Naïve Bayes algorithm is represented in Equation 2.8.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}) \tag{2.8}$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

Decision Tree (DT) is a supervised model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Given training vectors $x_i \in R^n$, $i = 1, \ldots, l$ and a label vector

$y \in R^l$, a decision tree recursively partitions the space such that the samples with the same labels are grouped together. The parameters are selected by minimising the impurity.

Random Forest (RF) is an ensemble learning model that operates by constructing a multitude of decision trees at training time and using averaging to improve the predictive accuracy and control over-fitting.

Neural Network Models are artificial networks that try to loosely model the way human brain works. The use of neural network models has recently received much attention in NLP. The feedforward neural network is one of the first and simplest types of artificial neural networks devised. It is made up of three components: the input layer, the hidden layers, and the output layer (see Figure 2.3). The features are the input to the network; this input is then multiplied by a weight matrix and passed through an activation function.

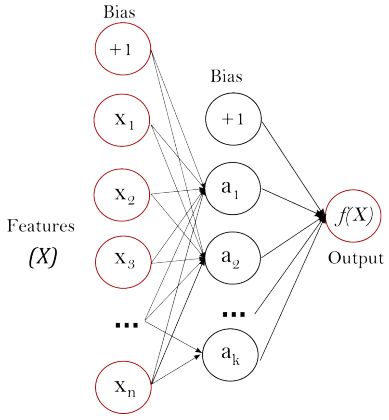


Figure 2.3: A simple neural network with one hidden layer

Multi-layer perceptron (MLP) consists of multiple layers of computational units, usually interconnected in a feed-forward way. In many applications the units of these networks apply the sigmoid function as activation. It is different from logistic regression, in that between the input and the output layers there can be one or more non-linear (so-called hidden) layers. The most common learning technique used in multi-layer networks is back-propagation in which the values of some predefined error-function is computed by comparing output values with the correct answer. The error is then fed back through the network and the weights are updated to reduce the value of the error function. This process is repeated for a large number of iterations so that the error converges to a small amount.

One popular non-linear optimisation function to update weights is gradient decent. In gradient decent the derivative of the error function with respect to the network weights is calculated and subtracted from weights such that the error decreases. The final weights are multiplied by input values to compute the output predictions.

2.4.3 Deep Neural Network Models

Recently popular deep neural network (Deep Learning) models can be used in both classification and tagging. They contain a larger number of hidden layers and also feedback connections between their components.

Recurrent Neural Networks (RNNs) are the most promising family of the

state-of-the-art neural networks, especially when we have sequential data in which the relationship between the components (words or sentences in the case of language processing) matters. The architecture of these networks are represented in Figure 2.4.

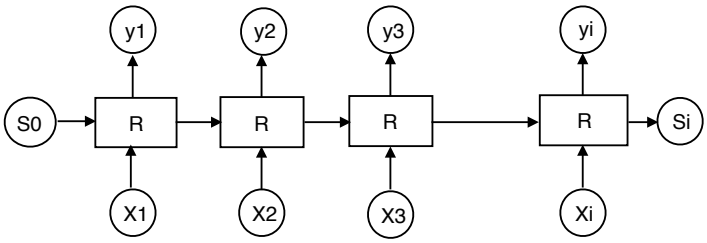


Figure 2.4: Graphical representation of a simple RNN.

The output of RNN networks are often vectors that can be fed into other network components that will try to predict final labels. In this sense RNNs are trained to produce informative representations for upper layers, i.e. they are used as 'feature extractors' (Goldberg, 2017). RNNs allow representation of arbitrarily sized sequential inputs in fixed-size vectors, while paying attention to the structured properties of the inputs.

In a high-level abstraction, as shown in Figure 2.4, the RNN is a function that takes as input an arbitrary length ordered sequence of n $d_{in} - dimensional$ vectors $x_{1:n} = x_1, x_2, \ldots, x_n$ and the initial state s_0, and returns as output a single d_{out} dimensional vector y_n. Each unit, R, takes as input a state vector s_{i-1} and an input vector x_i and returns a new state

vector s_i (Goldberg, 2017).

$$\begin{aligned} RNN(x_{1:n}; s_0) &= y_{1:n} \\ s_i &= R(s_{i-1}; x_i) \end{aligned} \tag{2.9}$$

The function R is the same across the sequence positions, but the RNN keeps track of the states of computation through the state vector s_i. RNNs are trained like any neural network by adding a loss function and using the back-propagation algorithm to compute the gradients with respect to that loss.

RNNs have different variations; the RNN-based architecture that we focus on in this study is LSTM (Hochreiter and Schmidhuber, 1997) which is a gated architecture devised to unravel the vanishing gradients problem (Pascanu et al., 2012). **LSTM** stands for **Long Short Term Memory** networks which provide more controlled memory access. In LSTM, the state vector s_i is split into two halves, where one half is treated as 'memory cells' and the other is working memory. At each input state, a gate is used to decide how much of the new input should be written to the memory cell, and how much of the current content of the memory cell should be forgotten. The architecture is represented in Figure 2.5 and the mathematical computations are detailed in Equation 2.10.

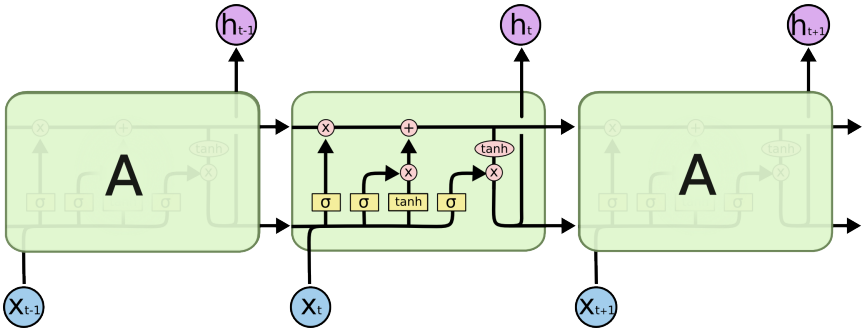


Figure 2.5: One of the components of LSTM architecture. The image is from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

$$
\begin{aligned}
s_t = R_{LSTM}(s_{t-1}, x_t) = [C_t; h_t] \\
C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \\
i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \\
f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \\
\tilde{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_C) \\
o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \\
h_t = o_t * tanh(C_t)
\end{aligned}
\tag{2.10}
$$

As can be seen, in each component of an LSTM network (A), instead of a single neural network layer, there are four. These layers are interconnected in a systematic way, controlling the information stored in or forgotten from memory. LSTM also has many small variants, exposition of which is out of the scope of this thesis. The Keras software package (Chollet et al., 2015) implements its most standard form which we use in the experiments in this thesis. The combination of two LSTMs that traverse the sequential data in

opposite directions is called bidirectional LSTM (biLSTM) and has proven to be very effective in language processing tasks. More details on the settings that we choose will be provided in the experiment sections of relevant chapters in this thesis.

One other recently popular neural network model which is also an effective feature extractor is **Convolutional Neural Network (CNN)**. A convolutional neural network is a combination of layers that function as convolving filters over local features in a large structure in order to capture important information for the prediction task. A CNN usually includes two consecutive operations: convolution and pooling. The convolution operation can be seen as a filter (function over each instantiation of a k-word sliding window) that passes through the input sentence.

The architecture of a convolution layer on a sample sentence is presented in Figure 2.6 which is from Goldberg (2017).

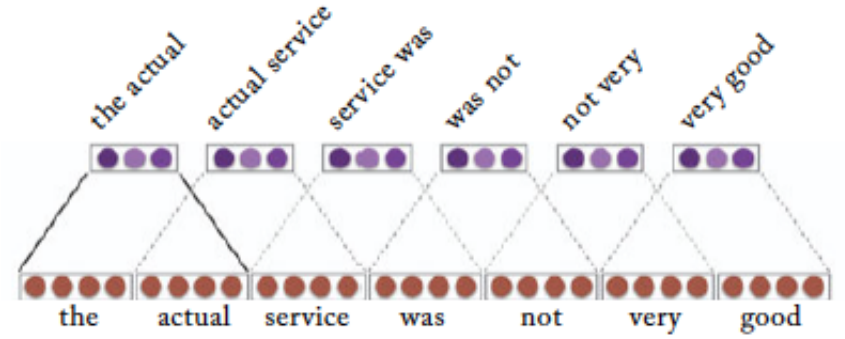


Figure 2.6: A narrow convolution with a window of size $k = 2$ and 3-dimensional output ($l = 3$), in the vector-concatenation notation.

Subsequently, a pooling operation is optionally used to combine the vectors resulting from the different windows into a single l-dimensional vector. This is achieved by taking the max or the average value observed in each of the l dimensions over the different windows. Pooling is usually used to compress or subsample the input (Hu et al., 2014). Since we do not want to filter out any information, in this thesis, we do not use pooling in our convolutional neural network layers and we refer to the network as **ConvNet**.

According to Kim (2014), if we consider $x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n$ to be a sentence of length n, where x_i is the k-dimensional word vector corresponding to the i-th word in the sentence and $\oplus$ is the concatenation operator, a convolution operation is defined as follows. Let $x_{i:i+j}$ refer to the concatenation of words x_i , x_{i+1}, ... , x_{i+j}. A convolution is a filter w which is applied to a window of h words to produce a new feature. In equation 2.11, a feature c_i is generated from a window of words $x_{i:i+h-1}$.

$$c_i = f(wx_{i:i+h-1} + b). \tag{2.11}$$

Here, b is a bias term and f is a non-linear function such as the hyperbolic tangent.

This convolution can be applied over the text resulting in m vectors $c_{1:m}$. It is also possible to apply multiple filters with different size and step sizes to allow the ConvNet to detect multiple features.

One of the most widely used works on using CNN in NLP is by Collobert and Weston (2008) in which they use CNN to predict part-of-speech tags,

chunks, named entity tags, semantic roles, semantically similar sentences, and also learn a language model.

2.4.4 Tagging Methodologies

Other than binary and multiclass classification, machine learning includes more complex structured prediction problems in which a label depends not only on features extracted for its corresponding data instance, but also on the predicted labels for other instances around. One example that illustrates the importance of this type of learning is the POS tagging problem in which many words are members of multiple parts of speech. The correct label for a word can often be deduced from the correct label of the word to the immediate left or right. For instance, the word *book* can be either a verb or a noun. In the sentence *I book this flight to the UK every month*, the word *I* is unambiguously a pronoun, and *this* a determiner. Using either of these labels, *book* can be deduced to be a verb, since nouns very rarely follow pronouns and are less likely to precede determiners than verbs. However, in the sentence *I bought that book*, the word *that* is unambiguously a determiner and can be used to infer that *book* is most probably a noun.

Structured prediction models such as tagging are widely modelled using probabilistic graphical models in which given input sequence $X_1, \ldots, X_n$, the values for output sequence $Y_1, \ldots, Y_n$ should maximise the following probability distribution (Equation 2.12).

$$p(X_1 = x_1, \ldots, X_n = x_n, Y_1 = y_1, \ldots, Y_m = y_m) \tag{2.12}$$

The key idea is to represent the family of the probability distribution using a graph. The vertices of the graph are random variables and edges represent statistical dependencies between two random variables.

In order to learn the structure of sequences (i.e. build a sequence labelling model), we need to have access to annotated datasets of running texts. Annotations in this case usually follow a specific scheme which is based on the so called IOB (short for inside, outside, beginning) tagging format. This format divides sentences into chunks of words. Labelling a word O means the word is outside any chunk. Labels B and I are often used as prefixes in tags where $B-$ indicates that the tag is the beginning of a chunk, and an $I-$ means that it is a continuation of a chunk. This tag representation scheme is originally presented by Ramshaw and Marcus (1999) for text chunking. Many variations of this format are subsequently discussed in the literature (e.g. IOE, IOBES), some including additional tags that differentiate Inside (I) and End (E). There is no consensus as to which scheme is better in general (Collobert et al., 2011).

Generative models such as Hidden Markov Models (HMMs) and Probabilistic Context-Free Grammars (PCFGs) are widely used to predict structures in NLP. However, the conditional model, linear-chain **Conditional Random Field (CRF)** has recently shown better performance in NLP (Yu et al., 2010). CRFs which are a type of discriminative undirected probabilis-

tic graphical models were introduced by Lafferty et al. (2001) for sequence labelling. Varieties of CRFs exist for different structured prediction problems (Turian et al., 2010).

In CRF, instead of modelling the joint probability $p(X, Y)$, the conditional probability $p(Y|X)$ is computed. Equation 2.13 defines the probability distribution over label sequences.

$$p_w(Y = y|X = x) = \frac{exp \sum_{i=1}^{n+1} w^T f(x, y_{i-1}, y_i, i)}{\sum_{y' \in \Lambda^n} exp \sum_{i=1}^{n+1} w^T f(x, y'_{i-1}, y'_i, i)} \tag{2.13}$$

where matrix w includes the weights used to construct a linear combination of the feature vectors computed using f. In order to learn the best w, the maximisation of this probability is generally performed using parameter estimation algorithms.

The features used in CRF are usually the word surface form, lemma, part of speech, and any shape, spelling, or morphological features of a word and its adjacent words. Turian et al. (2010) pioneered a successful work on augmenting CRF with word representation features to be used for chunking and named entity recognition. Linear scoring functions in CRF can now be replaced with neural networks (Goldberg, 2017).

2.5 Applications of Classification and Tagging in MWE Token Identification

Studies on token-based identification of MWEs start with disambiguating between different interpretations of expressions in their individual usages (Katz and Giesbrecht, 2006; Hashimoto and Kawahara, 2008). For instance, the expression *break the ice* has both literal and idiomatic interpretations. This sort of investigation requires more specialised corpora which feature sufficient instances of expressions in their different idiomatic/literal interpretation.

In this section, we first provide a review of the works that take this ambiguity into consideration. Next, we deal with more recent studies that consider identifying MWE tokens in general.

In order to identify the idiomaticity/noncompositionality of MWEs, Katz and Giesbrecht (2006) relied primarily on the local context of a token without considering linguistic properties of expressions. They represented different occurrences of an expression using LSA vectors and showed that the vectors of the expressions in their idiomatic sense are very different from those of the same expressions in literal sense. Based on this observation they classified a test expression token depending on whether it is more similar to the idiomatic or literal sense in training data.

Hashimoto and Kawahara (2008) have framed the task of idiom identification as sense disambiguation. After constructing a big corpus of Japanese idioms, they adopted a standard method in word sense disambiguation (WSD).

Specifically, they utilised SVM with a quadratic kernel along with the features commonly used in WSD. The features include surrounding word forms, their lemma and POS tags. Surrounding words were to some extent chosen based on Japanese grammar. Birke and Sarkar (2006) also adapted a slightly modified version of an existing word sense disambiguation algorithm for discriminating between literal and non-literal usages of verbs.

Fazly et al. (2009)'s work which is explained earlier in Section 2.3.2 also regards disambiguating between literal and idiomatic usages of expressions. While Fazly et al. (2009) have mainly focused on lexical and syntactic features to classify idiomatic and literal usages of expressions, most other works leverage context features (Katz and Giesbrecht, 2006).

Peng et al. (2014) proposed a topic modelling based approach in which they modelled the topics of text segments using LDA. They classified an expression usage in a given text segment as literal or idiomatic by using the topic term document matrix to project the expression into a topic space representation. Outliers within the topic space are labelled as idiomatic. They evaluated the efficacy of the approach on four expression types from the VNC-Tokens corpus (Cook et al., 2008). Experimenting with the same dataset, Salton et al. (2016) developed a distributional vector representation model to discriminate between sentences containing literal and idiomatic verb-noun expressions. They encoded the sentences containing expressions into their Sent2Vec (which uses RNN) distributed representations. They then employed KNN and different variations of SVM both for classifying sentences

of each expression separately and for general classification of all sentences.

In most of these works including Katz and Giesbrecht (2006), Hashimoto and Kawahara (2008) and Salton et al. (2016), the experiments and classifications are performed per expression and the evaluation results are reported for each expression individually.

Scholivet and Ramisch (2017) recently have tried to disambiguate a number of opaque French expressions using their contexts. They have proposed a CRF tagging approach using unigram and bigram features of the word forms and their POS. Tu and Roth (2011) have particularly considered the problem of in-context analysis of light verb construction (as a specific type of MWEs) using both statistical and contextual features. Their approach is supervised, but requires parsed data from English. Their contextual features include POS tags of the words in context as well as information from Levin's classes of verb components. These features have been reported to be particularly effective in recognising candidate LVCs whose surface structures are similar in both LVC and literal usages, as in their example below:

1. *He **had a look** of childish bewilderment on his face.*

2. *I've arranged for you to **have a look** at his file in our library.*

More recent studies focus on identification of MWE tokens in general or chunking a whole text for MWEs.[7] DiMSUM (Schneider et al., 2016) and

[7]The focus of these studies is not particularly on disambiguating between literal and idiomatic tokens, even though such systems might inherently take disambiguation into account.

PARSEME (Savary et al., 2017) are two notable workshops indicative of this recent interest in tagging corpora for MWEs.

Schneider et al. (2014a) have proposed a feature-rich sequence tagging model to identify all kinds of MWEs. The features used in their study are largely based on those of Constant et al. (2012), however they applied structured perceptron on the data tagged with their own proposed annotation scheme. They chose structured perceptron over CRF for its speed. As an important component, this system takes advantage of a group of features from various external lexicons. Particularly, they showed that the highest weighted features in the system are related to lexicon matching and also proper names. One of the strengths of these kinds of tagging methodologies is that they are straightforwardly able to deal with MWEs with gaps.

The impact of external lexical resources in identification of MWEs has been previously investigated by Constant and Tellier (2012) in their CRF tagging system. Constant et al. (2012) and Constant and Tellier (2012) both modelled the task using CRF. The features they have used include lexicon-based features, collocation-based features, word and POS n-grams, lowercase forms of the words, word prefixes and suffixes, whether the token is capitalised, has a digit, or is hyphenated. In Constant et al. (2012), the main idea is to integrate MWEs in parsing. They followed two approaches: one is to first employ a CRF-based system to identify MWEs as a pre-parsing step; the other is parsing with a grammar including MWE identification and then re-ranking the output parses using MWE-dedicated features. Both

approaches have been proved to be useful based on this work. In Constant and Tellier (2012), they showed that different ways of integrating lexicon-based features in the CRF model improve the results and compensate the use of a small training data.

Constant and Nivre (2016) proposed a transition-based system which jointly predicts syntactic dependency structure and lexical units (e.g. MWEs) for each sentence. In a standard dependency tree, both syntactic and lexical structures make the assumption that words and lexical units are in a one-to-one correspondence. The existence of MWEs violates this assumption and new representations are required to treat them as some kind of lexical units. The transition-based model for parsing MWEs is based on the arc-standard transition system for dependency parsing, first defined in Nivre (2004). The system uses a greedy search parsing algorithm and a linear model trained with an averaged perceptron to learn the best scored (optimal) dependency tree based on the MWE labels in the training data. In the results, they showed the better performance of their system over the CRF based approach (Le Roux et al., 2014) and the combination of graph based and CRF based system (Candito and Constant, 2014).

Using a simplified version of the transition based system, Al Saied et al. (2017)'s system is ranked first in the PARSEME shared task on identifying verbal MWEs. They extracted the transitions and predicted their types (i.e. whether they are associated with MWEs or not) using an SVM classifier.

Qu et al. (2015) have found word embedding representation of the words

in context very useful for tagging a text for MWEs. For classifying each expression as MWE or not, we also use word vector representations corresponding to the verb and noun components, along with the words in a window size of two on the right side of the expression (Chapter 3).

Legrand and Collobert (2016) have proposed a neural network based approach that learns fixed-size representations for arbitrary sized chunks which is able to classify these representations as MWE or not. More specifically, in their system, vectors of the words in arbitrary sized windows are concatenated to form vectors of arbitrary sized chunks. Then every possible chunk vector is projected onto a common vector space and the resulting vector representations are passed on to a classifier to tag them as MWE or not. The proposed neural network is trained by maximising the likelihood over the training data, using stochastic gradient ascent. They have shown better performance in MWE identification over the CRF-based approach in Constant et al. (2013). They also showed that their system (without relying on any variation of IOB-based tagging) performs on par with the IOBES-based model of Collobert et al. (2011) applied to MWE tagging. However, the system does not deal with MWEs with gaps while most systems trained on IOB-based tags have the capacity to learn non-continuous MWEs.

Gharbieh et al. (2017) made the first attempt towards using deep learning to tag a corpus for MWEs and reported state-of-the-art results for tagging MWEs in DIMSUM (Schneider et al., 2016). They formulated the task as classification and their convolutional neural network with three hidden layers

performs the best out of all neural network based systems they have tried. It is worth noting that the model combines handcrafted features with the embedding representations as the input to the system.

The presented systems in shared tasks DiMSUM and PARSEME give a broader picture on token-based identification of MWEs, .

2.6 Translation

The effectiveness of integrating bilingual MWEs into Statistical Machine Translation (SMT) systems has been studied in the literature (Ren et al., 2009; Pal et al., 2011). Finding translations of MWEs is challenging due to all the previously listed idiosyncratic properties (especially their semantic non-compositionality) of such expressions as discussed in Section 2.1. Salehi et al. (2014) took advantage of the peculiar translation behaviour of MWEs in order to identify them. Bilingual Extraction of MWEs can also be useful in construction of bilingual resources and to accelerate the work of lexicographers (Corpas Pastor, 2017).

One of the most common approaches for extracting translations both for single-word and multi-word tokens is to use alignments derived from SMT (Tiedemann, 1998) on parallel corpora. However, parallel corpora are out of reach for many languages. This can be alleviated by using comparable corpora (McEnery and Xiao, 2007).

Several studies have suggested methods for extracting parallel segments from comparable corpora in various different tasks, including bilingual lex-

icon construction (Rapp, 1999; Fung, 1997; Ismail and Manandhar, 2010; Bouamor et al., 2013), and sentence alignment for improving SMT (Ion, 2012; Smith et al., 2010; Pal et al., 2014). Corpus-based distributional similarity has been used in a bilingual context to automatically discover translationally-equivalent words from comparable corpora (Pekar et al., 2006; Rapp, 1999; Fung, 1997). It is not clear, however, whether a similar approach can be used for finding translations of multiword expressions.

NLP systems that need to translate MWEs sometimes use pre-existing lexicons of collocation translations (Mendoza Rivera et al., 2013). However, such lexicons do not provide translations of all collocations, as new combinations are created and used on a daily basis. Thus, it is important to develop a method that can automatically find translation equivalents for multiword expressions.

Bouamor et al. (2012) used distributional models to align MWEs in order to improve performance of a machine translation system. However, their method relies on sentence-aligned (parallel) corpora. Rapp and Sharoff (2014) also investigated the use of word co-occurrence patterns across languages to extract translations of single and multiword terms. Like Ismail and Manandhar (2010) they avoid using a large initial bilingual dictionary. While their approach delivers good results in finding the translations of single words, they did not report good results for MWEs. Even for single words their results only cover words that are, according to their frequency patterns, salient (keywords).

With regard to finding translation equivalents for single words, the pioneering work of Mikolov et al. (2013b) involves a supervised scenario in which a translation matrix can be learned from two sets of monolingually trained word2vec vectors (Mikolov et al., 2013a). Embedding representations for MWEs are still in the preliminary stage (Kordoni, 2017).

2.7 Summary

This chapter first illustrated the definitions and characteristics of MWEs outlined in the literature. Next, we detailed the datasets and resources available for computational treatment of MWEs. Then we started with the task of automatic extraction and identification of MWEs and explained different methodologies designed to model them. Machine learning methods and their applications relevant to our study are extensively discussed in this chapter. Finally, we studied some related work on automatic extraction of translation equivalents for MWEs. This chapter will be referred to for further information in studies of the following chapters.

Chapter 3

Verb-Noun Multiword Expressions: Resources and Identification

In this chapter, we first outline the development of new language resources for Italian and Spanish, namely corpora annotated with Italian and Spanish MWEs of the class: verb-noun expressions such as *fare riferimento*, *dare luogo* and *prendere atto* in Italian and *tener lugar*, *dar guerra*, *tomar partido* and *dar asco* in Spanish. Such collocations are reported to be a very frequent and productive cross-lingual class of MWEs (Cook et al., 2008). This chapter is structured as follows. After describing the annotation scheme in Section 3.1, we outline the development of our datasets of MWEs 'out of context' in Section 3.2 and 'in context' in Section 3.3. Next, we describe the pilot experiments that we have performed on the datasets of MWEs listed out of context in Section 3.4. Subsequently, in Section 3.5, we present our new approach for token-based identification of MWEs and report the experimental results on the Italian data we have compiled. Finally, we summarise the chapter in Section 3.6.

3.1 Scheme

The goal of this project is to investigate verb-noun MWEs in three languages: English, Spanish and Italian. There are some resources and datasets available for English, however, the need for such types of resources is greater for Italian and Spanish which do not benefit from the variety and volume of resources accessible for English. We compile new MWE datasets for Italian and Spanish and made them openly accessible to the research community.

In this study we focus on specific verb-noun combinations with no gaps between the components. This makes extensive investigation of these expressions possible. According to PARSEME multilingual corpus of verbal MWEs (Savary et al., 2018), more than 80% of verbal MWEs in Italian and almost 70% of verbal MWEs in Spanish are continuous (occur without any gaps). We also restrict the experiments to highly polysemous verbs, such as *give*, *take*, and *make* which have high occurrence in the formation of MWEs and exhibit a broad range of figurative meanings. The MWEs following this structure are either idioms or light verb constructions. We do not differentiate between these, instead we focus on their common characteristic of having no transparent meaning.

For the purpose of this study, we choose itWaC corpus (Baroni and Kilgarriff, 2006) for Italian. The corpus is made up of texts collected from the internet and is 1.5 billion words. We use the corpus provided by the SketchEngine (Kilgarriff et al., 2004) in which the corpus is tagged and lem-

matised with the TreeTagger tool (Schmid, 1995). For Spanish, we use the SpanishWaC, also available from SketchEngine. This corpus is gathered from the internet, tagged and lemmatised with the TreeTagger tool and includes roughly 98 million words (almost the size of the BNC corpus).

We compile two different resources: 1) lists of MWEs annotated out of context with a view to performing fast evaluation of the developed methodology (out-of-context mark-up) and 2) annotated MWEs along with their concordances (in-context annotation). The latter type of annotation is time-consuming, but provides the contexts for the annotated MWEs. The details of the two annotation exercises are further explained in the following section.

3.2 Annotated Lists

We first automatically compiled a list of verb-noun expressions, to be annotated by human experts. This is based on previous attempts at extracting a lexicon of MWEs, as in Villavicencio (2005). Annotators were not provided with any context making the task more feasible in terms of time. Human annotators were asked to label the expressions as MWEs only if they have sufficient degrees of idiomaticity. In other words, a verb-noun MWE does not convey literal meaning in that the verb is delexicalised.

In this phase, an expression type (rather than token) was evaluated based on its potential literal/idiomatic interpretations. For example in the expression *take a break*, the light verb *take* does not carry literal meaning and is an MWE, while *have coffee* will be marked as literal as the verb *have* bears the

literal meaning of drinking. Some expressions potentially have both kinds of interpretation e.g. *have a baby* which other than its transparent meaning, can have the idiomatic interpretation of *giving birth* in *she had a baby in ABC hospital*. We considered these three possibilities (literal, idiomatic, or both) for an expression type in this experiment.

The experiment initially covered Spanish and was performed after having tested the guidelines for two rounds of pilot annotations in English. More specifically, we first extracted verb-noun expressions in English from BNC and in Spanish from SpanishWaC. Two annotators marked up the expressions. By analysing the disagreement between annotators, and based on their feedback, we improved and finalised the guidelines. The finalised annotation task involves three tags: tag 1 (MWE) if the expression is idiomatic; tag 0 (non-MWE) if the expression is literal. We also introduced tag 2 for the expressions which in some contexts behave as MWEs and in others not, e.g. *have children*, which in some contexts means *to give birth* and hence can be an MWE. The finalised guidelines were applied to Italian annotations.

We focused on four of the most frequent verbs: *fare*, *dare*, *prendere* and *trovare* in Italian and *tener*, *hacer*, *formar* and *tomar* in Spanish.[1] Using SketchEngine (Kilgarriff et al., 2004), we extracted all the occurrences of these verbs when followed by any noun, from the itWaC corpus for Italian

[1]The verbs were selected from the list of most frequent light verbs, by native-speaker annotators who are knowledgable of the MWE phenomena. The idea was to cover as much productive and ambiguous verb-noun expressions as possible. The translations for the light verbs are not provided, since they are polysemous verbs with several different potential translations.

and SpanishWaC for Spanish. All extracted verb-noun expressions had the verb lemmatised.

After removing all occurrences with a frequency lower than 20, the extraction of verb-noun candidates featuring the above four verbs in Italian resulted in a dataset of 3,375 expressions. Two native speakers annotated every candidate expression with 1 for an MWE if the expression is idiomatic, with 0 for a non-MWE if the expression is literal, and 2 for the expressions which in some contexts behave as MWEs and in others do not. An example of this is *dare frutti*, which has a literal usage that means *to produce fruits* but in some contexts means *to produce results* and is an MWE.

The observed agreement between the annotators was 0.73. The observed agreement is simply the percentage of expressions annotated identically by the two annotators, without considering the chance agreement. Different coefficients have been defined in order to calculate chance-corrected agreements (Artstein and Poesio, 2008). In the case of this study with only two annotators, we did not register much variation between different coefficients. Therefore we chose to report the most common measure which is Cohen's Kappa coefficient (Cohen, 1960). The Kappa measure was only 0.40 for the annotation of Italian out-of-context expressions.

For Spanish, since the corpus is smaller, we set a lower threshold and removed the expressions with frequencies lower than 10. In several rounds of pilot annotations, we observed insufficient inter-annotator agreement. In the end, the first annotator marked up all 1,924 expressions and the sec-

ond annotated almost 33% (623 expressions), according to which the Kappa agreement was measured to be 0.36 and the observed agreement was 0.66.[2]

Even though this out-of-context 'fast track' annotation procedure saves time and yields a long list of marked-up expressions which could be useful in upstream NLP tasks, the results of this kind of annotation are not promising enough. Annotators often feel uncomfortable due to the lack of context. The low rate of agreement between annotators is indicative of the challenge. Also, we believe that idiomaticity is not a binary property; rather it is known to fall on a continuum from being completely semantically transparent, or literal, to entirely opaque, or idiomatic (Fazly et al., 2009). This makes the task of out-of-context marking-up of the expressions more challenging for annotators, since they have to choose a value according to all possible contexts of a target expression. This difficulty and the fact that there are many expressions that in some contexts are MWEs and in some contexts not, prompted us to initiate a subsequent annotation task and data preparation where MWEs are tagged in their contexts.

3.3 In-context Annotated Expressions

To annotate MWEs in context, in some previous studies the corpus is tagged for all possible occurrences of MWEs (Schneider et al., 2014a), or at least a focused category of them (Savary et al., 2017). Although the results would

[2]Since the inter-annotator agreement was insufficient and did not improve, we continued with annotations of only the first annotator for Spanish and we use the second annotation only for reporting the inadequate agreement.

be very interesting, given the time-consuming and labour-intensive nature of this task, it would not be feasible to annotate large corpora for all numerous variations of usages that a particular type of MWE could have.

We designed an annotation task, in which we provided a sample of all usages of any type of verb-noun expression constructed from the four verbs in focus. These usages were annotated as being MWE or not. The idea was to extract the concordances[3] around all the occurrences of a verb-noun expression and provide annotators with their context in order to be able to decide the degree of idiomaticity of the specific verb-noun expression.

For this purpose, we employed SketchEngine to list all the concordances of each verb when it is followed by a noun. We focused on four verbs in Italian and four verbs in Spanish as explained in Section 3.2. Each concordance included a verb in focus with almost ten words before and ten words after it. The SketchEngine returned $100,000$ concordances for each query. The query in our case was the verb in its lemmatised form when there was a noun in the window of one word after (on the right side of) the verb. We filtered out the concordances that include verb-noun expressions with frequencies lower than 50 for Italian and lower than 10 for Spanish, and randomly selected 10% of the concordances for each verb in both languages. Figure 3.1 shows one sample of a query in SketchEngine when we extract concordances for the verb *dare*.

[3]A concordance is a listing of each occurrence of a word (or pattern) in a text or corpus, presented with the words surrounding it (Wynne, 2008).

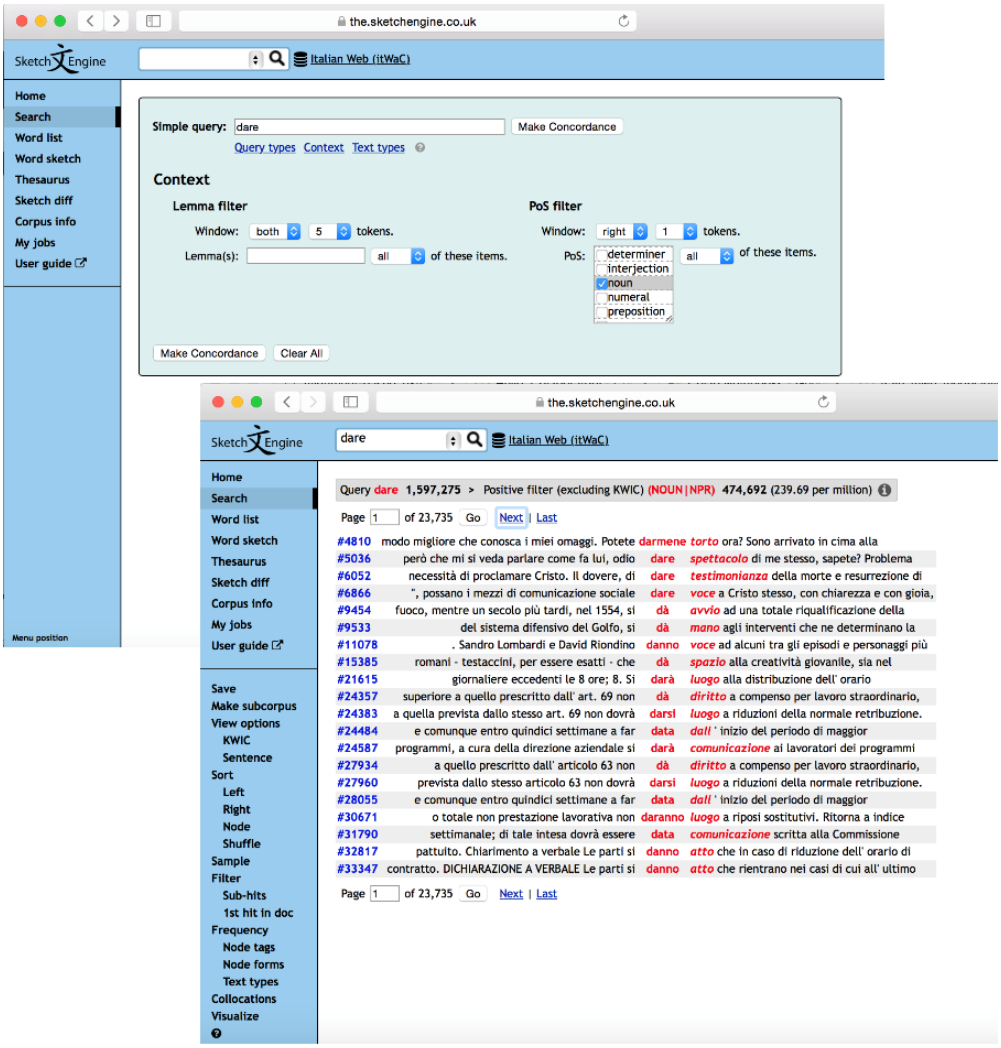


Figure 3.1: The SketchEngine interface for querying the verb *dare* with any noun in the context window of 1 on the right.

As a result, for Italian there were 30,094 concordances to be annotated. The two Italian annotators annotated all usages of verb-noun expressions in these concordances, considering the context that the expression occurred in, marking up MWEs with 1, and expressions which were not MWEs with 0.

We have noticed that a small number of very frequent expressions make up a large portion of our data (this is in accordance with Zipf's law). Examples are *prendere atto*, *fare parte*, *prendere parte*, *dare vita*, etc. These expressions have been almost always annotated as MWEs. Since these expressions are strongly salient, we believe it would be easy for any statistical system to identify them. Therefore they can be excluded from the annotation in order to collect and investigate more ambiguous expressions.

Accordingly for Spanish, we excluded the five most frequent expressions so that more time made available for annotators to mark up the more challenging expressions. As a result, the two Spanish native speakers annotated $3,965$ concordances. More details of this annotation task and the agreement rates are reported in Table 3.1. As seen in Table 3.1, the inter-annotator agreement was significantly higher when annotating the expressions in context compared to out-of-context annotation.

Table 3.1: Inter-annotator agreement for in-context annotation

	# of concordances	Kappa	Observed agreement
Italian	30,094	0.65	0.85
Spanish	3,965	0.55	0.79

In order to resolve the disagreement among annotators for the Italian data, we employed a third annotator who decided on the majority of cases of disagreement. This resulted in the finalised number of $20,030$ concordances. For Spanish, we ignored all cases of disagreements and considered only the

concordances on which both annotators agreed. These amount to 3,090 expressions.

One important feature of this dataset is that it represents various usages of each expression type. There are a number of expression types that, according to the annotations, occurred in both idiomatic and literal usages. For example, for Italian, according to the first annotator, among the 1,649 types of expressions in concordances, 530 (32%) of them are MWEs in some occurrences and non-MWEs in others. We propose an approach for automatic investigation of these cases in Section 3.5.

3.4 MWE Extraction

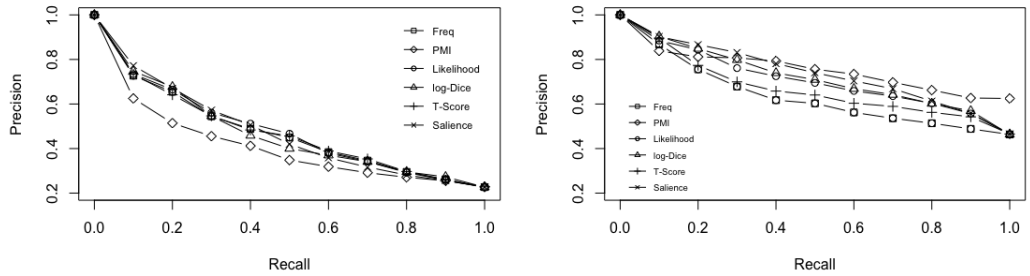
The most traditional approach to automatically recognise MWEs in any language is by examining corpora (Evert and Krenn, 2005; Ramisch et al., 2010; Villavicencio, 2005). Corpus based Association Measures (AMs) have commonly been used for MWE extraction (Ramisch et al., 2010). These measures have been proposed to determine the degree of compositionality, and fixedness of expressions. The more non-compositional or fixed an expression is, the likelier it is to be an MWE (Evert, 2008; Bannard, 2007). According to Evert (2008), there is no ideal AM for all purposes. We evaluate AMs as a baseline approach against the annotated data which we prepared. In this study we focus on fives AMs which are widely discussed to be among the best in identifying MWEs as explained in Chapter 2, Section 2.3.1.1. These are: PMI, log-likelihood, T-score, log-Dice and Salience, all as de-

fined in SketchEngine. We compare the performance of these AMs, and also frequency of occurrence (Freq) as the sixth measure, to rank the candidate MWEs. We evaluate the effect of these measures in ranking MWEs.

In the first experiment, the list of all extracted verb-noun combinations (as explained in Section 3.2), are ranked according to the above measures computed from itWaC for Italian and SpanishWaC for Spanish as reference corpora. With a view of performing the evaluation against the list of annotated expressions for Italian, we process all $2,415$ expressions for which the annotators agreed on tags 0 or 1. For Spanish, we process all $1,856$ expressions which the first annotator tagged with 0 or 1. After ranking the expressions by the AMs, we examine the retrieval performance of each AM by drawing the precision-recall curves. Precision-recall curves are known to be suitable for ranking retrieval tasks (Manning et al., 2008). Eleven-point Interpolated Precision (11-p IP) reflects the efficiency of a measure in ranking the relevant items (in this case, MWEs) before the irrelevant ones. To this end, the interpolated precision at the 11 recall values of 0, 10%, ..., 100% is calculated. As detailed in Manning et al. (2008), the interpolated precision at a certain recall level, r, is defined as the highest precision found for any recall level $r' >= r$ [4]. A composite precision-recall curve showing 11 points can then be graphed. Following the graph, we can see what the precision of every approach would be for different levels of recall. The results of these

[4]http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html

graphs comparing the different rankings are presented in Figure 3.2.



(a) (b)

Figure 3.2: 11-p IP of different measures for (a) Italian and (b) Spanish expressions.

According to Figure 3.2, the performance of different ranking scores are very close to each other for Italian, with Salience and Likelihood scoring the highest and PMI the lowest. However, for the Spanish data, PMI works the best followed by Salience and Likelihood. PMI, as explained in Chapter 2, is very sensitive to low frequency candidates and appears to not favour high frequency ones. This may be the reason why it does not work well for the Italian data where the corpus size is much bigger and the expressions in focus are of much higher frequencies. The difference in languages and selection of specific verbs with potentially different behaviour may also be an additional factor resulting in dissimilar performance level of some of the measures across languages (Evert, 2008).

To the best of our knowledge, this is the first time Salience has been applied to rank MWEs in these languages. We report its effectiveness on

ranking verb-noun MWEs in the languages in this study, and we use this measure further in our following experiments.

In order to identify usages of MWEs in context, the simplest approach is to first extract the canonical forms (type-based extraction of MWEs), and then locate different inflections of expression types. This approach has been employed by Cordeiro et al. (2016), in DiMSUM 2016 shared task for identification of MWEs and their system reported competitive results on the rather limited DiMSUM dataset. However, this approach ignores all expressions that have both literal and idiomatic usages. Investigating the literal and idiomatic usages of expressions in context is the focus of the following section.

3.5 Token-based Identification of MWEs

It is important to consider expressions at token level when seeking to establish whether they are MWEs. The reason is that there are expressions that in some cases occur with an idiomatic sense and with a literal sense in others. This could be determined by the context in which they appear. For example take the expression *play games*. It is opaque with regards to its status as an MWE and depending on context could mean different things. For example in the following sentences, in 1) it has a literal sense but is idiomatic in 2).

1) *He went to play games online.*

2) *Don't play games with me, I want an honest answer.*

A traditional classification model that does not have access to linguistic context proves to be insufficient in such cases.

Context of an expression has been shown to be discriminative in determining whether a particular token is idiomatic or literal (Fazly et al., 2009; Tu and Roth, 2011). However, in-context investigation of MWEs is generally an under-explored area. Fazly et al. (2009) use context information to identify syntactic restrictions of verb-noun MWEs. This approach is not applicable to the expressions considered in our study, which follow the specific structure of verb-noun expressions without any gaps. The expressions, which we analyse in this study, have various degrees of idiomaticity in the same structure. This makes the distinction between their idiomatic and literal usages more challenging.

Tu and Roth (2011) have particularly focused on the problem of in-context analysis of light verb constructions (as a specific type of MWEs), using both statistical and contextual features. Their approach is also supervised, but it requires parsed data from English. Their contextual features include POS tags of the words in context as well as information from Levin's classes of verb components. Our approach requires minimal pre-processing and is best suited to languages that lack ample tagged resources.

3.5.1 Our Proposed Context-based Approach

Our goal is to classify tokens of verb-noun expressions into literal and idiomatic categories. To this end, we propose a supervised approach that

utilises the context of the occurrences of expressions in order to determine whether they are MWEs. To extract context features we use word2vec, a state-of-the-art word embedding approach in distributional similarity (Mikolov et al., 2013c). Neural word embedding features extracted using unsupervised learning from unannotated corpora offer better generalisation than distributional word-level features (Turian et al., 2010; Collobert et al., 2011). By using word vectors, the training model is less prone to overfitting on exact word usages and would generalise to similar words.

We avoid employing language-specific features that are costly to develop in case of new languages and domains. We extract features from the raw corpus without any pre-processing. While we report the results for Italian, the approach is language-independent and can be applied to any resource-poor language.

Compared to literal verb-noun combinations, idiomatic combinations are expected to appear in more restricted lexical and syntactic forms (Fazly et al., 2009). One traditional approach in quantifying lexical restrictions is to use statistical measures (Ramisch et al., 2010). As our baseline approach, we employ statistical measures computed from expression components. Specifically, we focus on the best association measures based on our experiments in Section 3.4.

We represent our context features as follows. For each occurrence of an expression, we exploit the information contained in its concordance (context). Given each concordance, we extract vector representations for several of its

words to act as syntactic and lexical features.

We extract the vectors of the verb and the noun components in their raw form (i.e. their actual occurrences without performing any normalisation), with the intent to indirectly learn lexical and syntactic features for each occurrence of an expression. Our assumption is that the structure of the verb component is important in extracting the fixedness information for an expression. Also, the distributional representation of the noun component is informative since verb-noun expressions are known to have some degrees of semi-productivity (Stevenson et al., 2004).

Additionally, we extract vectors for co-occurring words around a target expression. We focus on the two words immediately following the verb-noun expression. The arguments of the verb and noun components that usually occur following the expression are expected to play a distinguishing role in these kinds of complex predicates (Samek-Lodovici, 2003). We expect verb arguments to occur in close vicinity on the right side of the expression. Example (1) shows the process of forming a feature vector for an example concordance of the expression *dare atto*.[5]

(1) del bilancio. Credo, quindi, che si debba <dare >atto alla Presidenza ed agli organi politici

$[vec(dare); vec(atto); vec(alla); vec(Presidenza)]$

[5]We experimented with window sizes of more than two and we observed no improvement.

The word vectors in this study come from the Italian word2vec embedding which is available online.[6] The generated word embedding is derived from Gensim's skipgram word2vec model with the window size of 10. It extracts vectors of size 300 for Italian words from Wikipedia corpus.

In order to construct our context features, given each occurrence of a verb-noun combination, we concatenate four different word vectors corresponding to the verb, noun, and their two following adjacent words. The concatenation preserves the original order, as in example (1). In other words, given each expression, the context feature consists of a combined vector with the dimension of 4 * 300 = 1200. Concatenated feature vectors are fed into a logistic regression classifier.

3.5.2 Experiment

We run our experiments on the Italian data that is described in Section 3.3.[7] In this experiment we consider the annotations of the first annotator as the gold-standard, as the agreement between the two is substantial.

We differentiate between expressions whose instances occur with a single fixed idiomatic or literal behaviour and the ones that show degrees of ambiguity in different potential usages. We partition the dataset in a way that accounts for both of these groups, and the experiments are run separately for each.

[6] http://hlt.isti.cnr.it/wordembeddings/

[7] For the Spanish data, we do not have enough number of expressions that have ambiguous literal/idiomatic occurrences on which both annotators agree.

3.5.2.1 Partitioning the Dataset

In this part of the experiment, the idea is to evaluate the effect of context features in order to identify the literal/idiomatic tokens of expressions, particularly for the type of expressions that are likely to occur in both senses. In our specialised data, around 32% of expression types have been annotated in both idiomatic and literal forms in different contexts. For this investigation, we divide the data into two groups:

(Group 1) Expressions with a skewed distribution of the two senses (e.g., with more than 70% of instances having either a literal or idiomatic sense).[8]

(Group 2) Expressions with a more balanced distribution of instances (e.g., with less than or equal to 70% of instances having either a literal or idiomatic sense).

We develop different baselines to evaluate our approach on these two groups as explained in the following sections.

3.5.2.2 Majority Baseline

We devise an informed and supervised baseline based on the idiomatic/literal usages of expressions in the gold-standard data. According to this baseline, a target instance vn_{ins}, of a test expression type vn, is assigned the label that it has received in the majority of vn occurrences in the gold-standard set. The baseline approach labels all instances of an expression with a fixed label (1 for MWE and 0 for non-MWE). This is a high precision model when working

[8]Expressions such as *dare inizio* 'to start' and *trovare cose* 'to find things' which most of the times occur as MWE and non-MWE respectively.

with Group 1, due to the consistent behaviour of instances there. However, this method is suitable to be used as a baseline for evaluating the results of our developed model over expressions of Group 2. A method that works better than this baseline demonstrates that it relies on linguisctic heuristics rather than following the majority of occurrences.

3.5.2.3 Association Measures as a Baseline

The data in Group 1 includes the expressions that mostly occur in either idiomatic or literal forms. These expressions are commonly categorised as being MWE or non-MWE based on association measures. These measures are computed by statistical analysis of the whole corpus, hence the values are the same for all instances of an expression. In other words, these methods do not have access to the contexts in which different instances of an expression could occur.

To evaluate our model over data in Group 1, these association measures are used as features to develop a baseline. We focus on two widely used association measures, namely, Log-likelihood and Salience, as defined in Chapter 2 and experimented with in Section 3.4. We also use frequency of occurrence as a statistical measure to rank MWEs. The statistical measures are computed using SketchEngine on the whole of itWac, and are then given to an SVM classifier to identify MWEs.

3.5.3 Evaluation

There are 1,480 types of expressions with 28,483 occurrences in Group 1 and 169 types of expressions with 1,611 occurrences in Group 2. For each group, we extract context features to train logistic regression classifiers.

Our proposed context features are vector representations of the raw form of the verb component, the raw form of the noun component, and a window of two words after the target expression. We refer to the combination of these vectors as the `Context` features, and apply a 5-fold cross validation approach to compute accuracies for each classifier. We split the dataset into five separate folds in a way that no instance of the same expression occurs in more than one fold. This is to make sure that the test data is sufficiently blind to the training data. The classifiers are compared against the baselines using different features. The results are reported in Tables 3.2 and 3.3.

3.5.4 Results and Analysis

Statistical measures are expected to be promising features for identifying MWEs among expressions with consistent behaviour. However, the results in Table 3.2 show that our `Context` features are more effective in MWE classification when applied over Group 1 and also for the entire data. Using the `Context` features alone shows statistically significant improvements over `Likelihood+Salience+Freq`, with $p < 0.05$ in Group 1 and $p < 0.001$ in all data.

The strong model performance with word context features leads us to

Table 3.2: Classification accuracies (%) using different features over Group 1 and the whole data.

Features	all data	Group 1
Freq	70.77	69.20
Likelihood	72.11	70.64
Salience	73.83	72.81
Likelihood+Salience+Freq	73.90	73.29
`Context` (word2vec)	**75.42***	**74.13**
Salience + `Context`	**78.40***	**80.13***
Likelihood+Salience+Freq+`Context`	**76.95***	**80.07***

believe that they contain information from external arguments of the verb and the noun constituents of expressions which helps boost classification accuracy. More experiments need to be done to confirm this and to find the best suitable window size for the word context around a target expression.[9]

We have also trained the logistic regression with the combination of the `Context` features and association measures in Table 3.2. According to the results, the combination improves the accuracy of our model in identifying idiomatic expressions, especially when applied to the consistent data in Group 1. The results lead us to believe that context features are even more useful in cases where we observe more consistent behaviour in the data and expect the best result from statistical measures. The better performance when using `Context` and statistical measures together, compared to when we use `Context` features alone, is also a remarkable observation visible in Table 3.2. This can be explained by the fact that, among all the data, iden-

[9]We have stablished through trial-and-error that a window size of two after a target expression leads to better results compared with no context or contexts of bigger size.

Table 3.3: Classification accuracies (%) over data in Group 2 compared to the majority baseline.

Model	Group 2
Majority Baseline	59.52
Logistic regression with `Context` features	**63.21**
Logistic regression with `Context`+Salience	54.37

tification of expressions that have skewed distribution of their interpretations (i.e, those which most of the times occur as either MWE or non-MWE) can still benefit from statistical measures as features. The accuracies marked by $*$ are for the cases that we see statistically significant improvement over the `Likelihood+Salience+Freq` baseline with $p < 0.001$.

Table 3.3 shows the results of our model for data from Group 2 compared to the majority baseline. Recall that the data instances in Group 2 are highly unpredictable in their occurrence as MWE or non-MWE. We expect that our supervised model using `Context` features be able to disambiguate between different instances of an expression. Here, our model (logistic regression with `Context` feratures) performs slightly better than the informed majority baseline.

Our experiment using the combination of `Context` and Salience (as the best statistical measure), for training over Group 2 expressions (Table 3.3), shows that the statistical measure is not helpful for the class of ambiguous expressions.

3.6 Summary

In this chapter, we have first described the compilation of datasets for processing verb-noun MWEs both out of context and in context. Then we have conducted experiments to rank expressions using different traditional statistical measures. Furthermore, we proposed a new approach for identifying the usages of idiomatic expressions in context. We applied the approach on the compiled Italian data, as explained in Section 3.3. We compared the results with baseline methodologies and outlined discussions on the experiments. We showed that in order to identify tokens of MWEs more effectively, lexical and syntactic context features derived from vector representations can be combined with traditional statistical measures.

Chapter 4

Modelling and Evaluation of Multiword Expressions in Context

As discussed in Section 2.3.2, automatic identification of Multiword Expressions (MWEs) in running text has recently received considerable attention from researchers in computational linguistics. In this chapter, we first discuss the two main approaches to framing the task of predicting MWEs in context: classification and tagging. We investigate why classification is more suitable than tagging for modelling MWEs in our data. Furthermore, the wide range of reported results for the task in the literature has prompted us to take a closer look at the algorithms and evaluation methods. We focus on the importance of train and test splitting and the distribution of expression types in validating the results, and propose an alternative method to perform train and test splitting.

4.1 Modelling MWE Identification

The focus of our study is on token-based identification of MWEs. The most evident solution is to go through the running text and tag any two or more words where the co-occurrence conveys idiomatic interpretation. However, it is not always feasible to traverse the whole of a large corpus. For this reason

we have gathered a specialised dataset of concordances of particular expressions as presented in Chapter 3. This dataset is a collection of sequences of words, each of which includes one instance of verb-noun expression to be categorised as literal or idiomatic. This problem can straightforwardly be framed as a classification task where the input is the collection of features extracted from sequences, i.e. the target expressions along with their contexts, and the output indicates whether the target expression is literal or idiomatic.

For manual evaluation, the difficulty of traversing the whole corpus is an obvious limitation. However, machine learning algorithms facilitate efficient investigation of each and every word in a corpus, and the resulting trained models tag sequences accordingly based on sequence labelling methodologies. Recent studies on token-based identification of MWEs are heading towards using structured sequence tagging models. Conditional Random Fields (CRF) in the work of Constant et al. (2013), and structured perceptron in the work of Schneider et al. (2014a) are two outstanding examples.

While most of the recent work on token-based identification of MWEs apply sequence tagging approaches with the so-called IOB labelling, Legrand and Collobert (2016) frame the problem as classification. They propose a neural network based model which is able to classify representations as MWE or not by learning fixed-size representations for arbitrary sized chunks. They have shown better performance in MWE identification than the CRF based approach in Constant et al. (2013).

The choice of the model based on the data is an important issue. Our data includes occurrences of specific verb-noun expressions with the context around them. This makes it possible to have sizeable datasets annotated for a specific type of MWE, enabling a more extensive evaluation. We design an experiment to see whether our task can benefit from sequence tagging compared to sequence classification. Specifically, we compare the results of a CRF tagger with a simple Naïve Bayes Classifier (NBC) in predicting the idiomaticity of the expressions. The idea behind our feature representation is similar to the model described in Chapter 3, with the difference that for CRF and NBC, we consider simple word forms (rather than the vectors) of the verb, the noun, and the two words after, as lexical context features. The experiments and results are further reported and discussed in Section 4.4.1.

Having observed and discussed the benefits of modelling the task as classification and in accordance with our proposed approach in Chapter 3, we continue to further develop and train models on our data using classification.

4.2 Evaluating MWE Identification

In the vast majority of previous work on supervised modelling of MWE tokens, data is randomly split into train and test sets. In a random splitting, it is possible for occurrences of the same expression type to occur in both train and test sets. Some examples are listed in Table 4.1. There are many instances where the expression almost always behaves idiomatically (e.g. *take part*, *make progress*), or literally (e.g. *eat food*, *give money*). In such cases

a model learns every feature related to the POS and lemma form of the expression, and can perfectly predict the correct tag for the expression in the test set (regardless of the expression being idiomatic or literal). In this way, it appears that the test data has an overlap with the training data.[1]

Table 4.1: An example of random train and test splitting of sentences containing MWEs. Instances and their annotations are selected from VNC-Tokens dataset in which the sentences are from BNC.

Train	Test
Finding *her* ***feet*** *she immersed herself in her role as a Soap Sud with all the ease of a small screen veteran*	*In just a couple of days you'll* ***find*** *your* ***feet*** *and get that special feeling that you belong in your Club*
Democracy is becoming a reality the possessors of new and increasing political power are ***finding*** *their* ***feet*** *not less abroad than in this country*	*A Member simply* ***gives notice*** *and eventually moves that the Bill be read a first time*
In fact they often demand their key worker when they may be still ***finding*** *their* ***feet*** *which is a bit of a pressure on them*	
Parliament should give fresh thought to enacting a provision placing an obligation upon a council tenant to ***give notice*** *to the council before being permitted to commence proceedings*	
First by ***giving notice*** *to the chairman of the appropriate committee*	

[1]Even in stratified cross-validation only the distribution of items of different classes are controlled to be balanced. However, the similarities between the items in train and test are ignored. In other words stratified splitting of the data does not take care of generalisation.

The weakness of a model would be illustrated when it makes incorrect predictions for the few instances where the expression occurs in the non-predominant sense. However, these cases are rare.

We believe that random splitting of data into train and test is the reason behind such disparity in the reported state-of-the art results in the literature: from the F-score of 64% with the DiMSUM dataset (Schneider et al., 2016), to 90% (Al Saied et al., 2017) for a dataset in the last PARSEME shared task (Savary et al., 2017). We find that in order to prevent the performance results from being misleadingly high, the distribution of the tokens between train and test should be controlled. Failure to do so can result in a kind of overfitting which may be overlooked during evaluation.

Having observed this issue, for evaluation we propose and perform type-aware train and test splitting. To this end, we divide expression types into train and test folds and gather all occurrences of each type into the same fold. This makes the predication rigorous, since the model performs cross-type learning. One interesting study that considers cross-type learning of MWEs is by Fothergill and Baldwin (2012). However, they did not deeply discuss evaluation and the general advantages and effects of cross-type classification on evaluation. Rather, they use the approach in order to learn better features from specialised MWE resources.

We propose type-aware splitting of the data as a supplementary benchmark for evaluating MWE identification. We design experiments to show the effectiveness of this kind of evaluation in assessing the generalisability of

models.

4.2.1 Methodology

The emphasis of our research is on token-based identification of MWEs, which we model as a classification, rather than a sequence labelling problem. To determine the idiomaticity of each verb-noun occurrence, we experiment with using solely context features without any sophisticated linguistic information. We do not exploit parsing, tagging or external lexicon-based information.

In order to construct context features, given each occurrence of a verb-noun combination, we follow the approach outlined in Section 3.5 and concatenate four different word vectors corresponding to the verb, noun, and their two following adjacent words while preserving the original order. Concatenated word vectors are fed into different classification models to be evaluated in terms of their performance.

The classification algorithms that have been used are Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Multi Layer Perceptron (MLP), and Support Vector Machine (SVM), as explained in Chapter 2. We have also experimented with neural network-based classification models. The best result is achieved with a combination of bidirectional Long Short-Term Memory network with a convolutional layer as a front-end (ConvNet+LSTM). These are also described in Section 2.4.3.

4.2.2 Evaluation Approaches

In all cases we measure classifier performance using 10-fold cross-validation.

4.2.2.1 Standard Splitting of Data into Train and Test

In the standard method of performing cross-validation, all of the data is randomly divided into k folds and then the model is repeatedly trained on the data of $k-1$ folds and tested on the remaining fold. The result is averaged among different k iterations. In our task, we find the result from this evaluation misleading. The repetition of the same expression in both train and test partitions helps the model predict those specific types of expressions well, whilst the model might not work for new unseen expressions in test. Even stratified cross-validation suffers from the same kind of overfitting. In standard stratified cross-validation, imbalance is addressed by controlling the distribution of labels alone, so that all folds have the same distribution of labels. As is the case with standard cross-validation, this method is not informed about the idiosyncratic distribution of types and tokens.

These methods of evaluation cannot precisely reflect the effectiveness of the model or the features used. They show better results for models that are more prone to overfitting. It is not particularly clear from this kind of evaluation whether a model, even if it performs well according to the performance metrics, could be generalised. The particular difficulty is in the learning of unseen and/or ambiguous expressions that have a balanced distribution of occurrences as literal or idiomatic. We show the performance computed using this type of evaluation for different classifiers in Table 4.4.

4.2.2.2 Type-aware Splitting of Data

We propose a custom cross-validation by splitting the expression occurrences into different folds based on their types/canonical forms. We split the expression types into k groups, and all occurrences of the expressions in the k^{th} group goes into the k^{th} fold. This method ensures that the model performs cross-type learning and generalises to tokens from unseen types in the test fold. In other words, the model is learning the features and general patterns and does not overfit on highly recurrent token occurrences.

In order to control for the distribution of the data in separate folds (i.e. keep the five fold sizes as equal as possible), we rank all expression types based on their frequency, in descending order. Starting from the top of the list, we select a type and place its occurrences in a separate fold until we cover all the folds. We repeat the same procedure until the list is exhausted.

The results for all classifiers evaluated using this approach are reported in Table 4.5.

4.3 Data

We first experiment with two similarly designed datasets in Italian and Spanish, and later on a standard available dataset for English. For these experiments, the type and token distribution of the expressions in the data is of high importance.

The Italian data, as described in Section 3.3, includes a large set of con-

cordances of verb-noun expressions. Each item in the dataset is one concordance of a verb-noun expression and the whole item is annotated with 1 if the verb-noun inside is an MWE and with 0 otherwise. The resulting data, after cleaning and resolving the disagreements, contains 18,540 concordances of 940 expression types.[2] The Spanish data, similar to Italian, is also presented in Section 3.3 and includes 3,090 concordances of 747 expression types.

For English, we employ a standard dataset called VNC-Tokens prepared by Cook et al. (2008)[3], as explained in Section 2.2. The dataset includes sentences from the BNC corpus including occurrences of Verb+Noun expressions. It is suitable for our task as it contains expressions with both skewed and balanced behaviour in being literal or idiomatic. Rather than concordances, it includes sentences from BNC containing occurrences of Verb+Noun expressions. Two native English speakers have selected the expression types based on whether they have the potential for occurring in both idiomatic or literal senses.

Although this dataset is slightly different from our Italian and Spanish data, which are extracted randomly, it features the same favourable pattern of containing different occurrences of same expression types that can be split into train and test. We find it worthwhile to investigate our observations on a differently collected but standard dataset. The Verb+Nouns in this dataset are not necessarily continuous. We ignore the cases where the Verb+Noun

[2]For more details on data collection and pre-processing, refer to Section 3.3.

[3]The dataset is available in `https://sourceforge.net/projects/multiword/files/MWE_resources/20110627/`

occurs in passive form and those that the annotators were unsure of, and this results in 2,499 sentences consisting of Verb+Noun expressions. The statistics of the data for all three languages are reported in Table 4.2.

Table 4.2: Distribution of the data

	Italian	Spanish	English
Expression types	940	747	53
Expression tokens	18,540	3,090	2,499
MWE tokens	10,804 (58.27%)	2,094 (66.57%)	1,981 (79.27%)

For all three datasets, we consider the same context features for classification: we extract the vectors of the verb, noun and the two words after the noun.

4.4 Results

In this section, we present experimental results of using classification versus sequence tagging for identifying MWEs. We then compare several classifiers using different train and test splitting methods. Finally, we analyse the effectiveness of neural network based word embeddings compared with count-based representations using one of the best classifiers.

4.4.1 Sequence Classification versus Sequence Tagging

The experimental data in this study can be processed thoroughly with standard classification approaches, since the goal is to predict the idiomaticity of an expression in a given context. However, Scholivet and Ramisch (2017)

have modelled such data with sequence tagging. We believe that since not all of the words in a sequence are going to be tagged, MWE identification using such data cannot benefit from sequence labelling. We have applied sequence tagging to the data to properly investigate the effects. Specifically, we consider the simple Naïve Bayes Classifier (NBC) as a baseline sequence classification methodology, and Conditional Random Field (CRF) as a sequence tagging approach. Both of the models use simple nominal features: the verb, the noun and the two words after the noun. The results are reported in Table 4.3 in terms of accuracy.

Table 4.3: Performance of sequence classification versus sequence tagging

	regular cross-validation			type-aware cross-validation		
	it	es	en	it	es	en
NBC	0.9504	0.9601	0.8560	0.7291	0.7298	0.6013
CRF	0.9165	0.9142	0.8176	0.6447	0.7199	0.6848

As can be seen in Table 4.3, CRF cannot even beat the simple NBC except in the case of English data (when we apply cross-type learning). This is because our data is naturally suited for sequence classification and cannot benefit from sequence labelling models. We do not want to tag each and every token in a sentence and only one expression in a sentence is the target. A classification approach better focuses on the features of a target expression inside the sentence.

4.4.2 Regular and Type-aware Evaluation

Evaluation performance for all classifiers using two different kinds of train and test splitting, namely regular (random) and our proposed type-aware, are reported in Tables 4.4 and 4.5. The columns of the tables represent the results for Italian (it), Spanish (es) and English (en). All traditional classifiers in this experiment use the same vectorised context features. The word vectors used in this study are available online.[4] The pre-trained Italian and Spanish word embeddings have been derived using Gensim's skipgram word2vec model with the window size of 10 to extract vectors of size 300. For English we use word embeddings of the same dimension trained using Glove (Pennington et al., 2014) algorithm available via `spaCy`. [5]

We also report the results from a more sophisticated neural network based architecture comprising of a bi-LSTM with an additional convolutional layer as a front-end (ConvNet+LSTM). For this architecture, the context window size is 2 (two words before and two words after the verb-noun expression).[6] Implementation details of these experiments can be found at `https://github.com/shivaat/VN-tokens-clf`.

When using regular cross-validation in which tokens are distributed into separate folds regardless of their types (Table 4.4), all classifiers show high performances with little difference. ConvNet+LSTM, in particular, performs

[4] `http://hlt.isti.cnr.it/wordembeddings/` for Italian and `https://github.com/Kyubyong/wordvectors` for Spanish

[5] `https://spacy.io/docs/usage/word-vectors-similarities`

[6] The difference in results was negligible when considering only the two context words on the right.

Table 4.4: Regular evaluation results: accuracy (standard deviation)

Classifiers	it	es	en
Majority Baseline	0.5827	0.6657	0.7927
LR	0.8869 (0.007)	0.9129 (0.011)	0.8651 (0.020)
DT	0.8905 (0.008)	0.9065 (0.017)	0.8799 (0.018)
RF	0.9218 (0.005)	0.9337 (0.019)	0.9024 (0.017)
MLP	0.9069 (0.006)	0.933 (0.009)	0.9056 (0.016)
SVM	0.9116 (0.005)	0.9207 (0.009)	0.7927 (0.021)
ConvNet+LSTM	**0.9220 (0.007)**	**0.9668 (0.01)**	**0.8860 (0.024)**

Table 4.5: Type-aware evaluation results: accuracy (standard deviation)

Classifiers	it	es	en
Majority Baseline	0.5827	0.6657	0.7927
LR	0.6909 (0.06)	0.8178 (0.074)	0.8092 (0.149)
DT	0.6048 (0.03)	0.7483 (0.078)	0.6327 (0.128)
RF	0.6337 (0.08)	0.7604 (0.097)	0.7321 (0.19)
MLP	0.7053 (0.06)	0.8319 (0.086)	0.7294 (0.169)
SVM	**0.7369 (0.07)**	0.8460 (0.093)	0.8062 (0.152)
ConvNet+LSTM	0.6601 (0.053)	**0.8681 (0.072)**	**0.8112 (0.106)**

the best. We believe this is the result of overfitting arising from random train and test splitting. However, we can see notable differences between classifiers in Table 4.5 where we cross-validate in such a way that no same expression type occurs in both train and test partitions.

In the case of cross-type learning (Table 4.5), the SVM classifier has shown the best results in identifying MWEs using vectorised context features for Italian, and a close second for the Spanish and English data where ConvNet+LSTM performs the best. The performance of this classifier is followed by that of MLP and LR for both Italian and Spanish. For English,

the results of SVM and LR are comparable. Computed performance for other classifiers like DT and RF drop sharply when we use our type-aware cross-validation. This is also the case for ConvNet+LSTM for Italian data. This experiment determines how well a classifier can generalise to different expression types. SVM and LR in particular are shown to be fairly suitable for cross-type identification of MWEs. MLP also performs relatively well overall. For the English data it is worth noting that VNC-Tokens is very imbalanced with the majority baseline of 0.7927 which is difficult to beat by classifiers.

Since type-aware cross-validation is a rigorous measure, expecting the model to learn general discriminative features across different MWE types, its resulting evaluation score not only shows the generalisability of the model, but also can be conceived as a measure of lower-bound performance for a system on the blind test data.

4.4.3 Effectiveness of Word Embedding Representation

One important feature of our method is the use of word embeddings to model context features. To specifically demonstrate the effect of neural network-based embeddings on classifiers in the task of identifying verb-noun MWEs, we perform an experiment using sparse bag-of-words count vectors with tf-idf weighting and make comparisons between the performance results from the two representations. Based on the previous experiments, we feed the vectors

to a Multi Layer Perceptron (MLP) which works reasonably well compared to other classifiers. Note that the execution time for the best performing model, SVM, is almost 5 times that of MLP which makes it impractical. The results of this comparison can be seen in Table 4.6.

Table 4.6: The accuracy of MLP in identifying verb-noun MWEs using word2vec and count-based embedding

	Accuracy (std.)		
	it	es	en
MLP w/ count-based embedding	0.6504 (0.0354)	0.7851 (0.042)	0.7002 (0.099)
MLP w/ word2vec	0.7053 (0.06)	0.8319 (0.086)	0.7294 (0.169)

The results in Table 4.6 show the improvement in performance when employing word embeddings rather than the vanilla count-based vectors for all three languages (this is less significant for English).

4.5 Discussion

In order to understand the argument behind type-aware evaluation and its applicability, we have to look at the distribution of data points. For instance, in the Italian data, the majority of data points belong to MWE types whose tokens invariably occur as idiomatic or literal only. In other words, if we plot the distribution of tokens with regard to the degree of idiomaticity of their corresponding types, we would see a skewed distribution (even after ignoring the 15 most frequent expressions), where only a small portion of tokens belong to MWE types whose usages can be fluid between literal and

idiomatic.

In such a scenario, a model easily overfits on the majority of the data, where labels have been assigned invariably. However, this skewedness is not necessarily reflected in the distribution of MWE labels where we might see a relatively balanced distribution of literal and idiomatic labels. In other words, there might be no severe class imbalance in the dataset, rather within-class imbalance (Ali et al., 2015).

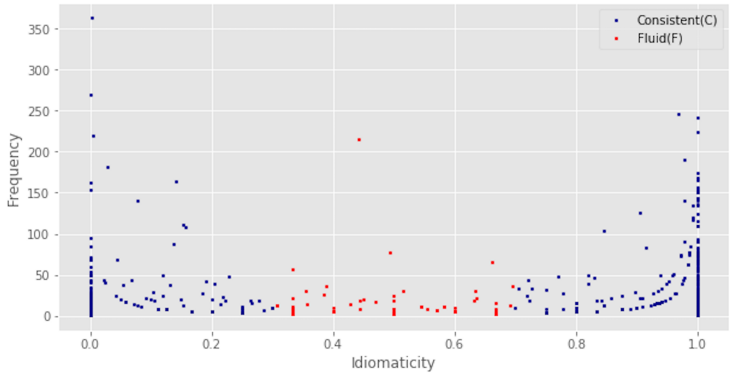


Figure 4.1: Distribution of expression types in the Italian dataset.

To illustrate the point, we assign two categories for MWE types, namely Consistent (C) and Fluid (F). Those types whose tokens occur more than 70% of the time as only literal or idiomatic, are tagged as C, and the rest are considered as F. Accordingly, Figure 4.1 shows the distribution of the expression types with regard to the behaviour of their corresponding tokens. As evidenced in Figure 4.1, the majority of expressions with higher token frequencies are from the sub-class C. For this reason, evaluation using a vanilla cross validation or even stratified cross validation would not provide

us with reliable results, since splitting of train and test disregards the within-class imbalance inherent in the data.

Since this is the case with data in real world, we propose type-aware train and test splitting as a supplementary approach for training models and evaluating the results. This way, we ensure that a model has the optimum ability for generalisation, learns general properties for MWEs, and is not merely based on memorising the words that construct MWEs.

It is worth noting that we have not used any linguistic or lexical features and we expect vector representation of context to be generalisable enough. Even with these generalisable features we observe substantial differences between regular and type-aware cross-validation. In cases where less generalisable features (e.g. word forms, POS tags, etc) are used, the need for a rigorous type-aware train and test splitting is even more pressing.

Regarding the previous data and models for MWEs, DiMSUM is one noteworthy shared tasks. DiMSUM includes a recent tagged corpus for MWEs with a fairly small size of $4,799$ sentences in train and $1,000$ in test, including all types of MWEs. With such limited data, we have observed only a few number of expressions of the form verb-noun occurring in both train and test. To give an example, from a selection of the six most frequent light verbs, their combinations with nouns occur only 13 times in the test data, out of which only 3 are MWEs. There are no repeated occurrences of these cases in both train and test data. Therefore, this data does not inherently lead to misleading results. In other words, a model that works well on this

data could be fairly generalised. We expect that the results of type-aware and random splitting would not be different from each other in this particular case.

Gharbieh et al. (2017) demonstrate better performance using deep neural network models when compared with traditional machine learning on DiMSUM. In our experiment of type-aware classification, SVM performs the best, even outperforming LSTM, ConvNet, and their combinations for Italian. Since neither DiMSUM or our data is big enough for a proper analysis with deep learning, more studies are required to find the most effective model to identify MWEs.

For token-based identification of MWEs in English, another standard dataset used in this study is VNC-Tokens (Cook et al., 2008). One advantage of this corpus is that the data is gathered particularly for the task of disambiguation between idiomatic and literal usages of expressions. Before annotation, they selected only the expressions that have the potential to occur in both idiomatic and literal senses. Although for this study we do not follow the original splitting of the data into train, development, and test sets (i.e. we perform our proposed way of splitting the data into train and test), the splitting of their data is type-aware. Therefore, an evaluation method applied to this data, is able to truly measure generalisation.

In the PARSEME shared task (Savary et al., 2017), which features the most recent multi-lingual data for MWEs, Maldonado et al. (2017) present statistics on the percentage of previously seen data from test sets of all lan-

guages (i.e. proportion of MWE instances in the test set that have also been seen in the training set). The correlation between these percentages and the results stress the need for proper train and test splitting. The experiments with the data for the PARSEME shared task would definitely benefit from such training and evaluation. Maldonado et al. (2017) reported that the datasets of some languages have high proportions of test data which were previously seen in the training data (e.g. for Farsi and Romanian). A model which is evaluated using random splitting shows misleadingly high performance.

4.6 Summary

In this chapter, we discussed the approaches to model MWE identification in context as well as their evaluation. We showed that MWEs in our data are best modelled using classification rather than tagging. We presented a new approach for evaluating the performance of systems for token-based identification of MWEs and investigated the generalisation power of type-aware splitting of data into train and test. The results are reported using several classifiers.

Chapter 5

Tagging Corpora for Multiword Expressions

In the previous chapter, we dealt with MWE identification in our compiled corpora. For our prepared data, the purpose was to disambiguate between different occurrences of the same expression types in addition to handling expressions with a more consistent idiomatic/literal behaviour. The corpus was task-specific as not all tokens of a sequence were tagged but only the target expressions. Not all datasets in the literature are like that.

Recently, new corpora tagged for MWEs have been developed in different languages. One of them is the collection of corpora (in almost 20 languages) collected and annotated for verbal MWEs within the scope of the PARSEME COST Action. This has been used for the 'PARSEME shared task on automatic verbal MWE identification' (Savary et al., 2017) which was held in conjunction with EACL 2017. This collection and the annotations were later improved in order to construct the datasets for the newer edition (edition 1.1) of this shared task to be held in conjunction with COLING 2018.

In this chapter, we present our novel approach to predict labels for MWEs in context and solve the problem as defined in the shared task on automatic verbal MWE identification. The chapter is organised as follows. Section 5.1

describes the shared task and its goals. In Section 5.2, we discuss the related studies on automatic sequence tagging including MWE identification in running texts. Furthermore, Section 5.3 demonstrates our proposed structured prediction approach. Section 5.4 details the datasets and how we use them. In Section 5.5, we explain the experiments including the baselines and our implementations. Finally, Section 5.6 reports the extensive evaluation and results of this study. The chapter is summarised in Section 5.7.

5.1 Task Description

The shared task on automatic identification of verbal multiword expressions (VMWEs) aims at identifying verbal MWEs in running texts (Savary et al., 2017). VMWEs are simply multiword expressions whose syntactic head in the prototypical form is a verb.[1] Due to their complex characteristics including discontinuity, non-compositionality, heterogeneity and syntactic variability, identification of verbal MWEs is challenging for NLP applications. They include different categories, such as idioms, verb-particle constructions (VPCs), light-verb constructions (LVCs), multi-verb constructions (MVCs), inherently reflexive verbs, etc. Brief definitions and examples for these categories are provided in Tables 5.1 and 5.2.

The shared task covers data for 18 languages for the first edition (1.0) and 20 languages for edition 1.1. Native speakers of different languages anno-

[1]This definition is from the shared task annotation guidelines in `http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=010_Definitions_and_scope/020_Verbal_multiword_expressions`.

tated both continuous and discontinuous sequences of lexicalised components of VMWEs. For instance, in *to **put** something **up***, the verb and the particle are lexicalised items that are annotated as integral parts of the VMWE. Each token in a sentence is tagged by annotators as part of a VMWE or not and each identified VMWE is assigned to one of the verbal MWE categories . The categories and hence the tags are more detailed in the second edition compared to the first. Tables 5.1 and 5.2 detail the tags for data of shared task edition 1.0 and 1.1 respectively. More explanation on shared task datasets are provided in Section 5.4.

Table 5.1: The tags used for annotating VMWEs in the shared task on automatic verbal MWE identification - edition 1.0.

VMWE categories	Examples
LVC: light verb constructions, in which the verb has little semantic content and the noun mostly determines the meaning of the expression.	*take a walk*
ID: idioms, which lack compositional meanings.	*kick the bucket*
IReflV: inherently reflexive verbs, in which a reflexive clitic modifies the meaning of the verb.	*find oneself*
VPC: verb-particle combinations, in which the combination of the verb and the particle has non-compositional meaning.	*turn on*
OTH: other verbal MWEs, which do not belong to any of the categories above.	*drink and drive*

The task is to devise an automatic system that can locate occurrences of VMWEs and recognise their categories. The system is evaluated by two statistical measures: one accounts for VMWEs that are identified exactly, i.e. when all components of an MWE token are correctly tagged (per VMWE, strict matching) and the other that concerns VMWEs that are identified

Table 5.2: The tags used for annotating VMWEs in the shared task on Automatic Verbal MWE Identification - edition 1.1.

Categories tags	Examples
LVC.full: LVCs in which the verb has no semantic content.	*make a decision*
LVC.cause: LVCs in which the verb adds a causative meaning to the noun.	*give control*
VIDs: Verbal expressions that have fully idiomatic interpretations.	*go bananas*
IRV: inherently reflexive verbs	*to **help oneself** to food*
VPC.full: fully non-compositional VPCs, in which the particle totally changes the meaning of the verb.	*give up*
VPC.semi: semi non-compositional VPCs, in which the particle adds a partly predictable but non-spatial meaning to the verb.	*eat up*
MVC: multi-verb constructions	*let go*
IAV: Inherently adpositional verbs (sometimes called prepositional verbs)	*come across*

partially (per token, fuzzy matching) i.e. word components are evaluated individually.

5.2 Background

Identifying MWEs in context using the aforementioned corpora is a sequence tagging task and hence can be framed as a structured prediction problem. Other problems that are modelled as structured prediction include automatic chunking (Tjong Kim Sang, 2000), Named Entity Recognition (NER) (Jansche, 2002), Semantic Role Labelling (Carreras et al., 2008). The models used for these problems can borrow some components from each other.

Structured learning is commonly accomplished using graphical models, among which conditional random field (CRF) is a standard technique. As

explained in Section 2.4.4, CRF considers labels for neighbouring words in order to tag each word in a sequence. It has been widely used in NLP tasks such as POS tagging (Lafferty et al., 2001), NER (McCallum and Li, 2003) and shallow parsing (Sha and Pereira, 2003). CRF has also been applied to MWE identification (Qu et al., 2015; Constant and Tellier, 2012; Maldonado et al., 2017).

Turian et al. (2010) augmented CRF with unsupervised word representations and applied the model to chunking and named entity recognition. The focus of the work is to incorporate word representations that are learned from unlabelled data to be used as extra features in supervised NLP systems. This would render the approach semi-supervised. For chunking, they employed CRFsuite (Okazaki, 2007) and modified its feature generation so that it receives new features. As a result, the model is flexible to receive different word embeddings as input features and can be used in other sequence labelling tasks. They evaluated the results by comparing the effect of using different word representations with each other and also with some previous baseline works. Since the model is well-implemented and robust, we use it as the benchmark system to identify MWEs in context and compare the results of our system with those from Turian's CRF.

The impact of word representation features has been investigated and discussed in the literature (Qu et al., 2015; Collobert et al., 2011; Turian et al., 2010). We consider using word2vec skip-gram (Mikolov et al., 2013c), its extended version, word2vecf (Levy and Goldberg, 2014) and its more

recent subword-enriched version (Bojanowski et al., 2016) in our models.

Neural network approaches such as recurrent neural networks (RNN) and convolutional neural networks (CNNs or ConvNets) have gained considerable recent traction, and are very effective in modelling sequences and extracting features from them (see Section 2.4.3). They are widely used in NLP tasks, especially when contextual information is of importance (Ma and Hovy, 2016; Lample et al., 2016; Rei and Yannakoudakis, 2016). These neural network architectures have the advantage of not relying heavily on hand-crafted features and domain-specific knowledge in order to learn effectively from available training corpora.

ConvNets are known as n-gram detectors (Goldberg, 2017) and are able to extract features that account for relations between neighbouring words. In ConvNet multiple filters can be applied to extract different local features across different window sizes. Long Short Term Memory networks (LSTMs) are able to capture long-range dependencies between components of sequences. The dependency features can be captured by LSTMs in both directions (left to right and right to left) and the two resulted context representations are concatenated to construct a bi-directional LSTM architecture. These architectures effectively encapture representations of words in their contexts, which is useful for numerous tagging applications.

Lample et al. (2016) have proposed a model combining LSTM with CRF in order to perform structured sequence tagging for NER. They have shown their results to be better than their implemented transition-based approach.

Ma and Hovy (2016) have also employed CRF on top of their devised neural network architecture. The success of these models in NER prompted us to devise similar architectures for tagging MWEs.

5.3 Our Proposed Approach

The first neural network architecture that we employ for our structured tagging system is a combination of two ConvNet and one LSTM layers. With the view to using ConvNets as n-gram detectors, we apply one convolutional layer with the convolutional window size of 2 (`conv1` in Figure 5.1) and the other with the size of 3 (`conv2` in Figure 5.1). We expect that most MWEs are combinations of 2 or 3 words. The features extracted from these two ConvNet layers are concatenated and given to a bi-directional LSTM to model the sequences. LSTM keeps track of context information and is relatively insensitive to the distance between tokens.

We use pre-trained word embeddings (as detailed in Section 5.3.1) to provide weights for the embedding layer in the network. We also have another optional input layer featuring POS information, which is further explained in Section 5.3.2. The architecture of the system is depicted in Figure 5.1.

Although ConvNets and LSTMs have been shown to be effective in capturing contextual features and therefore in representing sequences or phrases (Collobert et al., 2011), they are blind to the structure of output labels when simply combined with final dense layers. In standard sequence to sequence models, the independent classification decisions are limiting when there are

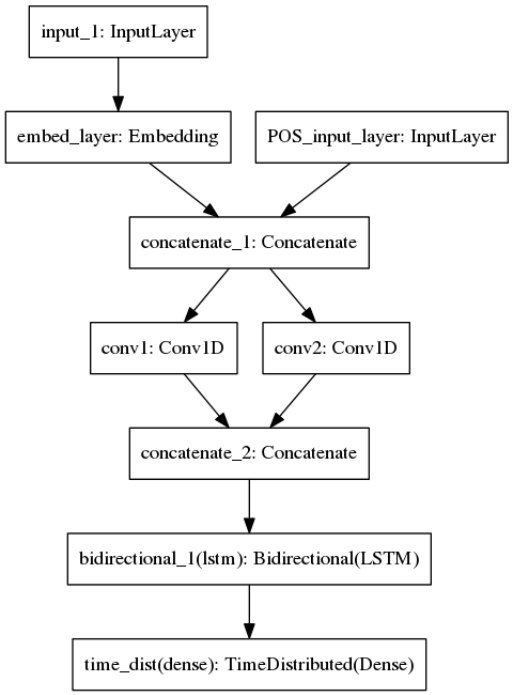


Figure 5.1: The architecture ConvNet+LSTM model.

strong dependencies across output labels. In the case of MWE identification, when we have the sentence *he made the usual mistake.*, a model should detect *made* and *mistake* together as components of an LVC. The model would benefit from being aware that an LVC is composed of two components. During training, it should be penalised for an invalidly tagged sentence with only one word as LVC.

In order for a system to effectively predict labels for sequences, in addition to access to features of input data, a knowledge of the structure of the output labels would enhance its performance. In structured learning, a method that predicts labels for a target token should consider labels of the neighbouring tokens along with their contextual features.

To this end, we propose another approach that combines LSTM and ConvNet with CRF. The idea is to use ConvNet and LSTM as feature generation layers of a neural network and CRF as the final output prediction layer. In this way, we benefit from the individual strengths of all three models in order to effectively tag a corpus for MWEs.

Addition of a CRF layer on top of a neural network has been experimented with, in NLP, by Lample et al. (2016) and Ma and Hovy (2016). We find the inclusion of ConvNets effective for this task based on previously reported results on applying deep learning for MWE identification (Gharbieh et al., 2017). Lample et al. (2016) and Ma and Hovy (2016) also augment their network with a character embedding layer. We sidestep this extra step with the help of pre-trained embeddings that take advantage of sub-word information (Bojanowski et al., 2016) as explained in Section 5.3.1. The architecture of our system is depicted in Figure 5.2.

As inputs for every token in a sequence, we use pre-trained word embeddings which are learned from large corpora as explained in 5.3.1. ConvNets then extract n-gram features using filters of size 2 and 3 on input word representations. Furthermore, the bi-directional LSTM learns the structures of sequences. The feature weights resulting from the bi-LSTM layer are linearly projected onto a layer whose size is equal to the number of distinct tags. Instead of using the Softmax output for the final layer, we use a CRF, as previously described, to take into account neighbouring tags, yielding the final predictions for every token. The parameter settings of different layers

are detailed in Section 5.5.2.

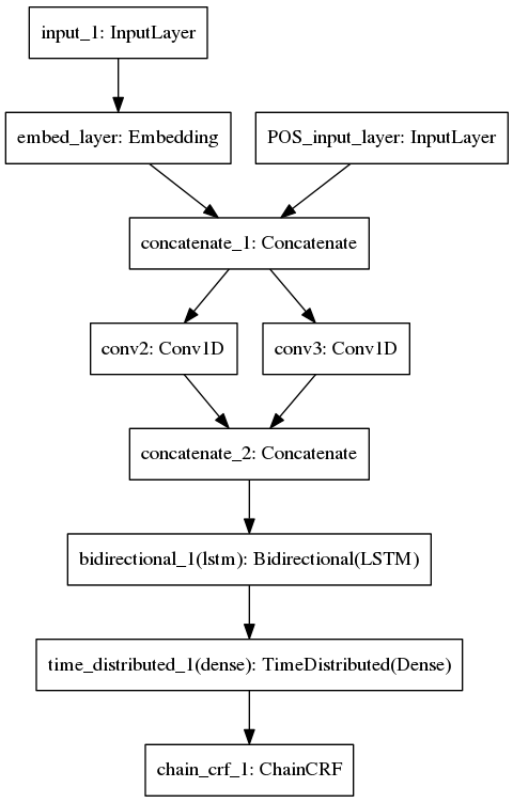


Figure 5.2: The architecture ConvNet+LSTM+CRF model.

5.3.1 Word Embeddings

Neural network based embeddings are the most standard input representations in deep learning methodologies. To learn a neural representation, an embedding layer can be added to the network architecture. This initial embedding layer is capable of iteratively learning the representation of the data for the task at hand. Another option is to use pre-trained embeddings, which is common in NLP. We believe that pre-trained embeddings learned

using unsupervised methods from larger data are more beneficial, especially in cases where the task data is not big enough. Therefore, in this study we use pre-trained embeddings as the input to our neural network architectures and keep them fixed by preventing the network from updating embedding weights.

Conventional unsupervised methods to construct embeddings from large corpora ignore the internal structure of words and assign a single distinct vector to each word in the corpus. When dealing with rare words in morphologically rich languages, a word with an infrequent inflected form might not receive a generalised representation. This is problematic in the case of rare or out-of-vocabulary words. To alleviate this limitation, Bojanowski et al. (2016)'s approach is an attempt at modelling morphology by integrating subword information. It can be considered as an extension of the continuous skip-gram model (Mikolov et al., 2013c) and we refer to it as `Subword WV`.

In `Subword WV`, word vectors are learned by representing words as character n-grams and then summing the vectors to derive a full representation for each word. This method is predicated on the hypothesis that character-level information (including affixes and grammatical rules) contained in the character n-grams help the model develop more generalised semantic representations and thereby represent rare words better. For each language we use a pre-trained `Subword WV` described in Bojanowski et al. (2016) as input to the neural network models. The distributed representations in dimension

300 are obtained with default parameters.[2]

Another embedding approach which we experiment with, in a smaller scale is word2vecf (Levy and Goldberg, 2014). The model as mentioned in Chapter 2 adapts word2vec to train dependency-based context vectors. In this thesis we train and test this model only for Spanish as a preliminary experiment.

We train spaCy's dependency parser (Honnibal and Johnson, 2015) on a sample of the Spanish part of the wikipedia English-Spanish comparable corpus[3] and then we apply word2vecf to extract vector representations. In this case, if we have a piece of dependency parsed text as in Figure 5.3, the idea is to have for each word its dependents as context words. Dependency relations are considered in both directions according to Levy and Goldberg (2014): direct dependency from a token to its head and the inverse relation which is from the head to the dependent token. Inverse relations are specified by `I`. As a result, for the example in Figure 5.3, pairs of word-contexts look like the example provided in Figure 5.4.

Many variations of word-context pairs can be constructed with regard to word forms, lemma and POS tags. In this study we extract vectors for word forms and consider the relations with regard to word forms, lemma forms and POS tags. The word2vecf method is then applied to these word-context pairs to extract dependency-based embeddings.

[2]The embeddings are available at `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

[3]The corpus is available at `http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/`.

ID	FORM	LEMMA	UPOS	DEPREL	DEPS
1	Auster	Auster	PROPN	2	nsubj
2	retreated	retreat	VERB	0	root
3	to	to	ADP	5	case
4	the	the	DET	5	det
5	kitchen	kitchen	NOUN	2	obl
6	to	to	PART	7	mark
7	prepare	prepare	VERB	2	advcl
8	the	the	DET	9	det
9	food	food	NOUN	7	obj
10	.	.	PUNCT	2	punct

Figure 5.3: An example of a dependency parsed text.

WORD	CONTEXT
Auster	retreated_nsubj
retreated	Auster_nsubjI
to	kitchen_case
kitchen	to_caseI
the	kitchen_det
kitchen	the_detI
kitchen	retreated_obl
retreated	kitchen_oblI

Figure 5.4: Sample of word-context pairs.

5.3.2 Features

We extract extra features from the data to be added as additional inputs to the neural network models. These include seven word shape features which can be informative for the identification of MWEs. These are binary features for each token that capture whether it starts with a capital letter, consists of all capital letters, the first character is a # or @, corresponds to a URL, contains a number or is a digit. These seven binary features are added at the end of embedding vectors for each word.

The PARSEME shared task data for most of languages are POS tagged. In order to enrich neural network models with POS information, one way is to construct one-hot representations for POS tags. For categorical variables, where no ordinal relationship exists, one-hot encoding is the general scheme to convert them into a coded form. A one-hot vector for a feature is a vector with the size of the number of all possible feature values. All values of the vector are zero except for the value corresponding to the target feature. We consider all POS tags for each language and construct on-hot vectors for them. The one-hot representation for POS tokens are given as an additional input (other than word2vec representation) to the neural network models.

5.4 Data

We focus on the data from both editions of the shared task on automatic identification of verbal multiword expressions. In the first edition, we experiment with the data of three languages, namely, Spanish, Italian, and French. For the second edition we report the results for six languages in this thesis. Datasets of most languages are provided with morphosyntactic data (parts of speech, lemmas, morphological features and/or syntactic dependencies).

In the first edition, the data in all languages are provided in CONLL-U format with a separate .tsv file (also adapted from the CoNLL format) for the tags. The annotated .tsv file has one token per line and an empty line indicating the end of a sentence. Lines with extra (optional) information about the sentences start with # before the beginning of each sentence. Each

token is represented by 4 tab-separated columns indicating (i) the position of the token in the sentence; (ii) the token surface form; (iii) an optional nsp (no space between the current and the next token) flag; and (iv) a VMWE code, which is the result of annotation for when the token is part of a VMWE.

In the second edition, both data and tags are combined in a .cupt format which is a small modification of CONLL-U and includes the following columns: Word ID, the position of the token in the sentence, the token surface form, LEMMA (column 3), UPOS (column 4), FEATS (column 6), HEAD and dependency relation, DEPREL (columns 7 and 8), MISC (column 10) and PARSEME:MWE (column 11) which is the manual annotation of expressions with the categories: IAV, IRV, LVC.full, LVC.cause, MVC, VID, VPC.full and VPC.semi.

The PARSEME:MWE tag is composed of the VMWE's consecutive number in the sentence and its category. Only the initial token in a VMWE is marked for the category, for example, the tag 2:VID shows that the token signals an idiom which is the second VMWE in the current sentence. In case of nested, coordinated or overlapping VMWEs, multiple codes are separated with a semicolon. Figure 5.5 depicts the file format for two sample of annotated sentences in English. Figure 5.6 presents a Spanish sentence which is annotated with overlapping MWEs.

As can be seen in Figures 5.5 and 5.6, there are minor differences between the two editions in the formatting of the annotations. We have different scripts for reading the data. In both cases we convert the annotation codes

```
# source_sent_id = . . 4045
# text = Worse yet, what is going on will not let us alone.
1   Worse   bad     ADJ     CMP         _   10  advmod          _   _   *
2   yet     yet     ADV     _           _   1   advmod          _   nsp *
3   ,       ,       PUNCT   Comma       _   1   punct           _   _   *
4   what    what    PRON    WH          _   6   nsubj           _   _   *
5   is      be      AUX     PRES-AUX    _   6   aux             _   _   *
6   going   go      VERB    ING         _   10  csubj           _   _   2:VPC.full
7   on      on      ADV     _           _   6   compound:prt     _   _   2
8   will    will    AUX     PRES-AUX    _   10  aux             _   _   *
9   not     not     PART    NEG         _   10  advmod          _   _   *
10  let     let     VERB    INF         _   0   root            _   _   1:VID
11  us      we      PRON    PERS-P1PL-ACC _ 10  obj             _   _   *
12  alone   alone   ADJ     POS         _   10  xcomp           _   nsp 1
13  .       .       PUNCT   Period      _   10  punct           _   _   *
```

Figure 5.5: Annotation of one sample sentence containing one VPC and a verbal idiom in the English data for the shared task edition 1.1.

into a labelling format similar to IOB.[4] To this end, the initial token of an MWE receives the tag `B-` plus its category. For example, for the token *going* in position 6 in the sentence of figure 5.5, the label is `B-VPC.full`. Other components of the expression receive the tag `I-` plus their category. In the case of Figure 5.5, *on* in position 7 gets `I-VPC.full`. Other tokens which are not part of a VMWE (the ones which are marked with `_` in edition 1.0 and `*` in edition 1.1), receive the tag `O`.

In our datasets, VMWEs rarely overlap. In the case of overlap we follow the same annotation scheme as in the shared task by separating multiple tags for a token with a semicolon. The token in position 2 in Figure 5.6 receives the tag `B-IReflV;B-ID`.[5]

[4]The standard IOB labelling is proposed for chunking where there is no gap between the components of one chunk. MWE identification task is different since an MWE may or may not be continuous.

[5]Our systems consider this combined tag as a separate category in the list of all possible tags for a data. This introduces some limitations for the learning system which does not recognise that a candidate MWE can adopt either of these labels individually. On the other hand, this representation can help the system learn what combinations of labels are acceptable.

```
# sent_id es-s5443
# orig_file_sentence 055#43
1     También          _     _
2     se               _     2:IReflV;3:ID
3     mostró           _     2;3
4     partidario       _     3
5     de               _     _
6     aplicar          _     _
7     el               _     _
8     denominado       _     _
9     “                _     _
10    modelo           _     .
11    regata           _     _
12    ”                nsp   _
13    ,                _     _
14    que              _     _
15    da               _     1:LVC
16    prioridad        _     1
17    de               _     _
18    entrada          _     _
19    a                _     _
20    aquellos         _     _
21    países           _     _
22    que              _     _
23    “                _     _
24    mejor            _     .
25    están            _     _
26    preparados       _     _
27    ”                _     _
28    por              _     _
29    encima           _     _
30    de               _     _
31    grupos           _     _
32    preseleccionados nsp   _
33    .                _     _
```

Figure 5.6: Annotation of one sample sentence containing two overlapping VMWE (an idiom for which the verb is reflexive) in the Spanish data for the shared task edition 1.0.

In this IOB-like labelling scheme, we differentiate between the beginning component of an expression and its other components. This distinction is not a requirement for the shared task evaluation (i.e. if any component of an expression is identified with the correct category it is considered as true positive). However, in the results we show that it is beneficial to have these different tags and perform some filtering based on them. Specifically, when we convert the labels from prediction results back to the shared task format,

we filter those cases that have label `I-` without a preceding `B-` and re-tag them as non-MWEs (by marking them with `*`).

Tables 5.3 and 5.4 summarise the sizes of the training/development/test sets for languages of our focus, which are: Spanish (ES), Italian (IT), French (FR) for both editions plus English (EN), German (DE) and Persian (FA) for edition 1.1.

Table 5.3: Sizes of the training/development corpora for the shared task data edition 1.0.

Language	Sents	Tokens	VMWE
ES-train	2502	102090	748
ES-test	2132	57717	500
IT-train	15728	387325	1954
IT-test	1272	40523	500
FR-train	17880	450221	4462
FR-test	1667	35784	500

5.5 Experiments

In order to tag the datasets for VMWEs we use sequence labelling methodologies. We employ both standard CRF and Turian et al. (2010)'s CRF as baselines and we compare them with the results obtained from the implementation of our proposed neural network based approach.

5.5.1 Baseline

Based on Turian et al. (2010)'s CRF approach we first use features that are generally used in CRF implementation as follows:

Table 5.4: Sizes of the training/development/test corpora for the shared task data edition 1.1.

Language	Sents	Tokens	VMWE
ES-train	2771	96521	1739
ES-dev	698	26220	500
ES-test	2046	59623	500
IT-train	13555	360883	3254
IT-dev	917	32613	500
IT-test	1256	37293	503
EN-train	3471	53201	331
EN-test	3965	71002	501
FR-train	17225	432389	4550
FR-dev	2236	56254	629
FR-test	1606	39489	498
DE-train	6734	130588	2820
DE-dev	1184	22146	503
DE-test	1078	20559	500
FA-train	2784	45153	2451
FA-dev	474	8923	501
FA-test	359	7492	501

- word and lemma form of the current token (in position i) and the tokens in the window of size 2 on the left and the right sides of the current word (w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, l_{i-2}, l_{i-1}, l_i, l_{i+1} and l_{i+2})

- POS of the current token and tokens in the window of size 2 on the left and the right sides of the target word (p_{i-2}, p_{i-1}, p_i, p_{i+1} and p_{i+2})

- bigrams of word and lemma forms including tokens in a window of size 1 around the current token ($w_{i-1}w_i$, w_iw_{i+1}, $l_{i-1}l_i$, and l_il_{i+1})

- POS bigrams and trigrams in the window of size 2 around the cur-

rent word ($p_{i-2}p_{i-1}$, $p_{i-1}p_i$, p_ip_{i+1}, $p_{i+1}p_{i+2}$, $p_{i-2}p_{i-1}p_i$, $p_{i-1}p_ip_{i+1}$, and $p_ip_{i+1}p_{i+2}$)

We use Turian's implementation which employs CRFSuite with the above features. We also augment it with pre-trained word embedding features to have embeddings for individual words in the widow of size 2 around the current word ($embed_{i-2}$, $embed_{i-1}$, $embed_i$, $embed_{i+1}$ and $embed_{i+2}$). The hyperparameter l2-regularisation sigma is set to 2 which is reported to be the optimal value for chunking based on Turian et al. (2010). We report the results of the CRF system with and without word representation features for the task of VMWE identification.

5.5.2 Neural Network Parameter Settings

The details of the layers which are depicted in Figures 5.1 and 5.2 are presented in this subsection. These parameters include the number of neurons in each hidden layer, the number of iterations before training is stopped, activation function, and more specifically, the filter size for ConvNet, and the dropout rate for LSTM.

In the first layer every token is represented by its vector from pre-trained embeddings concatenated with 7 word shape features. These are then fed to ConvNet layers. We use one ConvNet layer with 200 neurons and the filter size 2 and the second ConvNet layer with 200 neurons and the filter size 3. These two layers are then concatenated. Since most of the VMWEs are bigram or trigram combinations we find filter sizes of 2 and 3 to be the best

choice for extracting n-gram features. However, we also try filter size 5 for training and see no improvement on the validation set.[6]

We apply no dropout for the ConvNet layers and use rectified linear unit (`ReLU`) as the activation function which transfers the output z of the layer using the function $max(0, z)$. The output of the convolutional layers is given to a bi-directional LSTM layer with 300 neurons, dropout of 0.5 and recurrent dropout of 0.2. We use batch-size 100 for training the networks.

An embedding layer in a neural network architecture can be trainable in which case the weights become updated during the training. However, in all cases, we get better performance when we set the embedding layer not to be trainable. The pre-trained embedding weights which are derived from larger corpora have shown to be more effective in our task.

5.6 Evaluation

In this section, we present the evaluation of our systems for the datasets for VMWE shared tasks edition 1.0 (phase 1) and edition 1.1 (phase 2). In phase 1, predicting VMWE categories is not required and evaluation measures do not take them into account. In phase 2, however, evaluation results are reported for both identifying VMWEs in general and per category. The inputs to our neural network models in phase 1 are word embeddings. However, in phase 2 we feed one-hot representations of POS tags as additional inputs to

[6]We perform our experiment with ConvNet of filter size 5 for the Spanish data of the first edition. The decrease we see in the result is sufficient for us not to try filter size 5 for other datasets.

the systems.

5.6.1 Evaluation Measures

The performance of systems is measured by the standard metrics of precision (P), recall (R) and F1-score (F1). Measures are computed in two settings: one is strict matching (per-MWE) in which all components of an MWE are considered as a unit that should be correctly classified; the other is fuzzy matching (per-token) in which any correctly predicted token of the data is counted. Per-VMWE scoring considers every tagged or predicted VMWE as an indivisible instance, and calculates the ratio of the VMWEs that were correctly predicted (precision) and correctly retrieved (recall). Per-token scoring considers all possible bijections between the VMWEs in the gold and system sets, and takes a matching that maximises the number of correctly predicted tokens (Savary et al., 2017).

In a formal definition, as described in Savary et al. (2017), let $G = g_1, g_2, ..., g_{|G|}$, and $P = p_1, p_2, ..., p_{|P|}$ be the ordered sets of gold and predicted VMWEs in a given sentence, S, respectively. Let B be the set of all bijections and $N = max(|G|, |S|)$, where $g_i = \phi$ for $i > |G|$, and $s_i = \phi$ for $i > |S|$. We define TP_{max} the maximum number of true positives for any possible bijections in a sentence. For the example in Figure 5.7, TP_{max} counts the labels for bijections in lines 10, 11 and 12, as correctly classified labels (hence, the precision of 3/3 and the recall of 3/5).

Considering TP^j_{max}, G^j, S^j and N^j be the values of TP_{max}, G, S and N

Gold		Prediction	
9 Telecomunicaciones	*	9 Telecomunicaciones	*
10 se	1:IReflV;2:ID	10 se	1:IReflV
11 mostró	1;2	11 mostró	1
12 partidario	2	12 partidario	2:ID
13 de	*	13 de	*
14 completar	*	14 completar	*

Figure 5.7: A sample of gold and prediction example

for the j-th sentence, for a corpus of M sentences, precision and recall are defined as in Equation 5.1.

$$P = \frac{\sum_{j=1}^{M} TP_{max}^{j}}{\sum_{j=1}^{M} ||S^j||} \qquad R = \frac{\sum_{j=1}^{M} TP_{max}^{j}}{\sum_{j=1}^{M} ||G^j||} \tag{5.1}$$

The final $F1$-$measure$ is computed as $2PR/(P+R)$.

5.6.2 Phase 1

The aim of this evaluation phase is to compare the results of our model with CRF and Turian's approach as baselines. We present the results for different settings of both Turian's and our proposed neural network architecture. The experiments in this phase are performed for Spanish (ES), Italian (IT) and French (FR) datasets. We also compare the final test results with the best results obtained in the shared task. The best system in the first edition of the shared task was a transition based system (Al Saied et al., 2017) in which sequence labelling is done using a greedy transition-based parser. Their system can be considered as a state-of-the-art solution for identification of verbal MWEs. We compare the performance of our system with that (referred to as TRANSITION) as well.

In both Turian's and our models, for Spanish, we experiment with two

different embeddings including the pre-trained `Subword WV`, referred to in Section 5.3.1 (SWV in tables) and the dependency-based embedding that we train with the vector size of 300 (WVF). For other languages we only experiment with SWV.[7] We set the training epochs to 100 for neural networks.

Table 5.5: Test results for the data of shared task edition 1.0.

	Model	Per-token			Per-MWE		
		P	R	F1	P	R	F1
ES	CRF	76.17	33.36	46.40	66.80	33.80	44.89
	Turian (SWV)	72.62	35.84	48.00	62.37	34.80	44.67
	Turian (WVF)	68.02	40.03	50.40	59.47	40.20	47.97
	ConvNet+LSTM (SWV)	66.59	46.71	54.90	55.56	45.00	49.72
	ConvNet+LSTM (WVF)	64.51	51.15	**57.06**	55.21	49.80	**52.37**
	ConvNet+LSTM+CRF (WVF)	74.81	41.40	53.30	67.31	41.60	51.42
	TRANSITION	65.74	52.52	**58.32**	61.22	54.00	**57.38**
IT	CRF	69.70	10.84	18.76	66.67	10.40	17.99
	Turian (SWV)	69.70	14.45	23.94	64.15	13.60	22.44
	ConvNet+LSTM (SWV)	66.10	18.22	**28.57**	48.78	16.00	**24.10**
	ConvNet+LSTM+CRF (SWV)	59.34	16.97	26.39	49.66	14.40	22.33
	TRANSITION	61.34	33.78	**43.57**	53.54	31.80	**39.90**
FR	CRF	85.53	42.15	56.47	61.30	35.80	45.20
	Turian (SWV)	83.71	46.84	60.07	62.77	40.80	49.45
	ConvNet+LSTM (SWV)	80.34	51.26	**62.59**	69.48	47.80	**56.64**
	ConvNet+LSTM+CRF (SWV)	77.34	44.68	56.64	66.23	40.80	50.50
	TRANSITION	80.88	49.64	61.52	61.47	43.40	50.88

The performance results of our implemented systems for the shared task data of edition 1.0 are presented in Table 5.5. The first three rows for each language in the table demonstrate that enriching CRF with word embedding features improves the performance of MWE identification in accordance with

[7]Experiments with the dependency-based embeddings (WVF) is just a preliminary study that we added to this thesis only for Spanish. We study and experiment with it further in future.

the results for named entity recognition and other NLP tasks (Turian et al., 2010). Moreover, we can see the improved results from ConvNet+LSTM compared to the baseline in all three languages.

Table 5.6: Comparing the learning performance (in terms of F1) for different number of epochs (50 and 100).

		Per-token		Per-MWE	
		50	100	50	100
ES	ConvNet+LSTM (WVF)	56.20	57.06	49.95	52.37
	ConvNet+LSTM+CRF (WVF)	55.06	53.30	**52.49**	51.42
IT	ConvNet+LSTM (WV)	30.82	28.57	26.62	24.10
	ConvNet+LSTM+CRF (WV)	27.86	26.39	25.59	22.33
FR	ConvNet+LSTM (WV)	62.82	62.59	52.64	56.64
	ConvNet+LSTM+CRF (WV)	62.03	56.64	**55.23**	50.50

As for our proposed ConvNet+LSTM+CRF model, we do not see improvement in general across all languages in Table 5.5 over the initial ConvNet+LSTM model. We rerun the experiment with different training epochs to examine the impact of the number of iterations. According to the results of this experiment in Table 5.6, by using 50 epochs for training, in two out of the three languages (i.e. Spanish and French) ConvNet+LSTM+CRF has better performance over ConvNet+LSTM in terms of per-MWE F1 and hence in learning MWE structure. This shows that the added CRF layer can help the model learn MWEs faster in less number of epochs. In the case of the Italian data, higher number of epochs decrease the results in both systems.

For Spanish, we also compare the results of the systems using two differ-

ent embeddings, namely, WVF and `Subword WV` (SWV). According to Table 5.5, using WVF leads to a superior performance compared to when using SWV. This is noteworthy, as `Subword WV` is trained on the whole wikipedia. The strength of WVF over SWV can be seen in both Turian's and ConvNet+LSTM approaches. This shows that, for an MWE identification model, the information WVF obtains from a smaller dependency parsed corpus is more informative than a general SWV representation.

In Table 5.5, we further compare the results of our system to that of the transition based system (Al Saied et al., 2017) which is the winner of the VMWE shared task edition 1.0. Considering that system as the state-of-the-art, our model beats it in the case of the French data. According to the table, the performance of our system is also very close to TRANSITION in the case of Spanish, based on per-token F1 measure. In contrast to TRANSITION, we do not use the dependency parsing tags that are available for the dataset of most languages in the shared task. This information is not always available in real-word applications when dealing with running text, and a model that operates independently of this information has its particular advantages. We further demonstrate the effectiveness of our model in Section 5.6.3.

We improve these systems by adding POS representations to the models and apply the models to the data of the second shared task in the following section.

5.6.3 Phase 2

In this section, we focus on the second edition of VMWE identification shared task. We build on our previous experiments and augment the systems with one-hot representations of POS tags as additional inputs. We compare the results obtained from the two proposed neural network systems on the validation data for different languages. We choose the better system based on validation, to be evaluated on the blind test data. The languages for which we perform the training and parameter optimisation using the development data are English, Spanish, Italian, French, German and Farsi. For all languages both train and validation data sets are provided except for English for which only one set of training data is available. For English we perform 5-fold cross validation to obtain the validation results. The results of the validation phase are presented in Table 5.7.

The significant difference between results of the two systems is mostly in the case of per-token evaluation, based on which ConvNet+LSTM usually performs better. In three out of the six languages, ConvNet+LSTM+CRF shows better results in terms of per-MWE F1 score. In German, we also see a slight improvement in terms of per-token F1 score. This makes comparison of the two systems and choosing the best one more difficult, since even in cases where per-MWE is slightly higher for ConvNet+LSTM+CRF, this small improvement is accompanied by the same or higher amount of drop in per-token F1. In these cases there is a trade-off between per-token and per-MWE

Table 5.7: Development results for the data of shared task edition 1.1.

	Model	Per-token			Per-MWE		
		P	R	F1	P	R	F1
ES	ConvNet+LSTM	71.61	58.44	**64.36**	54.81	49.00	51.74
	ConvNet+LSTM+CRF	71.19	52.87	60.68	58.37	47.40	**52.32**
EN	ConvNet+LSTM	55.39	31.92	**40.50**	35.34	26.59	**30.34**
	ConvNet+LSTM+CRF	52.03	27.12	35.65	35.59	23.87	28.57
IT	ConvNet+LSTM	64.76	43.16	**51.80**	43.72	33.60	38.00
	ConvNet+LSTM+CRF	59.40	43.00	49.88	49.29	34.81	**40.80**
FR	ConvNet+LSTM	85.81	68.82	**76.38**	77.32	66.14	**71.29**
	ConvNet+LSTM+CRF	74.69	67.13	70.71	67.24	62	64.52
DE	ConvNet+LSTM	70.75	45.32	55.25	47.67	38.72	42.73
	ConvNet+LSTM+CRF	60.27	54.08	**57.01**	41.99	45.51	**43.68**
FA	ConvNet+LSTM	91.80	78.00	**84.34**	81.74	73.25	**77.26**
	ConvNet+LSTM+CRF	91.86	76.54	83.50	81.33	73.05	76.97

evaluation measures.

Based on the overall results (average of per-token and per-MWE scores) obtained in the validation phase, we choose to run the first model, ConvNet+LSTM, for all but one language (DE) for the blind test data.

In Table 5.8, for different languages we report the results of our chosen systems compared to two of the best participating systems in the shared task.

In the shared task, the participating systems are categorised into two tracks: closed and open. The closed track includes the systems which use only the provided training/development data, their VMWE annotations and morpho-syntactic tags (e.g. POS and dependency parsing information). The open track includes the systems that use additional resources such as MWE lexicons, symbolic grammars, WordNets, raw corpora, word embeddings, lan-

Table 5.8: Test results for the data of shared task edition 1.1.

	Model	track	Per-token			Per-MWE		
			P	R	F1	P	R	F1
ES	ConvNet+LSTM	open	38.33	53.57	**44.69**	31.65	48.8	**38.39**
	TRAPACC_S	closed	35.96	44.43	39.75	29.54	40	33.98
	CRF-Seq-nocategs	closed	37.47	41.74	39.49	30.87	36	33.24
EN	ConvNet+LSTM	open	60.36	18.77	28.63	48.40	18.16	26.42
	Milos	open	37.32	31.83	34.36	33.81	32.73	**33.27**
	TRAPACC	closed	42.23	28.98	**34.37**	38.4	28.74	32.88
IT	ConvNet+LSTM	open	67.55	49.30	**57**	49.09	43.55	46.15
	TRAVERSAL	closed	74.42	42.11	53.78	63.09	40.32	**49.2**
	TRAPACC	closed	61.54	30.34	40.64	52.43	30.44	38.52
FR	ConvNet+LSTM	open	82.94	57.73	**68.08**	72.39	54.22	**62**
	TRAVERSAL	closed	84.72	48.76	61.9	77.19	44.18	56.19
	Deep-BGT	open	78.88	56.45	65.8	57.81	49.8	53.51
DE	ConvNet+LSTM+CRF	open	69.7	40.82	**51.49**	54.15	37.95	44.63
	Deep-BGT	open	77.92	37.64	50.76	60.94	36.35	**45.53**
	TRAPACC_S	closed	61.13	42.26	49.97	53.26	39.36	45.27
	ConvNet+LSTM	open	*76.22*	*41.74*	***53.94***	*62.38*	*39.96*	***48.71***
FA	ConvNet+LSTM	open	93.87	74.3	**82.95**	86.12	71.86	**78.35**
	GBD-NER-resplit	closed	84.13	78.62	81.28	78.23	77.45	77.83
	GBD-NER-standard	closed	84.54	75.65	79.85	78.11	74.05	76.02

guage models trained on external data, etc. The information about the systems belonging to closed or open track is provided in Table 5.8. We compare the systems of both tracks against each other in the same table.

According to Table 5.8, ConvNet+LSTM(+CRF) have the highest results in 5 out of 6 languages in terms of per-token F1 and 3 out of 6 languages based on per-MWE F1 score. The performance of our systems in Spanish and French is significantly, and in Farsi slightly, better than the other systems. For Italian, ConvNet+LSTM ranked the best in terms of token-based

F1 and the second, based on per-MWE F1 score. Surprisingly, for English the performance of ConvNet+LSTM is considerably worse than the best participating systems. This might be due to the use of more informative features by other systems, since English is a resource-rich language.[8]

The results that we obtain for the German data is contrary to our expectation. ConvNet+LSTM+CRF is applied to this language not only due to better performance on the validation data but also based on the assumption that CRF helps the model achieve better results in terms of per-MWE measure. However, to our surprise, the system does not rank well based on per-MWE and works best in terms of per-token F1. After the blind test evaluation phase we try ConvNet+LSTM also for German to see whether it affects performance. As can be seen in Table 5.8, the corresponding row for ConvNet+LSTM for DE clearly shows a better result for this model.

Although our system is considered 'open track' based on the shared task definition, we have not used any knowledge-based resources. The only data that we employ in addition to the shared task data is unsupervised generic pre-trained word embeddings without any adaptation to accommodate for MWEs. In most other shared tasks (e.g. CONLL, SemEval, etc), using such generic unlabelled resources are allowed in the closed track. It is worth noting that we do not use the dependency parsing tags that might not be available for some languages. The results of our system seems to be partic-

[8]At the time of submission of this thesis, the results of the shared task are anonymously submitted. We are not aware of the details of these systems.

ularly promising for resource-poor languages such as Farsi.

In the following sections we evaluate the performance of our system more extensively in various detailed settings.

Performance per categories of VMWEs

In Table 5.9, we present the results of our system for different categories of VMWEs in several languages.[9] As can be seen in the table, there is almost always a correlation between the amount of gold standard MWEs of a category in training data and the performance of the model for that category in test data. Higher number of MWEs in a category obviously helps the model to learn that category better. This is only violated in the case of English, where `LVC.full`s are the most overrepresented, however our system does not perform very well in identifying them (with token-based F1 score of only 11.23).

The model performs well for the category of reflexive verbs in almost all languages. This can be explained by the fact that the expressions of this category usually have one fixed token (i.e. *se* in Spanish) plus some verbs that can be reflexive. Other than reflexive verbs, there is no category for which our method works well across all languages. This can be explained by the fact that we do not use any linguistic features or expert knowledge, which makes the model not only language-independent, but also category-independent to some extent.

[9]For German, we report the results of the system, LSTM+CRF, which is proved to be the best after evaluation on the test data.

Table 5.9: Performance of our system per categories of VMWEs in terms of token-based and MWE-based F1 scores.

		gold proportion	predicted proportion	token-based	MWE-based
ES	IAV	13%	16%	40	36.96
	IRV	24%	30%	46.79	45.63
	LVC.cause	6%	1%	7.69	5.88
	LVC.full	17%	15%	30.29	22.11
	MVC	21%	33%	41.84	34.64
	VID	19%	6%	32.82	30.77
EN	IAV	9%	3%	12.84	12
	LVC.cause	7%	0%	0	0
	LVC.full	33%	14%	11.23	6.25
	MVC	1%	0%	0	0
	VID	16%	9%	11.76	14.74
	VPC.full	29%	69%	46.67	45.09
	VPC.semi	5%	6%	11.94	5.41
IT	IAV	8%	8%	60.87	41.56
	IRV	19%	22%	59.36	56.54
	LS.ICV	2%	1%	34.78	16.67
	LVC.cause	5%	5%	64.58	58.33
	LVC.full	21%	20%	56.28	45.60
	MVC	1%	0%	26.67	28.57
	VID	41%	41%	43.59	34.03
	VPC.full	5%	3%	47.37	42.11
FR	IRV	22%	27%	71.39	69.57
	LVC.cause	3%	2%	9.3	9.09
	LVC.full	32%	30%	62.82	58.09
	MVC	1%	0%	40	40
	VID	43%	41%	67.84	62.47
DE	IRV	8%	11%	46.63	29.73
	LVC.cause	0%	0%	0	0
	LVC.full	8%	4%	13.01	7.27
	VID	37%	35%	46.73	38.10
	VPC.full	42%	47%	65.86	62.60
	VPC.semi	5%	3%	24.39	18.18
FA	LVC.full	100%	100%	82.89	78.43

For further discussion, more analysis is required on the effect of other phenomena in the distribution of categories in train and test, including the amount of seen/unseen expressions or continuous/discontinuous expressions in each category.

Effects of specific phenomena in VMWEs identification

Following the shared task evaluation, we also assess our system focusing on several specific phenomena in VMWEs including continuity, single or multi-token, and seen or unseen VMWEs. Continuity regards whether the lexicalised components of a VMWE are adjacent (e.g. *make sense*) or non-adjacent (e.g. ***make*** *perfect* ***sense***). Performance of the system is reported for continuous and non-continuous, and single and multi-token VMWEs separately. The other phenomena regards the novelty of expressions in the test data. For this the performance for seen (the ones that occur in both train and test data) and unseen expressions (which only occur in the test data) are reported individually. This reflects the generalisability of the model, that is, its ability to predict correct labels for new unseen data.[10]

The results in Tables 5.10 and 5.11 show the MWE-based F1 scores for different groups when considering these phenomena.

According to Table 5.10, our system consistently performs better in detecting continuous VMWEs compared to discontinuous ones. Whilst the system is generally capable of detecting discontinuous VMWEs to an accept-

[10]Lexicalised components of VMWEs are lemmatised to check whether they are seen or not.

Table 5.10: Proportion of VMWEs in each group: Continuous (C), Discontinuous (D), Multi-token (M) and Single Token (S) and F1 scores of our system for each group individually.

	C-D				M-S			
	gold	prediction	C	D	gold	prediction	M	S
ES	72-28%	93-7%	41.90	19.29	100-0%	91-9%	40.60	0
EN	59-41%	97-3%	37.24	1.9	99-1%	79-21%	28.17	0
IT	67-33%	76-24%	50.45	35.56	100-0%	81-19%	50.64	0
FR	56-44%	64-36%	69.10	51.43	100-0%	94-6%	63.60	0
DE	54-46%	69-31%	54.51	40.12	70-30%	55-45%	45.80	53.92
FA	79-21%	85-15%	83.18	56.10	100-0%	97-3%	79.47	0

able extent, low performance in the case of discontinuous VMWEs in English is notable, which is in accordance with the results for English LVCs in Table 5.9. Occurrences of English LVCs are mostly gappy with distances of more than one word, which can explain lowered performance.

The right side of Table 5.10 shows different results for single-token and multi-token VMWEs. Single-token entries are generally rare, and the only language where they occur often is German in which our system identify them equally well and even better than multi-token VMWEs.

Table 5.11: Proportion of Seen and Unseen VMWEs in gold standard and prediction data and F1 scores of our system for the two groups individually.

	Seen-Unseen			
	gold	prediction	Seen	Unseen
ES	59-41%	52-48%	53.58	19.90
EN	29-71%	45-55%	46.70	16.45
IT	64-36%	50-50%	71.27	12.50
FR	52-48%	60-40%	82.92	36.32
DE	53-47%	57-43%	72.48	20.00
FA	66-34%	67-33%	87.73	59.74
Average			69.11	**27.49**

Table 5.11 shows the results of our system on seen and unseen VMWEs separately. The system achieves respectable results on unseen data compared to seen expressions. The proportion of unseen data in the validation phase did not adversely affect performance. For a model to generalise well into unseen expressions, it is important that it learns general patterns rather than memorising exact occurrences of tokens. In the case of Farsi, for which the high performance on unseen data is considerable, all the VMWEs annotated are LVCs. LVCs in Farsi are very common and are usually constructed from a small number of verbs plus nouns. More than half of the verbs (almost all the new verbs according to Bateni (1989)) in Farsi are LVCs. This can be observed by comparing the number of sentences and VMWEs in Farsi in Table 5.4, where annotated VMWEs (which are all LVCs) outnumber the sentences in development and test datasets.

We roughly compare the average performance of our system on seen and unseen data with the average performance of the best system in the shared task. The F1 score of 27.49, that we achieved for unseen VMWEs is significantly higher than 19.71, by the best participating system. We applied our system to all languages in the shared tasks while we report the results for the languages of our focus in this thesis. Since one of the main purposes of the shared task is development of language-independent and cross-lingual VMWE identification systems, we applied our best system, ConvNet+LSTM for all languages. it is worth noting that the average performance of our system on unseen data for all languages was 28.46.

The remarkable strength of our system compared to all participating systems in the shared task is its generalisation power. Increased generalisation power could be the result of using large generic embedding representations that are trained unsupervised on large corpora and are informed about morphological word formation rules (Bojanowski et al., 2016).

Effects of labelling format

The shared task evaluation does not take into account the beginning and continuation of the expression components. A system can identify any part of the expression (by ignoring the order) and that will be considered as a true positive for token-based evaluation. However, we train our models using IOB-like labelling which differentiates between `B_LVC` and `I_LVC` as an example. When we convert the labels back to the shared task format, we filter out cases that an `I` is labelled without a preceding `B`. Our technique considers any individual `I` as a mistake of the model, which seems to have, in some cases, improved the results.

In Table 5.12, we report the results following the shared task scheme where there is no distinction between different components of VMWEs. We compare this with the results following our own labelling scheme in which the components are labelled based on an IOB-like format and some filtering has been performed if the beginning-continuation order is not satisfied. The results are reported for the ConvNet+LSTM approach on the test data. We focus a sample of selected languages to analyse possible effects on the whole

data.

Table 5.12: Effects of labelling format on performance.

	Model	Per-token			Per-MWE		
		P	R	F1	P	R	F1
EN	with filtering	60.36	18.77	28.63	48.40	18.16	**26.42**
	no filtering	57.74	20.24	**29.97**	39.57	18.16	24.90
FR	with filtering	82.94	57.73	68.08	72.39	54.22	**62**
	no filtering	80.73	60.12	**68.92**	63.53	54.22	58.50
DE	with filtering	76.22	41.74	53.94	62.38	39.96	**48.71**
	no filtering	73.64	44.41	**55.41**	54.32	40.36	46.31
FA	with filtering	93.87	74.3	82.95	86.12	71.86	**78.35**
	no filtering	93.74	75.38	**83.57**	83.53	71.86	77.25

According to Table 5.12, the results computed without filtering non-complete MWEs are slightly higher than the results for our filtering-based approach in terms of token-based F1 score. Performing no filtering increases the recall while it adversely affects precision in all cases. It is somewhat expected that when we don't apply the filter, we accept more tokens as part of MWEs; the correct ones help the recall and the incorrect ones impair the precision. Considering the per-MWE evaluation measures, the recall would not change as is expected (because by accepting the tokens which are part of an incomplete MWE we don't help the coverage of the identification system), but the considerable drop in precision causes a significant decrease in per-MWE F1 scores.

Based on these results, we believe that this IOB-like labelling scheme is more effective for automatically tagging VMWEs. It is also interesting to

apply more complex labelling scheme in future.

5.7 Summary

In this chapter, we targeted the shared task on automatic identification of verbal MWEs. We proposed our ConvNet+LSTM+(CRF) neural network based model for sequence tagging. According to the experiments, we found that the CRF should be considered as an optional layer, since it does not help the model in general. The ConvNet+LSTM model using pre-trained word embeddings outperformed all the baselines. The results of our system on the blind test data of the shared task showed the best F1 scores in several different languages and also by macro-averaging among all of them. The significantly higher performance of our system over other participating systems on unseen data and hence the best generalisation ability of our system is noteworthy.

Chapter 6

Extracting Translation Equivalents for Multiword Expressions Using Comparable Corpora

In this chapter, we explore a different area within our study of MWEs, namely, finding translation equivalents for them. Our primary focus is on the methodologies for finding translations for verb-noun collocations, and we no longer consider their idiomatic behaviour. After outlining the motivation for finding translation equivalents in Section 6.1, we discuss previous work that investigates the task for single and multi-word units in Section 6.2.

The first proof-of-concept stage of the study pursued in this chapter, covers English and Spanish and focuses on a particular subclass of MWEs: verb-noun collocations such as *take advantage*, *make sense*, *prestar atención* and *tener derecho*. In Section 6.3, we propose a novel approach to extract translation equivalents for verb-noun collocations from comparable corpora. The detail about our developed comparable corpora is outlined in Section 6.4. We report our evaluation and results in Section 6.5.

In Section 6.6, we perform comprehensive evaluation on different configurations of English and Spanish corpora. In particular, we investigate the impact of the size and quality of comparable corpora (in terms of their simi-

larity) on the specific task of extracting translation equivalents of verb-noun collocations. This study sheds some light on the more general and perennial question: what matters more? the quantity or quality of corpora? We finally summarise the chapter and our findings in Section 6.7.

6.1 Motivation

The focus of our study in this chapter is on verb-noun collocations as a particular type of MWEs. While there are many studies on the automatic extraction of collocations from monolingual text (Evert, 2005; Wanner, 2004; Smadja, 1993), only a few have drawn on bilingual resources for the automatic treatment of collocations (Corpas Pastor, 2017; Mendoza Rivera et al., 2013; Bouamor et al., 2012). Finding translations is known to be more difficult for collocations than for words (Corpas Pastor, 2017).

NLP systems that translate collocations, often do so using pre-existing lexicons constructed from collocation translations (Mendoza Rivera et al., 2013). However, as new combinations of words are created and used on a daily basis, such lexicons do not provide translations of all collocations. Thus, it is important to develop a method that can automatically find translation equivalents for multi-word collocations.

Parallel corpora are the standard resources used in Machine Translation and other multilingual NLP applications (Tiedemann, 1998). Unfortunately, parallel corpora are not widely available and do not cover all domains. An alternative and more promising approach would be to use comparable corpora

as they can be compiled from the web in a relatively straightforward way, making use of available purpose-built tools.

We propose an approach to find translation equivalents for collocations using comparable corpora. The idea is to use distributional similarity across bilingual resources. By 'equivalent expressions' or 'equivalents' we refer to expressions which are translations of each other across languages. One of the premises in this methodology is that equivalent expressions are expected to appear in the same or similar contexts across languages.

6.2 Background

Due to the scarcity of parallel corpora, comparable corpora are now increasingly used as an alternative resource in a number of multilingual applications, which include but are not limited to, machine translation (Smith et al., 2010; Rapp et al., 2016), word translation (Rapp, 1999; Gaussier et al., 2004; Pekar et al., 2006; Vulić and Moens, 2012), term extraction (Fung, 1997; Daille and Morin, 2005; Saralegui et al., 2008), bilingual document similarity (Sharoff et al., 2015; Jagarlamudi and Daumé, 2010), cross-lingual coreference resolution (Green et al., 2011), name entity transliteration (Udupa et al., 2008; Klementiev and Roth, 2006), automatic identification of cognates and false friends (Mitkov et al., 2007), and testing the validity of translation universals (Corpas Pastor et al., 2008).

The studies on the subject of extracting parallel segments from comparable corpora are further discussed in Section 2.6. Constructing and compari-

son of distributional context vectors is known to be the standard approach to bilingual lexicon extraction from comparable corpora (Bouamor et al., 2013). In order to find similarities among the words of different languages, Mikolov et al. (2013b) proposed a supervised approach derived from word embeddings. Mikolov's model takes as input a list of translationally equivalent words for whom word embeddings were individually learned in each language. Based on these separate word embeddings, it constructs a translation matrix that maps words in one language to another. Translation equivalent for any new word can be obtained by multiplying the vector for that word by the translation matrix. This model is designed to translate single words across languages. There are also more recent studies on bilingual lexicon induction for single words (Vulić and Moens, 2015; Vulić et al., 2016; Zhang et al., 2017). However, there is no consensus on how to train embeddings for multiword units (Kordoni, 2017).

6.3 Distributional Similarity Across Languages

According to the distributional similarity hypothesis, meaning is a function of distribution, in that words that co-occur tend to share semantic content (Harris, 1954). The idea is succinctly elucidated by Firth (1957), with the sentence "you shall know a word by the company it keeps". By the same token, terms that are translation equivalents may share common concepts in their contexts. These shared concepts are in turn expressed by words/terms that are translation equivalents in the two languages. For example, we might

expect to see the Spanish expression *poner en marcha* co-occurring with words, such as *problema*, *decisión* and *mercado*, and the potential English translation of it, *to launch*, co-occurring with the translations of the Spanish context words, i.e., *concern*, *decision*, *market*, respectively.

Distributional similarity has been widely used in finding pairs (words or terms) that are semantically similar; however, the applications have mainly focused on similar pairs within a single language. We use an extended version of a state-of-the-art distributional similarity method to identify translation equivalents for collocations. Specifically, we define context in a bilingual space by pairing words that are translations of each other. This context is shared between the two languages. Note that we do not rely on a clean bilingual lexicon. Instead, we take the word pairs from a noisy bilingual lexicon, which is automatically learned by using a word alignment tool.[1]

6.3.1 Word Vector Representation

To represent words using context vectors, we follow the `word2vec` method proposed by Mikolov et al. (2013c). The method employs the patterns of word co-occurrences within a small window to predict similarities among words. The idea is to represent each word as a dense vector (a.k.a. word embeddings) extracted from various training methods, which in turn have been inspired by neural-network language modelling (Collobert et al., 2011).

The word embedding approach which is employed in this study, uses a

[1]We use the lexicon derived from applying GIZA++ on the Spanish-English portion of the Europarl.

neural network to learn low-dimensional word vectors from raw monolingual text. The standard implementation of `word2vec` constructs bag-of-words contexts for all single-word terms that appear in a training corpus. We adapt the model to our task of finding translation equivalents for multi-word collocations by: (i) treating sequences of words as single units/terms, and (ii) defining bilingual contexts by drawing on a core set of known translation pairs. To do this we use the flexible word embedding approach proposed by Levy and Goldberg (2014) that allows us to define bilingual contexts.

Regular word2vec models are based solely on linear contexts. In the work of Levy and Goldberg (2014), the skip-gram model with negative sampling introduced by Mikolov et al. (2013c), is generalised to include arbitrary contexts. That is the reason we also refer to this model as generalised word2vec. Although this generalised version of `word2vec` was originally used to extract dependency-based word embeddings, its ability to accept arbitrary contexts makes it possible to be easily adapted to our specific task of bilingual vector construction for multi-word collocations.

6.3.2 Bilingual Phrase Vector Representation

In standard `word2vec`, using a window of size k around a target word w, $2k$ context words are produced: k words before and k words after w. We base our context extraction on this standard, with the difference that we extract only specific words rather than all the words in the context window. Our essential context words come from a bilingual dictionary. Specifically, we

focus on nouns as the most important components of meaning. In most of the previous work in the literature on the study of word vector representations, nouns are the focus of the evaluation since they are less semantically ambiguous (Mikolov et al., 2013b; Zhang et al., 2017). We use a core lexicon of paired English-Spanish nouns as our bilingual context terms.[2] The generalised `word2vec` model (called `word2vecf`) [3] can then be trained on these pairs, resulting in the vectors of the two languages to be defined over the same space (of paired English-Spanish nouns), and to be comparable.

6.3.3 Translation Equivalent Extraction

Given a target collocation s from the source language (e.g., Spanish), our goal is to find the best translation equivalent in the target language (e.g., English). First, we identify a set of candidate translations for s, from a Spanish-English comparable corpus that we automatically build by pairing documents from the two languages. Next, we rank these candidates according to their semantic similarity to the target collocation. The following subsections explain these two steps in more detail.

Candidate Extraction. To extract candidate translations for a collocation, we examine a set of automatically paired comparable documents from the two languages. Specifically, for each collocation s, we examine all target language documents that are paired to the source language documents con-

[2]Our window size is 5 and the words in the window that are not in the bilingual core lexicon are ignored.

[3]The software is available in the websites of the authors at `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`

taining s. As candidate translations for s, we take a set of frequent unigrams, bigrams, and trigrams (which are verb combinations) appearing in these documents. [4] The details of compiling and pairing documents in comparable corpora are explained in section 6.4.

Ranking Candidates using Cross-lingual Similarity. We construct a cross-lingual vector representation for each collocation s, and for each of its candidate translations, drawing on our proposed approach for defining a cross-lingual semantic space (see Section 6.3 above). The winning candidate is the one that has the highest similarity to the collocation s.

6.4 Comparable Corpora

While parallel corpus is a corpus that contains source texts and their translations, a comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness (McEnery and Xiao, 2007). These include the same proportions of the texts of the same genres, domains, and sampling period in a range of different languages. The sub-corpora of a comparable corpus might or might not be translations of each other. Rather, their comparability lies in their similarity.

[4] We set the frequency threshold to 10 in our experiments.

6.4.1 Compilation

We build a corpus of comparable English-Spanish documents from various news sources on the Web. News texts are rich sources of shared content, and hence have commonly been used to construct comparable corpora (Fung, 1998; Munteanu and Marcu, 2005; Aker et al., 2012). To build the corpus of comparable English-Spanish documents, we collect news feeds from a variety of news sources, including the ABC news,[5] Yahoo news,[6] CNN news,[7] Sport news,[8] and Euronews[9] in both Spanish and English languages. We use a tool from the ACCURAT project[10] to extract comparable documents from the news texts (Aker et al., 2012).

ACCURAT also comes with a tool, called DictMetric, which is designed to measure the comparability levels of document pairs via cosine similarity (Su and Babych, 2012). The tool is specifically proposed to provide a data for extracting parallel segments with high performance. To measure the comparability of two documents in different languages, one language gets translated to the other. The tool translates non-English texts into English by using lexical mapping from the available GIZA++ based bilingual dictionaries.

Since the proportion of overlapped lexical information in two documents is the key factor in measuring their comparability, the tool converts the texts

[5] `http://www.abc.es` and `http://www.abc.net.au`
[6] `http://es.noticias.yahoo.com` and `http://uk.news.yahoo.com`
[7] `http://cnnespanol.cnn.com` and `http://cnn.com`
[8] `http://www.sport.es/es` and `http://www.sport-english.com/en`
[9] `http://es.euronews.com` and `http://euronews.net`
[10] `http://www.accurat-project.eu`

into index vectors and then computes the comparability score of document pairs by applying cosine similarity measure on the index vectors. Using the ACCURAT toolkit, we compute the comparability of all pairs of Spanish and English documents.

6.4.2 Size and Quality of Comparable Corpora

The size of the corpora whether monolingual, parallel or comparable, is often regarded as a decisive factor for the performance of NLP tasks or applications, the expectation being that the larger the corpora used, the better the performance of the tasks or applications exploiting them. It would be noteworthy to establish whether the size of the corpora is an important factor irrespective of their quality and whether even sufficiently large data of inferior quality could deliver better results than smaller data of better quality.

In order to answer this question in our study of automatic translation of multiword expressions using comparable corpora, in Section 6.6, we experiment with corpora of different sizes and quality and compare the results. To the best of our knowledge, this is the first study which seeks to shed a light on this fundamental question.

In the study reported in Section 6.6, our premise is that quality of comparable corpora is directly related to their comparability: the more comparable they are, the better their quality is deemed. For the purpose of the study, we operationalise the concept of comparability and quality through similarity: the more similar two corpora are, the more comparable they are. In

other words the comparable corpus built on these corpora would be of higher quality.

6.5 Experiment 1: Mining for Translations of Collocations from Comparable Corpora

In this section, we report on our experiments on finding translation equivalents for verb-noun collocations from comparable corpora.

6.5.1 Corpora

In this experiment, we focus on news documents from July to December 2015. By using the ACCURAT toolkit, we extract and pair documents with the comparability score (cosine similarity) of higher than 0.45, and consider them as aligned. This results in 16,436 English documents (with around 11 million word tokens) and 11,468 Spanish documents (with around 6 million word tokens). Each English document is paired to at least one Spanish document; equally, there is at least one paired English document for every Spanish document.[11]

6.5.2 Target Group

Our methodology is to use bilingual word vector representation to find translations for collocations across comparable corpora. To report the results, we focus on nine highly frequent verbs in English and six in Spanish. These verbs

[11]The comparable corpora that we prepared is available on `https://github.com/shivaat/EnEsCC`.

tend to frequently combine with many different nouns in their direct object positions to form multi-word collocations. The verbs are: *take*, *have*, *make*, *give*, *get*, *find*, *pay*, *lose* in English, and *tener*, *dar*, *hacer*, *formar*, *tomar*, *poner* in Spanish. We extract all occurrences of these verbs followed by a noun, from the whole news corpora. This process results in 1,007 English, and 930 Spanish Verb+Noun collocations. Among these candidate expressions, only 162 English expressions and 187 Spanish expressions occur with frequency higher than 9 in our paired comparable documents. We run the experiments only on these expressions.

6.5.3 Vector Construction

Recall that to construct vectors for our English and Spanish expressions, we need a seed list of paired context words (a.k.a., the bilingual context pairs). For this purpose, we use a subset of the word alignments resulting from applying GIZA++ on the English-Spanish Europarl parallel corpus (Koehn, 2005). Specifically, we only consider pairs of frequent nouns that have an alignment probability of higher than 0.2, where frequent nouns in a language are those that appear 50 times or more in Europarl. As a result we have a list of 4,700 bilingual contexts.

For learning the vectors, we use the following corpora to extract word co-occurrence statistics: the monolingual English and Spanish components from the Europarl, and the English and Spanish components of our news corpora. We index all the English and Spanish verb combinations (unigrams, bigrams,

trigrams) according to their occurrences with the context word pairs. Specifically, from the window of 10 words around a target expression, we capture any word that exists in our bilingual context pairs (focusing on the relevant language given the language of the target expression). The word2vecf software is then used to train vectors on the indexed corpus. We then apply our methodology to find translations for collocations in both directions: Spanish to English, and English to Spanish.

Note that we focus on finding translations for Verb+Noun combinations. We assume that for most such expressions, the translation equivalent is either a Verb (unigram), a Verb+Noun (bigram), or a Verb+Noun with an intervening word, such as a determiner or an adjective (trigram). We thus consider as our candidate translations all unigram verbs, bigram Verb+Noun, and trigram Verb+Noun combinations with an intervening word. For this purpose, the corpus is pre-processed and bigram and trigram Verb+Noun expressions are annotated as single tokens. For every expression from the source language (e.g. Spanish), our goal is to find the five most cross-lingually similar verb or Verb+Noun combination in the target language (e.g. English).

6.5.4 Evaluation and Results

We evaluate our methodology called `bi-word2vec` by comparing it with a baseline approach which is explained below and we call it **co-occurrence Jaccard**. We evaluate the two approaches also on loosely comparable corpora.

Co-occurrence Jaccard. We implement a simple distributional similarity approach as our baseline. Given two expressions (from the two languages), we measure their similarity by comparing their corresponding sets of bilingual context pairs using a context window of size 10. We use the Jaccard coefficient to measure similarity. The baseline is applied to our comparable corpora in order to find translation candidates for each expression, relying on the above simple similarity measure to rank these candidates.

Using Loosely Comparable Corpora. We also perform experiments to investigate the advantage of using comparable corpora with high level of similarity for finding the candidate translations of an expression. Accordingly, we add noisy alignments to our accurately aligned documents. Specifically, for each source-language (e.g. Spanish) document, paired with several highly similar target-language (e.g. English) documents, we align an extra set of 2,000 randomly selected target-language documents.[12] This process results in a larger but noisy corpus of comparable documents. Our goal here is to understand whether using a larger set of documents that may contain more candidate translations is helpful, despite the noise. That is, we intend to understand whether a method like word2vec is sufficiently robust to noise, and hence capable of finding good translations from documents that are not perfectly aligned. If that is the case, then we can avoid the rather expensive process of building highly accurate comparable corpora. We apply both the baseline and our proposed approach (the one that uses word2vec) to this

[12]Note that we add noise in both Spanish-English and English-Spanish directions.

noisy data, and compare the results with those on the smaller corpora with more accurately aligned documents.

6.5.5 Results and Discussion

We ask a human expert to rate the top-ranked translations produced by each of the methods for each expression. We ask the expert to give a rating of 1, if there is at least one good translation within the top 5 rank in the list; otherwise, the list is given a rating of 0. We also have 25% of the resulted translation lists annotated by a second annotator. The inter-annotator agreement, in terms of Kappa, is 0.80. The measure is computed for finding translations of both English and Spanish expressions.

Note that we use a similarity measure to rank the candidate translations of each expression. By using different threshold values for this similarity, we get ranked lists of varying sizes. The higher this threshold, the smaller the number of the resulting translation candidates, and hence the higher the number of expressions for which we may not have any good translations. In other words, we can trade off accuracy (precision) for coverage (recall). We thus set the similarity thresholds to different values in order to measure accuracy for varying degrees of coverage (from around 10% to around 80%). Doing so gives us a better understanding of the overall performance of each method.

Table 6.1 shows accuracy and coverage values for finding translations of the Spanish expressions; Table 6.2 gives the results for English expres-

sions. Note that we show the results for both the baseline (`Co-occurrence Jaccard`) and the `bi-word2vec` method, using both corpora of comparable documents: the smaller and less noisy corpus of highly-comparable documents (referred to as paired CC), and the larger and noisy corpus (referred to as CC + noise).

Table 6.1: The accuracy of the baseline compared to the `bi-word2vec` approach in extracting translations of Spanish Expressions.

	coverage	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
paired CC	baseline	82%	55%	24%	22%	18%	16%	12%
	bi-word2vec	50%	46%	**40%**	**36%**	**34%**	**32%**	**33%**
CC + noise	baseline	78%	50%	24%	18%	14%	13%	8%
	bi-word2vec	44%	45%	**38%**	**37%**	**30%**	**33%**	**32%**

Table 6.2: Comparing the accuracy of the baseline with the `bi-word2vec` approach in extracting translations of English Expressions.

	coverage	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
paired CC	baseline	79%	52%	46%	35%	26%	22%	18%
	bi-word2vec	39%	37%	34%	36%	**34%**	**29%**	**31%**
CC + noise	baseline	70%	50%	24%	22%	18%	12%	13%
	bi-word2vec	38%	34%	**31%**	**39%**	**39%**	**32%**	**31%**

As can be seen in the first rows of both tables, the baseline accuracy/-precision is high when we limit the method to a very low coverage/recall, but drops quickly as we increase coverage. Note that when coverage is low, many expressions do not have any translation equivalents. But those that do have candidates, have a few accurate ones, and hence it is easy for a simple

method such as the baseline to pick the best.

Compared to the baseline, the `bi-word2vec` approach is more stable across the different degrees of coverage in both translation directions: in fact, the performance of `bi-word2vec` drops only slightly when we move from a coverage of 30% to almost 80%. More importantly, even for a very high degree of coverage (i.e., 70%–80%) `bi-word2vec` performs much better than the baseline in terms of accuracy (33% compared to 12% for Spanish-to-English, and 31% versus 18% for English-to-Spanish). The better performances of `bi-word2vec` over the baseline which are shown in bold are all statistically significant with $p < 0.01$ for Spanish and $p < 0.10$ for English.

Next, we compare the results using the two corpora. Investigating the baseline approach over the two corpora, we observe that in almost all coverage ranges the performance of the model drops when using the noisy paired documents. This can be seen in both Table 6.1 and Table 6.2 for both Spanish to English and English to Spanish translations. Then we compare the results of `bi-word2vec`: Interestingly, the performance of `bi-word2vec` is reasonably close in the two different corpora, even though the CC + noise has a much higher degree of noise. When using the larger noisy corpora, in some cases, which are underlined in the table, `bi-word2vec` results in better accuracy. This is an interesting result, suggesting that even using a large but noisy corpus of comparable documents, we can find reasonable translations for multiword collocations by relying on a robust and accurate method such as `bi-word2vec`.

Some examples of translations extracted using `bi-word2vec` from our paired documents are shown in Table 6.3. Correct translations, if there are any, are shown in bold.

Table 6.3: Translation equivalents (Spanish to English) extracted using `bi-word2vec` from our paired documents for two different ranges of coverage: 10%-20% and 70%-80%.

Source expression	Freq	coverage: 10%-20%	coverage: 70%-80%
tener lugar	440		**take-place**, have-no-idea, describe-the-situation, have-no-place, open-the-possibility
tomar medidas	81		have-no-choice, **measure**, have-no-reason, become-a-member, have-no-idea
poner fin	323		have-no-idea, tell-the-truth, make-no-mention, reach-the-end, **put-a-stop**
hacer falta	64		make-no-mention, **lack**, have-no-information, have-the-time, shelve
hacer referencia	82		have-no-knowledge, open-the-possibility, have-no-intention, make-no-mention, insert
dar apoyo	13		
hacer eco	70	**echo**	use-social-media, refute, follow-statement, time-the-announcement, **echo**
tener éxito	38	**have-success**	rediscover, **have-success**, lay-the-foundations, change-strategy
hacer justicia	37	get-justice, **bring-to-justice**,	get-justice, deliver-justice, demand-justice, **bring-to-justice**
tomar fotografas	20	**take-photos**, snap-pictures, orbit-lab, **take-pictures**, orbit,	**take-photos**, snap-pictures, catch-a-glimpse, take-full-advantage, **take-pictures**
tener mayoría	24	**win-a-majority**	convince-voters, gain-popularity, govern-coalition, serve-two-terms, face-accusations
hacer preguntas	43	make-no-mention,	betray, impress, reshape, admire, astonish
hacer cambios	22	need-a-change,	need-a-change,
tener cáncer	17	**have-cancer**, die-of-cancer, leave-flowers, lay-flowers, battle-cancer,	mourn, die-of-cancer, **have-cancer**, leave-flowers, undergo-treatment,

As can be seen in Table 6.3, keeping the recall low not only results in higher precision, but also yields better translations. For example in the case of *tener mayoría*, the better translation, *win a majority* is seen at lower recall

level, but lost at a higher level of recall. We can also see that there is not much difference in how the system deals with expressions of different frequencies; for example, the system has difficulty in finding exact translations for even a fairly frequent expression like *hacer referencia*.

Semantically Coherent Collocations. Our experimental Verb+Noun combinations (that we try to find translations for) include a range of expressions, from frequent combination of words (*get things*), to multi-word verbal units (*make reference*), to more idiomatic expressions (*take place*). It is thus interesting to find out whether the performance of our method differs when applied to different types of expressions. For this, we take a subset of expressions from each language that has been annotated as a semantically-coherent MWE by two annotators. This selection process results in 80 Spanish and 101 English expressions. Table 6.4 shows accuracy of the `bi-word2vec` method for both Spanish and English subsets when coverage is set to around 80% (using the cleaner comparable corpora for finding candidates).

Table 6.4: The accuracy of the `bi-word2vec` approach in extracting translations of multiword collocations from comparable corpora.

	accuracy	
	Spanish	English
`bi-word2vec` approach	48%	44%

The results in Table 6.4 show that, for both languages, accuracy is higher when we focus on these subsets (48% versus 33% for Spanish expressions,

and 44% versus 31% for English). This list excludes literal expressions like *dar respuesta*, *tener hijos* and *dar apoyo* for which word-for-word translation might give better results.

6.6 Experiment 2: What Matters More: The Size of the Corpora or their Quality?

The main objective of this experiment is to measure the effect of various configurations of comparable corpora on the task of translation equivalent extraction. We experiment with corpora of different size and quality in order to establish their impact on the performance. We employ the same methodology as explained in Section 6.3 and is used in the experiment in Section 6.5.

6.6.1 Corpora

Two comparable corpora are used for our experiments. One is a collection of aligned documents from English and Spanish Wikipedia [13]. It includes around 673,000 document pairs with 456.6 million English and 316.2 million Spanish tokens. The documents are aligned one by one using the language links in Wikipedia pages; therefore, they are accurately aligned based on their contents and regarded as high quality corpora in terms of comparability.

We compile the other comparable corpora from various news sources on the web using the ACCURAT toolkit (Pinnis et al., 2012; Skadiņa et al., 2012; Su and Babych, 2012) as explained in Section 6.4.1. For this experiment, we

[13] `http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/`

collect news feeds in both Spanish and English from July 2015 to February 2016. In addition to the news feeds listed in Section 6.4.1, RSS feeds of Ultimahora[14] and Europapress[15] for Spanish are also added to ensure the Spanish data is more balanced. The downloaded data from online news (1.5 GB) consisted of 200,000 documents in English and 112,000 documents in Spanish.

These documents are classified with a view to building English-Spanish comparable corpora. For the purpose of this study we operationalise comparability via similarity. Similarity is automatically computed with the help of the ACCURAT tool, DictMetric, which as explained in Section 6.4.1, translates documents of one language to the other, converts texts into index vectors and compares document vectors from the two languages by employing cosine similarity.

By varying comparability thresholds, we generate comparable corpora of different size and quality. Recall that in this study, quality of comparable corpora is the degree of their comparability, which in turn is modeled with their degree of similarity. It is expected that higher comparability thresholds would yield more accurate alignments between documents but also results in generation of smaller corpora. To this end, we set the comparability threshold to five values from 0.5 to 0.1. The number of paired documents in each of these five sets are reported in Table 6.5.

[14] `https://ultimahora.es`
[15] `http://www.europapress.es`

Table 6.5: The number of paired documents in the news comparable corpora in both directions.

	CC 0.5	CC 0.4	CC 0.3	CC 0.2	CC 0.1
es-en	6,544	15,417	31,856	63,703	114,313
en-es	9,337	22,123	46,798	94,896	195,175

6.6.2 Target Expressions

In this study, we experiment with the most productive and widely used verbs in Verb + Noun combinations. We focus on eight highly frequent verbs occurring before nouns in English, and six such verbs in Spanish. [16] All such occurrences are extracted from our paired documents (a.k.a. comparable corpora). We only select occurrences with frequencies higher than 3 in the aligned documents of similarity threshold 0.5. This process results in a list of 220 English and 210 Spanish Verb + Noun collocations.

6.6.3 Experimental Setup

For learning vectors, as explained in Section 6.5.3, the monolingual English and Spanish components from the Europarl, and the English and Spanish components of our news corpora are used to obtain co-occurrence statistics. All English and Spanish verb combinations (unigrams, bigrams, trigrams) are indexed according to their occurrences with the context word pairs. Specifically, words (nouns) that exist in our bilingual context pairs are identified within a context window of length 10 around a target expression. As a result,

[16] English verbs: *take*, *have*, *make*, *give*, *get*, *find*, *pay*, *lose*; Spanish verbs: *tener*, *dar*, *hacer*, *formar*, *tomar*, *poner*.

expression-context pairs are generated.

Similar to the previous experiment, two approaches are experimented with to construct vectors:

co-occurrence Jaccard: `co-occurrence Jaccard` is our baseline approach, as explained in Section 6.5.4. It measures Jaccard similarity of corresponding context words of two expressions in order to find their similarity.

bi-word2vec: `bi-word2vec` is our proposed approach for vector representation of expressions and is detailed in section 6.3.2. In this study, the word2vecf software is used to train vectors on the indexed corpora.

Experimenting with both types of vectors, we apply our methodology (as explained in Section 6.3.3) to find translations for collocations in both directions: Spanish to English, and English to Spanish. Note that we focus on finding translations for Verb+Noun combinations. We assume that for most such expressions, the translation equivalent is a Verb (unigram), a Verb+Noun (bigram), or a Verb+Noun with an intervening word such as a determiner or an adjective (trigram). For every expression from the source language, our goal is to find the five most similar Verb or Verb+Noun combinations (bigram or trigram) in the target language.

6.6.4 Gold Standard

For the purpose of the evaluation, we prepare a list of correct translations for the candidate collocations from online dictionaries such as Wordrefernce [17],

[17] `https://www.wordreference.com/es/translation.asp`

Linguee [18], Spanish Central [19], and Reverso Dictionary [20]. We also ask a human expert to examine and rate the top-ranked translations of four sample result lists and mark the correct translations. We extend the gold standard list with the correct translations marked by the annotator. Note that our approach might extract correct translations which are not on the list.

6.6.5 Evaluation: Size vs. Quality of Comparable Corpora for Translating MWEs

This is the first study seeking to establish the impact of different comparability thresholds which control the quality of the selected paired documents. The threshold values are directly proportional to the quality in terms of comparability and inversely proportional to the size. Higher comparability threshold implies better quality but also means smaller corpora. Lower comparability thresholds generate larger corpora of inferior quality. The experiments in Section 6.5.5 suggest that bigger corpora even if noisy can potentially help finding better translation equivalents. The performance of the task of finding translation equivalents is evaluated by applying the two distributional similarity approaches on the five groupings of paired documents based on their comparability (referred to as CC0.5, CC0.4, CC0.3, CC0.2, CC0.1).

For each expression, we use a similarity measure to rank the candidate translations. By setting different threshold values for this similarity, we ob-

[18] `https://www.linguee.com/english-spanish`
[19] `http://www.spanishcentral.com`
[20] `https://dictionary.reverso.net/spanish-english/`

tain ranked lists of varying lengths. The higher this threshold, the smaller the number of the resulting translation candidates, and hence the higher the number of expressions for which we may not have any good translations. In other words, we trade accuracy for coverage. In our experiments we set the similarity thresholds to different values in order to measure accuracy for three degrees of coverage (20%, 50% and 80%). These different configurations offer a meaningful picture of the overall performance of a method on each comparable corpus.

Table 6.6 displays the accuracy and coverage values for finding translations of both Spanish (es) and English (en) expressions.

Table 6.6: The accuracies compared on different sets of comparable corpora.

	coverage	20%		50%		80%	
		es	en	es	en	es	en
Co-occurrence Jaccard	CC 0.5	37%	36%	15%	15%	10%	9%
	CC 0.4	42%	52%	24%	24%	13%	14%
	CC 0.3	63%	65%	29%	31%	16%	15%
	CC 0.2	**64%**	**67%**	**40%**	35%	**20%**	20%
	CC 0.1	45%	58%	38%	**40%**	20%	**23%**
bi-word2vec	CC 0.5	38%	28%	25%	17%	11%	14%
	CC 0.4	32%	34%	34%	23%	24%	18%
	CC 0.3	**37%**	**48%**	**36%**	**31%**	28%	24%
	CC 0.2	37%	44%	34%	30%	**31%**	25%
	CC 0.1	31%	43%	24%	29%	27%	**27%**

As illustrated in Table 6.6, usually, the choice of lower comparability threshold yields better results provided that the threshold is not lower than a specific threshold value (e.g. 0.2 for Spanish when using `co-occurrence`

`Jaccard`). The larger size of the corpora 'matters' up to that point, as long as the corpora exhibit minimal quality (e.g. comparability 0.2). In these cases the accuracies drop when the threshold is set to 0.1. This trend generally holds for both distributional similarity approaches. However in the case of `bi-word2vec`, the optimal threshold for comparability is 0.3 for lower coverages and shows the importance of quality. With `bi-word2vec` for Spanish, the more drastic drop in accuracy is when we use CC 0.1 rather than CC 0.2. This is only violated in higher coverages for English for which size pose as a counterpoise to quality.

A general conclusion from these results is that size indeed matters and the larger the size, the better the performance, as long as the quality is above a minimal comparability threshold.

The performance of the models is further evaluated on the accurately aligned Wikipedia comparable corpora and reported in Table 6.7. In terms of accuracy, `bi-word2vec` does better than the simple `co-occurrence Jaccard` at establishing the translations of Spanish expressions (es). On the other hand, the simple co-occurrence Jaccard fares better at finding the translations of English expressions (en).

Table 6.7: Accuracies (%) of models in finding translations from aligned Wikipedia comparable corpora.

coverage	20%		50%		80%	
	es	en	es	en	es	en
co-occurrence Jaccard	64	84	58	68	40	46
bi-word2vec	74	78	63	55	48	43

The Wikipedia aligned corpus is almost seven times bigger than our news corpora. To compare these two, we focus on a sample of the Wikipedia documents which are of comparable size with our news corpora (specifically, on 96, 193 document pairs). Figure 6.1 shows that, for both English and Spanish expressions, the translation results from the Wikipedia aligned corpora (wikiJ and wikiW) are significantly higher than our automatically paired comparable corpora (ccJ and ccW). As the Wikipedia aligned corpus is deemed to be of better quality in terms of comparability, the above finding confirms that quality does have an impact.

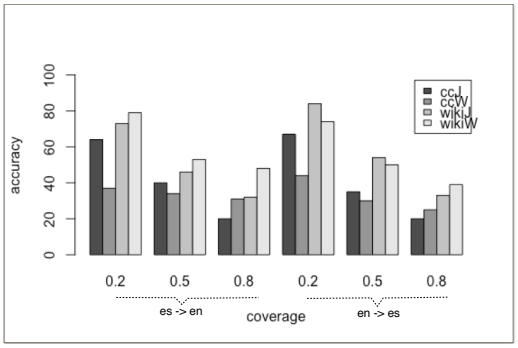


Figure 6.1: Accuracy of models in finding translation equivalents using: ccJ (co-occurrence Jaccard on our comparable corpora), ccW (`bi-word2vec` on our comparable corpora), wikiJ (co-occurrence Jaccard on wikipedia comparable corpora), wikiW (`bi-word2vec` on wikipedia comparable corpora).

Finally, according to the obtained results, the simple `co-occurrence Jaccard` approach performs very well at finding translations for English expressions. It appears that this approach delivers promising results for highly frequent expressions (e.g. *have time*) for which the `bi-word2vec` approach suggests semantically related but incorrect translations (e.g. *tomar tiempo*,

haber tiempo, *pasar mucho tiempo*).

6.6.6 Comments on Assessing Size and Quality

We study the particular task of automatic translation from comparable corpora and show that the employment of larger aligned corpora results in identifying translation equivalents with higher accuracy. At the same time, the importance of quality of the corpora cannot be underestimated. If the quality of comparable corpora is under a specific minimal threshold, the performance deteriorates. Therefore, we can conclude that both quantity and quality matter with comparable corpora of larger size delivering better performance as long as they are of minimal quality.

6.7 Summary

In this chapter, we proposed a new bilingual distributional similarity-based model derived from neural word embedding approach in order to find translation equivalents for verb-noun collocations from comparable corpora. Our model features bilingual context in order to extract vectors. Rather than extracting vectors for single words, we devised representations for the combination of verbs and nouns. We developed English-Spanish comparable corpora and cross-lingually aligned their documents based on similarity. Our model for mining translation equivalents of verb-noun collocations is compared with a baseline model in the first experiment of this chapter. We showed the efficacy of the proposed method over the baseline.

We further conducted a comprehensive experiment on the effect of size and quality of comparable corpora on the performance of the model. We conclude that size and quality both play a part in the model performance. However, for the quality to be sufficiently high, the requirement is that the aligned corpus be of reasonable size.

Chapter 7

Conclusions and Future Work

The main goal of this thesis was twofold: identification of multiword expressions and translation equivalent extraction for them. We researched methods for identification of MWEs from several different perspectives including data preparation, task modelling and evaluation techniques. We also targeted publicly available datasets for MWE identification and trained and evaluated our new proposed systems on them. In the case of automatic translation of MWEs, we focused on resource-poor languages and extracted translations from comparable corpora.

This chapter first summarises the research questions, proposed methods and findings of this thesis and then discusses the limitations and outlines future research directions.

7.1 Automatic Identification of MWEs

The first research question regarded the problem of scarcity of gold standards for MWEs.

RQ1: To what extent can the challenges with the availability and appropriateness of gold-standards for computational treatment of MWEs be resolved?

We addressed this question in Chapter 3 by preparing datasets of verb-noun MWEs in two languages: Italian and Spanish. Having done pilot annotations, we noted the difficulties in out-of-context annotations of MWEs, and initiated in-context annotation of verb-noun MWEs in Spanish and Italian. This is to our knowledge the first compiled data for these two languages. We extensively discussed the effectiveness of these datasets for disambiguating literal/idiomatic usages of verb-noun expressions. We made the datasets publicly available online.[1]

The second research question involved disambiguating between idiomatic and literal usages of expression types.

RQ2: How can we disambiguate between different occurrences of the same expression type which are idiomatic in some contexts but literal in others?

The focus was on the expression types whose token occurrences have identical surface realisations but have different literal or idiomatic interpretations depending on their contexts. Distributional similarity methods are commonly used for this task and have been previously experimented with

[1]urlhttps://github.com/shivaat/itVN

alongside syntactic and/or lexical features. We do not find syntactic and lexical properties of expression components helpful for our task, since our target expressions did not have much variation in their different occurrences.

The novelty of our approach lies in effective use of neural word embeddings as features in a classification scenario. We evaluated our system on our prepared Italian data and presented the better performance over baselines. Our method showed particularly better results over majority baseline in the case of expression types with more ambiguous behaviour.

The third research question was posed to inquire about the reliability of previously proposed ways of modelling MWEs and their evaluation.

RQ3: What is the reason behind the significant variation in reported results for MWE identification and is there a better way of modelling MWEs and more reliable evaluation methodology?

Various models can be found in the literature for identifying MWEs. We argued that choosing an appropriate model for a target dataset is important and concluded that for our specialised data which includes concordances of a selection of verb-noun expression types, a classification model would be more effective than a tagging approach. This effectiveness was quantitatively shown by devising an experiment where we applied both a simple classification and a standard tagging approach on our datasets and achieved better results from classification compared to tagging.

Regarding evaluation, we argued that many of the favourable results re-

ported for MWE expression identification might be due to overfitting. Even if test sets are controlled to be blind to training sets, there would inevitably exist a considerable overlap between train and test. The overlap comes from occurrences of the same expression types in different usages. This would be problematic in cases when an expression type has consistent behaviour, with most of the cases having either idiomatic or literal interpretation. This phenomena can also be understood as within-class imbalance that arises when each class has subsections that are not equally represented. In this case, for example, majority of the expressions tagged as MWE have a consistent behaviour, almost always occurring as MWE (belong to the subcategory of being consistent). The opposite category is underrepresented and the majority class baseline has a high performance.

Conventional cross-validation techniques easily overfit on over-represented category and the results from this cross-validation fail to reflect this issue. Therefore, evaluations of this kind lead to misleadingly high results. To counter this issue, we propose a novel approach using type-aware cross-validation in which we distribute expression types into separate folds. As a result, a classification model performs cross-type learning. This learning and evaluation model has the following advantages.

- Its performance results effectively represent generalisation power of the model. A model with high performance in type-aware cross validation can guarantee effective learning of new unseen expressions.

- The results from type-aware cross-validation can be considered as lower-bound performance for a model on blind test data.

The fourth research question is about the recent direction in identifying MWEs which is framing the task as a sequence tagging problem.

RQ4: In the case of automatic identification of MWEs in running text or tagging corpora for MWEs which is a recent direction in NLP studies on MWEs, to what extent can we improve the state-of-the-art?

We answered this question by proposing a deep neural network architecture to perform structured sequence tagging. Inspired by the recent synthesis of structured prediction models and recurrent neural networks for the task of Named Entity Recognition (NER), we introduced a similar architecture adapted to the task of MWE identification. We incorporated convolutional neural network models into our system since they are known to act as ngram detectors and can extract informative features for MWE identification.

We proposed two neural network architectures: one is a combination of two ConvNet and one LSTM layers, and the other adds a CRF layer to the combination of LSTM and ConvNets. To the best of our knowledge, this is the first experiment with a hybrid MWE tagger that integrates traditional structured prediction and neural network models into a unified architecture.

We focused on the data provided by the recent shared task on 'automatic identification of verbal multiword expressions', which includes datasets of

several languages annotated for verbal MWEs (VMWEs). The aim was to compete with other teams in order to build the best system for tagging VMWEs in corpora. The systems were evaluated in each and among all languages, based on various different perspectives, i.e. their ability to tag: VMWEs in general, their individual components, their different categories, continuous and discontinues VMWEs, multi-token and single-token VMWEs, and seen and unseen VMWEs.

We proposed a slight alteration to the labelling format of the data to make it more similar to IOB labelling which is standard for structured prediction and tagging. Then we applied our neural network models using pre-trained word embeddings and one-hot representations of POS tags as input. We do not use any task-specific domain knowledge beyond generic POS tags and this helps the model learn features independently of language.

The pretrained embeddings used in our experiments are based on an embedding technique proposed by Bojanowski et al. (2016) which is designed to model morphology by integrating subword information and has been shown to generalise well to rare words. In the case of Spanish, we also experimented with dependency based embeddings which were trained on a dependency parsed dataset. The embeddings resulted in a slightly better performance, however since dependency parsed texts are not always available for resource-poor languages, we continued working with Bojanowski et al. (2016)'s embedding representations for the rest of our experiments.

We conducted extensive evaluation for a selected number of languages in

this thesis. We first compared the results of our model with strong baselines on the first edition of the shared task data for three languages. We implemented two baseline models; one is standard CRF and the other is CRF augmented with word embeddings. Our models significantly outperformed the baselines. We then evaluated our models on validation datasets of the second edition of the shared task for six languages. Our ConvNet+LSTM model overall performs better than ConvNet+LSTM+CRF. We conclude that a simple addition of a CRF layer to a well-performing ConvNet+LSTM network does not necessarily improve the results.

Finally, we compared the results of our best model (ConvNet+LSTM) with those of all participating systems on blind test data. Our system outperformed the best participating system on five out of six languages in terms of the token-based evaluation measure and four out of six languages in terms of the MWE-based metric. ConvNet+LSTM performs well in detecting both single-token and multi-token VMWEs. The system works significantly better for continuous VMWE over discontinuous ones. The low performance is mostly in the case of English discontinuous VMWEs. Taking a closer look at the data, we concluded the underlying reason is the the prevalence of long distance between components of English LVCs which constitute a high proportion of English VMWEs.

Our system performs significantly better on unseen VMWE types than other systems in the shared task (higher F1 score of 7.78 in a rough comparison). Increased generalisation power could be the result of using large

generic embedding representations that are trained unsupervised on large corpora and are informed about morphological word formation rules.

7.2 Automatic Mining for Translations of MWEs

The fifth research question regarded the approaches to find translation equivalents for MWEs from comparable corpora.

RQ5: Since parallel data is limited, can we determine a new approach to extract translation equivalents for verb-noun multiword expressions that works better than methods used in previous studies on translation equivalent extraction?

To address this question, we first compiled English-Spanish comparable corpora from the news. By using an available toolkit, the documents of the corpora are aligned based on their comparability (which is computed with the cross-lingual similarity between documents). The aligned comparable corpora are available to the community for future research as there is no counterpart as of yet.

Then, we proposed a new bilingual distributional similarity approach to extract translations from corpora. We worked with a recently devised word embedding methodology with arbitrary contexts and adapted the model to our task in two ways by: 1) constructing vectors for sequences of words rather than single words, 2) defining bilingual contexts by drawing on a core

set of known translation pairs. The model traverses document alignments in order to find the best translations for candidate verb-noun expressions. We implemented a distributional similarity method as a baseline, demonstrating how our model achieves better results. We also discovered that our approach has a higher degree of robustness in processing of noisy comparable corpora compared to the baseline.

We further investigated the effects of size and quality of comparable corpora on the performance of extracted translation equivalents. The goal was to, for the first time, find an answer to the question: what matters more? the quantity or quality of corpora? We assorted the aligned documents based on comparability thresholds. By setting higher comparability thresholds, we had fewer aligned documents in a category, hence a smaller corpus. The results showed that both size and quality of comparable corpora are important for finding better translations. Specifically, when we decrease the comparability threshold to gather a higher number of aligned documents with lower quality, we steadily see improvement in the results. But this continues only up to a point. If the quality of comparable corpora is under a specific minimal threshold, performance deteriorates.

We also applied our model to the accurately aligned Wikipedia comparable corpus which is of fairly large size and reported the improvement of translation accuracies compared to our aligned comparable corpora. We conclude that automatically aligned comparable corpora using current methodologies are not yet as effective in finding translation equivalents in comparison with

knowledge-based aligned corpora such as Wikipedia.

7.3 Future Work

Multiword Expressions has been a broad and long-lasting research topic in NLP. To study the different idiosyncratic behavior and applications of MWEs, various directions can be pursued. In this section, we mostly outline future directions related to the methodologies and discussions particular to this thesis.

In Chapter 3, we devised datasets for identifying verb-noun MWEs in context for Italian and Spanish. The datasets contain occurrences of continuous verb-noun expressions along with their concordances. One limitation of the datasets is their restriction to continuous verb-noun expressions. Such corpora were helpful in the specific task of disambiguating between different usages of the same expression types. However, in general, datasets annotated for both continuous and discontinuous MWEs are more useful in practice. Constructing annotated corpora of this kind is a necessary future direction. Furthermore, to investigate the effects of contextual information, we suggest inclusion of complete sentences around target expressions instead of concordances.

In the same chapter (Chapter 3), we proposed an approach for disambiguating between literal and idiomatic usages of verb-noun MWEs. The study has been done in a traditional classification scenario, where for each

data entry, features were extracted and classification was performed independently of other entries. A possible future direction is to study the effectiveness of tagging methodologies and investigate ways to improve them. In processing expression types whose token occurrences tend to occur invariably as either idiomatic or literal, tagging methodologies might also suffer from bias to majority class. In order for such a study to be conducted, one needs to make sure that a target tagged corpus contains enough instances of expressions with both occurrences of literal or idiomatic.

In Chapter 4, we discussed modelling and evaluation of MWEs in context. We proposed type-aware train and test splitting and evaluation as a supplementary approach for evaluating MWE identification. The purpose was to effectively assess the generalisation power of a classification approach. One area of future work that is of interest to us, is to investigate the utility of type-aware cross-validation in parameter optimisation. We devised this study and the experiments for a classification scenario. Type-aware train and test splitting (i.e. categorising expression types) would be more challenging in a tagging scenario. Since promising tagged datasets have been introduced recently, we suggest such a study be performed on MWE tagged corpora.

In Chapter 5, we proposed a deep neural network approach to tag corpora for MWEs. However, deep learning models are still under-explored. We expected to see higher improvements by adding CRF to our ConvNet+LSTM model. The effective integration of CRF layers into RNN and CNN architectures should be more investigated in near future. We used pre-trained

word embeddings as input to our networks. Many embedding methodologies, word, subword or character based, have been developed. It would, however, be very interesting to devise an embedding approach that can represent MWEs, rather than single words. NLP researchers would benefit from flexible MWE embeddings.

In order to perform structured prediction we slightly modified the labelling format of the data to make it more similar to IOB format in which the beginning part of an expression is tagged differently from other components. We suggest introduction of labelling formats that differentiate between intervening words that occur in-between gappy MWEs and other outside non-MWE tokens. Schneider et al. (2014a) made the first attempt towards devising such a format. We plan to re-run our experiments using their formatting method. However, the approach does not consider tokens that might belong to two different categories of MWEs, hence we suggest more investigations on labelling formats.

Since the MWE-tagged running texts are fairly small in many languages, one interesting future direction is to apply the model cross-lingually (and language independently) on all tagged corpora and investigate the results on blind test data.

In chapter 6, we proposed a new bilingual distributional similarity approach to extract translations for verb-noun MWEs. We used an extended version of the standard word2vec with bilingual contexts to construct vectors for verb-noun expressions. The vectors defined on the same bilingual context

space can then be compared to each other. One limitation of this approach is that it uses a loosely aligned list of nouns from the two languages. We suggest further research in order to devise an approach that can benefit from comparable corpora to build this bilingual context independently.

We have performed our experiments with continuous verb-noun expressions which consist of only two words. The task is more challenging if we consider discontinuous MWEs. To tackle this task in such cases, we suggest incorporating dependency parsing information that can establish some links between components of an expression. More extensive studies on dependency-based embeddings (Levy and Goldberg, 2014) would be helpful in this regard.

Bibliography

Aker, A., E. Kanoulas, and R. Gaizauskas (2012). A light way to collect comparable corpora from the web. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Al Saied, H., M. Constant, and M. Candito (2017). The ATILF-LLF system for Parseme shared task: a transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain, pp. 127–132. Association for Computational Linguistics.

Ali, A., S. M. Shamsuddin, and A. L. Ralescu (2015). Classification with class imbalance problem: a review. *International Journal of Advances in Soft Computing and its Application 7*(3), 176–204.

Artstein, R. and M. Poesio (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics 34*(4), 555–596.

Baldwin, T. (2005). Deep lexical acquisition of verb-particle constructions. *Computer Speech and Language 19*(4), 398–414.

Baldwin, T., C. Bannard, T. Tanaka, and D. Widdows (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and*

treatment-Volume 18, pp. 89–96. Association for Computational Linguistics.

Baldwin, T., J. Beavers, E. Bender, D. Flickinger, A. Kim, and S. Oepen (2005). Beauty and the beast: What running a broad-coverage precision grammar over the bnc taught us about the grammar—and the corpus. *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, 49–70.

Baldwin, T. and S. N. Kim (2010). Multiword expressions. In *Handbook of Natural Language Processing, second edition.*, pp. 267–292. CRC Press.

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pp. 1–8. Association for Computational Linguistics.

Bannard, C., T. Baldwin, and A. Lascarides (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pp. 65–72. Association for Computational Linguistics.

Baroni, M. and A. Kilgarriff (2006). Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics:*

Demonstrations, EACL '06, Stroudsburg, PA, USA, pp. 87–90. Association for Computational Linguistics.

Bateni, M. R. (1989). Farsi zabāni aghim?'(persian a sterile language?).

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin (2003). A neural probabilistic language model. *Journal of machine learning research 3*(Feb), 1137–1155.

Biemann, C. and M. Riedl (2013). Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling 1*(1), 55–95.

Birke, J. and A. Sarkar (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-06*, pp. 329–336.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016). Enriching word vectors with subword information. *CoRR abs/1607.04606*.

Bouamor, D., N. Semmar, and P. Zweigenbaum (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Bouamor, D., N. Semmar, and P. Zweigenbaum (2013). Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, pp. 759–764. Association for Computational Linguistics.

Brooke, J., V. Tsang, G. Hirst, and F. Shein (2014). Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 753–761.

Candito, M. and M. Constant (2014). Strategies for contiguous multiword expression analysis and dependency parsing. In *ACL 14-The 52nd Annual Meeting of the Association for Computational Linguistics*. ACL.

Carreras, X., K. C. Litkowski, and S. Stevenson (2008). Semantic role labeling: An introduction to the special issue. *Computational Linguistics 34*(2).

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Church, K. W. and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics 16*(1), 22–29.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

Collobert, R. and J. Weston (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Pro-*

ceedings of the 25th International Conference on Machine Learning, ICML '08, New York, USA, pp. 160–167. ACM.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research 12*, 2493–2537.

Constant, M., M. Candito, and D. Seddah (2013). The LIGM-Alpage Architecture for the SPMRL 2013 Shared Task: Multiword Expression Analysis and Dependency Parsing. In *Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, Seattle, United States, pp. 46–52.

Constant, M., G. Eryiğit, J. Monti, L. van der Plas, C. Ramisch, M. Rosner, and A. Todirascu (2017). Multiword expression processing: A survey. *Computational Linguistics 43*(4), 837–892.

Constant, M. and J. Nivre (2016). A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, pp. 161–171. Association for Computational Linguistics.

Constant, M., A. Sigogne, and P. Watrin (2012). Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jeju Island, Korea, pp. 204–212. Association for Computational Linguistics.

Constant, M. and I. Tellier (2012). Evaluating the impact of external lexical resources into a CRF-based multiword segmenter and part-of-speech tagger. In *8th International Conference on Language Resources and Evaluation (LREC'12)*, pp. 646–650.

Cook, P., A. Fazly, and S. Stevenson (2008). The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 19–22.

Cordeiro, S., C. Ramisch, and A. Villavicencio (2016). Ufrgs&lif at semeval-2016 task 10: Rule-based mwe identification and predominant-supersense tagging. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, pp. 910–917. Association for Computational Linguistics.

Corpas Pastor, G. (2017). Collocations in e-bilingual dictionaries: from underlying theoretical assumptions to practical lexicography and translation issues. In *Collocations and other lexical combinations in Spanish. Theoretical and Applied approaches (S. Torner, E. Bernal (ed.))*., Routledge, Abingdon, pp. 173–199.

Corpas Pastor, G., R. Mitkov, N. Afzal, and V. Pekar (2008). Translation universals: do they exist? a corpus-based nlp study of convergence and simplification. In *Proceedings of the AMTA'2008 conference*, pp. 75–81.

Daille, B., S. Dufour-Kowalski, and E. Morin (2004). French-english multi-

word term alignment based on lexical context analysis. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Daille, B. and E. Morin (2005). French-english terminology extraction from comparable corpora. In *International Conference on Natural Language Processing*, pp. 707–718. Springer.

de Caseli, H. M., C. Ramisch, M. d. G. V. Nunes, and A. Villavicencio (2010). Alignment-based extraction of multiword expressions. *Language resources and evaluation 44*(1-2), 59–77.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American society for information science 41*(6), 391.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology 26*, 297–302.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS 19*(1), 61–74.

Evert, S. (2005). *The statistics of word cooccurrences : word pairs and collocations*. Ph. D. thesis, Universitt Stuttgart, Holzgartenstr. 16, 70174 Stuttgart.

Evert, S. (2008). Corpora and collocations. In *Corpus Linguistics. An International Handbook*, Volume 2, pp. 1212–1248.

Evert, S. and B. Krenn (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language 19*(4), 450–466.

Farahmand, M. and R. Martins (2014). A supervised model for extraction of multiword expressions, based on statistical context features. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014)*, Gothenburg, Sweden, pp. 10–16. Association for Computational Linguistics.

Farahmand, M., A. Smith, and J. Nivre (2015). A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, Denver, Colorado, pp. 29–33. Association for Computational Linguistics.

Fazly, A. (2007). *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph. D. thesis, Department of Computer Science, University of Toronto.

Fazly, A., P. Cook, and S. Stevenson (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics 35*(1), 61–103.

Fazly, A. and S. Stevenson (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceed-*

ings of the Workshop on a Broader Perspective on Multiword Expressions, MWE '07, Stroudsburg, PA, USA, pp. 9–16. Association for Computational Linguistics.

Fazly, A. and S. Stevenson (2008). A distributional account of the semantics of multiword expressions. *Italian journal of linguistics 20*(1), 157–180.

Fellbaum, C. (Ed.) (1998). *WordNet: an electronic lexical database*. MIT Press.

Firth, J. (1957). *A Synopsis of Linguistic Theory, 1930-1955*.

Fothergill, R. and T. Baldwin (2012). Combining resources for MWE-token classification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, Stroudsburg, PA, USA, pp. 100–104. Association for Computational Linguistics.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Fung, P. (1997). Finding terminology translation from non-parallel corpora. In *Proceeding of the 5th Workshop on Very Large Corpora, 1997*, pp. 192–202.

Fung, P. (1998). A statistical view on bilingual lexicon extraction: from

parallel corpora to non-parallel corpora. In *Machine Translation and the Information Soup*, pp. 1–17. Springer Berlin Heidelberg.

Gaussier, E., J.-M. Renders, I. Matveeva, C. Goutte, and H. Déjean (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 526. Association for Computational Linguistics.

Gelbukh, A. and O. Kolesnikova (2010). Supervised learning for semantic classification of spanish collocations. In J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and J. Kittler (Eds.), *Advances in Pattern Recognition*, Berlin, Heidelberg, pp. 362–371. Springer Berlin Heidelberg.

Gharbieh, W., V. Bhavsar, and P. Cook (2016). A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions, MWE@ACL 2016, Berlin, Germany, August 11, 2016.*

Gharbieh, W., V. Bhavsar, and P. Cook (2017). Deep learning models for multiword expression identification. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, Vancouver, Canada., pp. 54–64. Association for Computational Linguistics.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies 10*(1), 1–309.

Granger, S. and F. Meunier (2008). *Phraseology: an interdisciplinary perspective*. John Benjamins Publishing Company.

Green, S., N. Andrews, M. R. Gormley, M. Dredze, and C. D. Manning (2011). Cross-lingual coreference resolution: A new task for multilingual comparable corpora.

Green, S., M.-C. de Marneffe, and C. D. Manning (2013). Parsing models for identifying multiword expressions. *Computational Linguistics 39*(1), 195–227.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Hashimoto, C. and D. Kawahara (2008). Construction of an idiom corpus and its application to idiom identification based on wsd incorporating idiom-specific features. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 992–1001. Association for Computational Linguistics.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural Computation 9*(8), 1735–1780.

Honnibal, M. and M. Johnson (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 1373–1378. Association for Computational Linguistics.

Hu, B., Z. Lu, H. Li, and Q. Chen (2014). Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pp. 2042–2050.

Ion, R. (2012). Pexacc: A parallel sentence mining algorithm from comparable corpora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Ismail, A. and S. Manandhar (2010). Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 481–489. Association for Computational Linguistics.

Jackendoff, R. (1997). *The architecture of the language faculty*. Number 28. MIT Press.

Jagarlamudi, J. and H. Daumé (2010). Extracting multilingual topics from unaligned comparable corpora. In *European Conference on Information Retrieval*, pp. 444–456. Springer.

Jansche, M. (2002). Named entity extraction with conditional markov models and classifiers. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pp. 1–4. Association for Computational Linguistics.

Katz, G. and E. Giesbrecht (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and*

Exploiting Underlying Properties, MWE '06, Stroudsburg, PA, USA, pp. 12–19. Association for Computational Linguistics.

Kilgarriff, A., P. Rychly, P. Smrz, and D. Tugwell (2004). The Sketch Engine. In *EURALEX 2004*, Lorient, France, pp. 105–116.

Kim, S. N. and T. Baldwin (2005). Automatic interpretation of noun compounds using wordnet similarity. In R. Dale, K.-F. Wong, J. Su, and O. Y. Kwong (Eds.), *Natural Language Processing – IJCNLP 2005*, Berlin, Heidelberg, pp. 945–956. Springer Berlin Heidelberg.

Kim, S. N. and T. Baldwin (2006). Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, Stroudsburg, PA, USA, pp. 491–498. Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751.

Klementiev, A. and D. Roth (2006). Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 817–824. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, Phuket, Thailand, pp. 79–86.

Kordoni, V. (2017). Beyond words: Deep learning for multiword expressions and collocations. In *Proceedings of ACL 2017, Tutorial Abstracts*, pp. 15–16. Association for Computational Linguistics.

Kordoni, V., C. Ramisch, and A. Villavicencio (2011). Proceedings of the ACL workshop on multiword expressions: from parsing and generation to the real world (MWE 2011). Association for Computational Linguistics.

Krenn, B. and S. Evert (2001). Can we do better than frequency? a case study on extracting pp-verb collocations. *Proceedings of the ACL Workshop on Collocations*, 39–46.

Lafferty, J. D., A. McCallum, and F. C. N. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, San Francisco, CA, USA, pp. 282–289. Morgan Kaufmann Publishers Inc.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer (2016). Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pp. 260–270.

Le Roux, J., A. Rozenknop, and M. Constant (2014). Syntactic parsing

and compound recognition via dual decomposition: application to french. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1875–1885.

Lebret, R. and R. Collobert (2014). Word embeddings through hellinger pca. *EACL 2014*, 482.

Legrand, J. and R. Collobert (2016). Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016)*, Berlin, Germany.

Levy, O. and Y. Goldberg (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, pp. 302–308. Association for Computational Linguistics.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 317–324. Association for Computational Linguistics.

Losnegaard, G. S., F. Sangati, C. P. Escartn, A. Savary, S. Bargmann, and J. Monti (2016). Parseme survey on mwe resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Ma, X. and E. Hovy (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1064–1074. Association for Computational Linguistics.

Maldonado, A., L. Han, E. Moreau, A. Alsulaimani, K. D. Chowdhury, C. Vogel, and Q. Liu (2017). Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions* , MWE '17, pp. 114–120. Association for Computational Linguistics.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to information retrieval*. Cambridge University Press.

Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press.

McCallum, A. and W. Li (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 188–191. Association for Computational Linguistics.

McCarthy, D., B. Keller, and J. Carroll (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* -

Volume 18, MWE '03, pp. 73–80. Association for Computational Linguistics.

McEnery, A. and R. Xiao (2007). Parallel and comparable corpora: What is happening. *Incorporating Corpora. The Linguist and the Translator*, 18–31.

Mendoza Rivera, O., R. Mitkov, and G. Corpas Pastor (2013). A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In *Workshop on Multi-word Units in Machine Translation and Translation Technology*, Nice, France.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*.

Mikolov, T., Q. V. Le, and I. Sutskever (2013b). Exploiting similarities among languages for machine translation. *CoRR abs/1309.4168*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013c). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Mitkov, R., J. Monti, G. Corpas Pastor, and V. Seretan (2018). *Multiword Units in Machine Translation and Translation Technology*, Volume 341. John Benjamins Publishing Company.

Mitkov, R., V. Pekar, D. Blagoev, and A. Mulloni (2007). Methods for

extracting and classifying pairs of cognates and false friends. *Machine translation 21*(1), 29.

Mitkov, R. and S. Taslimipoor (n.d.). What matters more: the size of the corpora or their quality? the case of automatic translation of multiword expressions using comparable corpora. In G. Corpas and C. J. P. (Eds.), *Computational Phraseology*. John Benjamins.

Moirón, B. V. and J. Tiedemann (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*.

Monti, J., V. Seretan, G. Corpas Pastor, and R. Mitkov (2018). Multiword units in machine translation and translation technology. In *Multiword Units in Machine Translation and Translation Technology*. John Benjamins Publishing Company.

Morin, E. and B. Daille (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation 44*(1-2), 79–95.

Munteanu, D. S. and D. Marcu (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics 31*(4), 477–504.

Nasr, A., C. Ramisch, J. Deulofeu, and A. Valli (2015). Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the*

7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, pp. 1116–1126. Association for Computational Linguistics.

Nivre, J. (2004). Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, pp. 50–57. Association for Computational Linguistics.

Nivre, J. and J. Nilsson (2004). Multiword units in syntactic parsing. In *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA 2004)*, pp. 39–46.

Okazaki, N. (2007). Crfsuite: a fast implementation of conditional random fields (crfs).

Pal, S., T. Chakraborty, and S. Bandyopadhyay (2011). Handling multiword expressions in phrase-based statistical machine translation. In *Proceedings of the 13th Machine Translation Summit*, pp. 215–224. MT Summit 2011.

Pal, S., P. Pakray, and S. K. Naskar (2014). Automatic building and using parallel resources for smt from comparable corpora. *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, 48–57.

Pascanu, R., T. Mikolov, and Y. Bengio (2012). Understanding the exploding gradient problem. *CoRR abs/1211.5063*.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, Stroudsburg, PA, USA, pp. 13–18. Association for Computational Linguistics.

Pecina, P. (2008). A machine learning approach to multiword expression extraction. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, 54–61.

Pekar, V., R. Mitkov, D. Blagoev, and A. Mulloni (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation 20*(4), 247–266.

Peng, J., A. Feldman, and E. Vylomova (2014). Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 2019–2027. Association for Computational Linguistics.

Pennington, J., R. Socher, and C. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Pinnis, M., R. Ion, D. Ştefănescu, F. Su, I. Skadiņa, A. Vasiļjevs, and B. Babych (2012). Accurat toolkit for multi-level alignment and information extraction from comparable corpora. In *Proceedings of the ACL*

2012 System Demonstrations, ACL '12, Stroudsburg, PA, USA, pp. 91–96. Association for Computational Linguistics.

Qu, L., G. Ferraro, L. Zhou, W. Hou, N. Schneider, and T. Baldwin (2015). Big data small data, in domain out-of domain, known word unknown word: The impact of word representations on sequence labelling tasks. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL)*, pp. 83–93.

Ramisch, C. (2014). *Multiword expressions acquisition: A generic and open framework*. Springer.

Ramisch, C., L. Besacier, and A. Kobzar (2013a). How hard is it to automatically translate phrasal verbs from english to french? In *MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technology*.

Ramisch, C. and A. Villavicencio (2018). Computational treatment of multiword expressions. In R. Mitkov (Ed.), *Oxford Handbook of Computational Linguistics* (2nd ed.). Oxford University Press.

Ramisch, C., A. Villavicencio, and C. Boitet (2010). mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta.

Ramisch, C., A. Villavicencio, and V. Kordoni (2013b). Introduction to the

special issue on multiword expressions: From theory to practice and use. *ACM Trans. Speech Lang. Process. 10*(2), 3:1–3:10.

Ramshaw, L. A. and M. P. Marcus (1999). *Text Chunking Using Transformation-Based Learning*, pp. 157–176. Dordrecht: Springer Netherlands.

Rapp, R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 519–526. Association for Computational Linguistics.

Rapp, R. and S. Sharoff (2014). Extracting multiword translations from aligned comparable documents. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014*, Gothenburg, Sweden, pp. 83–91.

Rapp, R., S. Sharoff, and P. Zweigenbaum (2016). Recent advances in machine translation using comparable corpora. *Natural Language Engineering 22*(4), 501–516.

Řehůřek, R. and P. Sojka (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50. ELRA.

Rei, M. and H. Yannakoudakis (2016). Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Volume 1, pp. 1181–1191.

Ren, Z., Y. Lü, J. Cao, Q. Liu, and Y. Huang (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, Stroudsburg, PA, USA, pp. 47–54. Association for Computational Linguistics.

Rondon, A., H. Caseli, and C. Ramisch (2015). Never-ending multiword expressions learning. In *Proceedings of the 11th Workshop on Multiword Expressions*, Denver, Colorado, pp. 45–53. Association for Computational Linguistics.

Sag, I. A., T. Baldwin, F. Bond, A. A. Copestake, and D. Flickinger (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, London, UK, pp. 1–15. Springer-Verlag.

Salehi, B. and P. Cook (2013). Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, Atlanta, Georgia, USA, pp. 266–275. Association for Computational Linguistics.

Salehi, B., P. Cook, and T. Baldwin (2014). Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 472–481.

Salehi, B., P. Cook, and T. Baldwin (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, pp. 977–983. Association for Computational Linguistics.

Salton, G., R. Ross, and J. Kelleher (2016). Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 194–204. Association for Computational Linguistics.

Samek-Lodovici, V. (2003). The internal structure of arguments and its role in complex predicate formation. *Natural Language & Linguistic Theory 21*(4), 835–881.

Saralegui, X., I. San Vicente, and A. Gurrutxaga (2008). Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. In *LREC 2008 workshop on building and using comparable corpora*.

Savary, A., M. Candito, V. Barbu Mititelu, E. Bejček, F. Cap, S. Čéplö,

S. R. Cordeiro, G. Eryigit, V. Giouli, M. van Gompel, Y. HaCohen-Kerner, J. Kovalevskaite, S. Krek, C. Liebes kind, J. Monti, C. Parra Escartín, L. van der Plas, B. QasemiZadeh, C. Ramisch, F.-d.-r.-c. Sangati, I. Stoyanova, and V. Vincze (2018). PARSEME multilingual corpus of verbal multiword expressions. In S. Markantonatou, C. Ramisch, A. Savary, and V. Vincze (Eds.), *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*. Berlin, Germany: Language Science Press.

Savary, A., C. Ramisch, S. Cordeiro, F. Sangati, V. Vincze, B. Qasemizadeh, M. Candito, F. Cap, V. Giouli, I. Stoyanova, et al. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pp. 31–47.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50.

Schneider, N., E. Danchik, C. Dyer, and N. A. Smith (2014a). Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *TACL 2*, 193–206.

Schneider, N., D. Hovy, A. Johannsen, and M. Carpuat (2016). SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM).

In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@ NAACL-HLT)*, pp. 546–559.

Schneider, N., S. Onuffer, N. Kazour, E. Danchik, M. T. Mordowanec, H. Conrad, and N. A. Smith (2014b). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, pp. 455–461.

Scholivet, M. and C. Ramisch (2017). Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain, pp. 167–175. Association for Computational Linguistics.

Schone, P. and D. Jurafsky (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 100–108.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics 24*(1), 97–123.

Sha, F. and F. Pereira (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 134–141. Association for Computational Linguistics.

Sharoff, S., P. Zweigenbaum, and R. Rapp (2015). Bucc shared task: Cross-language document similarity. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pp. 74–78.

Skadiņa, I., A. Aker, N. Mastropavlos, F. Su, D. Tufis, M. Verlic, A. Vasiļjevs, B. Babych, P. Clough, R. Gaizauskas, N. Glaros, M. L. Paramita, and M. Pinnis (2012). Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics 19*, 143–177.

Smith, J. R., C. Quirk, and K. Toutanova (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 403–411. Association for Computational Linguistics.

Stevenson, S., A. Fazly, and R. North (2004). Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, Stroudsburg, PA, USA, pp. 1–8. Association for Computational Linguistics.

Su, F. and B. Babych (2012). Measuring comparability of documents in

non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, EACL 2012, Stroudsburg, PA, USA, pp. 10–19. Association for Computational Linguistics.

Taslimipoor, S., A. Desantis, M. Cherchi, R. Mitkov, and J. Monti (2016a). Language resources for Italian: towards the development of a corpus of annotated Italian multiword expressions. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016), Napoli, Italy.*

Taslimipoor, S., R. Mitkov, G. Corpas Pastor, and A. Fazly (2016b). Bilingual contexts from comparable corpora to mine for translations of collocations. In *The 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, pp. 115–126. Springer.

Taslimipoor, S., O. Rohanian, R. Mitkov, and A. Fazly (2017). Investigating the opacity of verb-noun multiword expression usages in context. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain, pp. 133–138. Association for Computational Linguistics.

Taslimipoor, S., O. Rohanian, R. Mitkov, and A. Fazly (2018). Identification of multiword expressions: A fresh look at modelling and evaluation. In

S. Markantonatou, C. Ramisch, A. Savary, and V. Vincze (Eds.), *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*, 299–317. Berlin, Germany: Language Science Press.

Tiedemann, J. (1998). Extraction of translation equivalents from parallel corpora. *Proceedings of the 11th Nordic conference on computational linguistics*, 120–128.

Tjong Kim Sang, E. F. (2000). Text chunking by system combination. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, Stroudsburg, PA, USA, pp. 151–153. Association for Computational Linguistics.

Tsvetkov, Y. and S. Wintner (2014). Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics 40*(2), 449–468.

Tu, Y. and D. Roth (2011). Learning english light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, Portland, Oregon, USA, pp. 31–39. Association for Computational Linguistics.

Turian, J., L. Ratinov, and Y. Bengio (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the*

48th annual meeting of the association for computational linguistics, pp. 384–394. Association for Computational Linguistics.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research 37*, 141–188.

Uchiyama, K., T. Baldwin, and S. Ishizaki (2005). Disambiguating japanese compound verbs. *Computer Speech & Language 19*(4), 497–512.

Udupa, R., K. Saravanan, A. Kumaran, and J. Jagarlamudi (2008). Mining named entity transliteration equivalents from comparable corpora. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 1423–1424. ACM.

Villavicencio, A. (2005). The availability of verb–particle constructions in lexical resources: How much is enough? *Computer Speech & Language 19*(4), 415–432.

Vulić, I., D. Kiela, S. Clark, and M.-F. Moens (2016). Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 188–194. ACL.

Vulić, I. and M.-F. Moens (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for*

Computational Linguistics, pp. 449–459. Association for Computational Linguistics.

Vulić, I. and M.-F. Moens (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. 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)*, Volume 2, pp. 719–725.

Wanner, L. (2004). Towards automatic fine-grained semantic classification of verb-noun collocations. *Natural Language Engineering 10*(2), 95–143.

Wynne, M. (2008). Searching and concordancing. In *Handbook of Corpus Linguistics*, pp. 706–737. De Gruyter Mouton.

Yu, D., S. Wang, and L. Deng (2010). Sequential labeling using deep-structured conditional random fields. *IEEE Journal of Selected Topics in Signal Processing 4*(6), 965–973.

Zhang, M., H. Peng, Y. Liu, H.-B. Luan, and M. Sun (2017). Bilingual lexicon induction from non-parallel data with minimal supervision. In *AAAI*.

Appendix A

Annotation Instructions for Verb-Noun MWEs In Context

This exercise consists of annotating Multiword Expressions (MWEs) to prepare a gold-standard for evaluating a system which uses Natural Language Processing methodologies for automatic identification of MWEs. For the time being, the study is restricted to verb-noun expressions only and the focused verbs are mostly light verbs (e.g. *take* and *make*) in three languages, English, Spanish and Italian. The Corpora are tagged with Part-of-Speech which were used to extract verb-noun expressions.

Multiword Expression Definition and Annotation Criteria

For the purpose of this task, we shall annotate expressions as MWEs only if they exhibit a sufficient degree of idiomaticity. In other words, they are expressions which do not convey literal meanings. There may be two possible cases of idiomaticity:

- The verb is delexicalised. For example, the expressions *take a decision*, *tomar decisiones* or *prendere una decisione* will be tagged as MWEs as in such constructions the light verb *take* (*tomar* or *prendere*) do

not carry literal meaning and rather refer to the process of arriving at a specific decision (possibly after deliberation). Also, the expressions *take breaks* and *hacer descansos*, are MWEs. In these cases, the verbs are used as light verbs and not in their literal senses. These Light Verb Constructions (LVCs) are MWEs.

- The noun is also delexicalised. The expression is thoroughly idiomatic and the meaning of the whole expression does not have anything to do with the meaning of the components: e.g. *take place* in English or *tener lugar* in Spanish.

Contrary to these, expressions like, *give a present*, *have a coffee*, *tomar café* or *prendere un caffé*, *tener libros*, *dar dinero* will not be marked as MWEs because they have fully transparent meanings and can be interpreted literally.

The general rule will therefore be that for an expression to be an MWE, the meaning of the expression cannot be solely predicted from the simple composition of the meanings of the verb and the noun that form it. That is, the meaning is not fully compositional. Expressions will be annotated as MWEs if they feature any of the following properties.

- MWEs are not fully compositional (Idioms are largely non-compositional; LVCs are semi-compositional): e.g. the meaning of the expression *take place* cannot be inferred from the meaning of the components *take* and *place*.

- MWEs are restricted with respect to the syntactic forms they appear in: e.g. *gave a groan* but not *a groan was given.*

- MWEs have some degree of lexical restrictiveness: e.g. *shoot the wind* does not have an idiomatic meaning like *shoot the breeze*, even though *wind* and *breeze* are semantically similar

- MWEs are not fully productive: e.g., we do not say *give a gripe*, while we can say *give a groan/cry/yell.*

Verb-noun MWEs include Idioms and LVCs. In LVCs, the verb component does not contribute much of its 'basic' meaning – e.g., in *give a groan*, *give* does not mean 'transfer of possession'. LVCs differ from idioms in that they are semantically more transparent because of a strong semantic connection to the noun constituent – e.g., *give a groan* can be roughly paraphrased by *groan.* LVCs are semi-compositional since their meaning can be mainly predicted from the noun constituent.

Having idiomatic expressions and LVCs as MWEs on one side and literal combinations as non-MWEs on the other side, there are also expressions with in-between levels of semantic transparency, such as *give confidence* (referring to an abstract transfer, as opposed to a physical transfer). These cases in which meaning does not denote a physical action, but there is semantic information linked to the verb, will not be marked as MWEs. This is the case of *give confidence/dar confianza*, which does not denote a physical action of 'giving' something, but it is one of the meanings of *give/dar* (metaphorical

or abstract meaning). The same happens with *have ideas*/*tener ideas* or *take a meaning*, *put prices*, or *tomar café* which is metaphorically connected to the prototypical meaning of the verb.

Annotation Format

In this annotation phase, the task consists of annotating a list of concordances including a marked verb and a following noun. Each concordance is going to be annotated targeting the usage of the containing Verb+Noun expression.

There are TWO possible tags for any given verb-noun combination in its concordance: 0 and 1.

1 We will annotate with "0" all usages which are not MWEs.

2 We will annotate with "1" all usages which we believe that are MWEs.

Important Remarks

- Annotators shall judge the expressions only in their consequent forms in the extracted context/concordance. That is, the noun follows the verb with no other element appearing in between. If, for instance, a verb-noun combination is not an MWE but it would be one, if an element (such as a determiner) appeared between them, we will annotate this expression as "0".

- If a verb-noun expression could never occur without a following preposition (e.g. *give rise* could never occur without 'to' after it or *formar*

parte without 'de'), annotate such expression with "1" and then add 'PP' in the adjacent column.

- If a verb-noun expression would only be an MWE when it occurs with another word (other than a preposition or particle)[1] that makes it a three/four/more-word expression (e.g. *rain cats and dogs*), then annotate it with "0" and add "INC" in the adjacent column.

- Sometimes there are words which were misspelt in the corpus. If this is the only issue, annotate the expression as "1" and add a remark about the right spelling in the comments column.

- If the verb-noun expression conveys a literal meaning but with the abstract sense of the verb i.e. in *give confidence / dar confianza*, please annotate it with "0" and then add "ABS" in the comments column. Please consider that some idiomatic expressions might contain an abstractive verb, those are MWEs. Put "0" if they convey literal meaning but with the abstract sense of the verb.

- We know that this is not a trivial task and that some issues may arise during the annotation. If you are unsure about the idiomaticity of an expression, use your instinct as a native speaker. We will use the inter-annotator agreements to determine the cases which shall be revised once the first annotation phase finishes.

[1]This other word might occur with some distance from the verb-noun in the sentence

APPENDIX A. ANNOTATION INSTRUCTIONS FOR VERB-NOUN MWES IN CONTEXT

Table A.1 depicts some examples of verb-noun annotations in English.

Table A.1: Some sample annotations for English.

Concordance	Verb-Noun	Annotation	COMMENT
Pictures. Their president, Gordon Stulberg, < became > president of Twentieth Century-Fox and	become president	0	
you look at page forty one there you've < got > people playing bowls at the top there	get people	0	
project and it is things like that about < giving > people confidence to join arts in a way	give people	0	
importunate - terminates. If she didn't, she'd < have > babies annually from puberty until death	have babies	1	
ends in tears.' 'I'm not trying to < have > babies with her, blast it!' bellows the	have babies	0	
between these two types of behaviour, and < gives > rise to unstable behaviour as the injected	give rise	1	pp
n't made an arrangement to meet, even, she < had > visitors coming for the night, official	have visitors	0	
. I want to stay in football. I have not < lost > confidence in my ability. 'I have experience	lose confidence	0	ABS
you feel guilty. We did nothing wrong. We < made > love , that's all.' All... it was	make love	1	
it! Hurriedly she forced herself to < pay > attention , surreptitiously edging away	pay attention	1	
but you did say on that base oh well you < pay > tax but obviously you must do as well then	pay tax	0	
exactly. went unemployed and then decided to < take > advantage . Yeah I mean we do get people	take advantage	1	
let's not beat about the bush here. We are < taking > responsibility for people who we were not	take responsibility	1	
client, it is not part of the firm's role to < take > decisions for a client. We should	take decisions	1	
their head. C we've, I mean we've, we are < having > discussions , er, with the health authority	have discussions	1	