Automatic acquisition of lexical-semantic relations.

Gathering information in a dense representation.

Silvia Necşulescu

TESI DOCTORAL UPF / ANY 2015

DIRECTORA DE LA TESI

Núria Bel
Departament de Traducció i Ciències del Llenguatge
Institut Universitari de Lingüística Aplicada

upf. Universitat
Pompeu Fabra
Barcelona

Abstract

Lexical-semantic relationships, for instance hyperonymy, meronymy and cohyponymy, between words are key information for many Natural Language Processing tasks, which require this knowledge in the form of lexical resources. The aim of this thesis is to automate the development of these resources, addressing the acquisition of relation instances: given a target semantic relation and a corpus of texts in a language, the system outputs word pairs holding the target relation. State of the art systems rely on word pair representations based on patterns of contexts where two related words co-occur to detect their lexical semantic relation. This approach is hindered by data sparsity because to observe in a corpus related words co-occurring is a pre-condition to detect their lexical semantic relation. Even when mining very large corpora, not every related word pair co-occurs or not frequently enough. Therefore, our main goal was to investigate novel representations to predict relations between words, even when these are never found in the same sentence in a given corpus. Our intuition was that these representations should contain information about patterns of context combined with information about the meaning of words involved in the relation. These two sources of information have to be the basis of a generalization strategy to be able to provide information even for words that do not co-occur. We provide two novel representations that proved to overcome the data sparsity issue as demonstrated by a gain of more than 20 points in recall.

Resum

Les relacions lexicosemàntiques entre paraules, com per exemple la hiperonímia, la meronímia i la cohiponímia, són una informació clau per a moltes tasques del Processament del Llenguatge Natural, que requereixen d'aquest coneixement en forma de recursos lingüístics. L'objectiu d'aquesta tesi és l'automatització del desenvolupament d'aquests recursos, tractant l'adquisició d'instàncies d'aquestes relacions: donada una relació semàntica particular i un corpus de textos en una llengua, el sistema produeix parells de paraules que mantenen aquesta relació semàntica. Els sistemes actuals utilitzen representacions basades en patrons dels contextos en que les dues paraules relacionades coocorren per detectar la relació lexicosemàntica que s'hi estableix. Aquest enfocament s'enfronta a problemes de falta de dades, ja que una precondició per detectar la relació entre elles és trobar coocurrències d'aquestes paraules en el corpus. Fins i tot en el cas de treballar amb corpus de grans dimensions, hi haurà parells de paraules relacionades que no coocorreran, o no amb la freqüéncia necessària. Per tant, el nostre objectiu principal ha estat proposar noves representacions per predir relacions entre paraules, fins i tot quan aquestes no apareixen a la mateixa frase en un corpus en particular. La intuïció era que aquestes representacions noves havien de contenir informació sobre patrons de context, però combinada amb informació sobre el significat de les paraules implicades en la relació. Aquestes dues fonts d'informació havien de ser la base d'una estratègia de generalització que oferís informació fins i tot quan les dues paraules no coocorrien. Així, proposem dues representacions noves que han mostrat resoldre el problema de la manca de dades, com demostra el fet que aconsegueixen augmentar la cobertura en més de 20 punts.

Resumen

Las relaciones léxico-semánticas entre palabras, por ejemplo hiperonimia, meronimia y cohiponimia, son una información clave para muchas tareas del Procesamiento del Lenguaje Natural, que requieren de este conocimiento en forma de recursos lingüísticos. El objetivo de esta tesis es la automatización del desarrollo de estos recursos, tratando la adquisición de instancias de estas relaciones: dada una relación semántica particular y un corpus de textos en una lengua, el sistema produce pares de palabras que mantienen esa relación semántica. Los sistemas actuales utilizan representaciones basadas en patrones de los contextos donde co-ocurren las dos palabras relacionadas para detectar la relación léxico-semántica entre ellas. Este enfoque se enfronta a problemas de falta de datos ya que una precondición para detectar la relación entre ellas es encontrar co-ocurrencias de esas palabras en el corpus. Incluso en el caso de trabajar con corpus de grandes dimensiones, habrá pares de palabras relacionadas que no co-ocurrirán o no con la frecuencia necesaria. Por tanto, nuestro principal objetivo ha sido proponer nuevas representaciones para predecir relaciones entre palabras, incluso cuando éstas no aparecen en la misma frase en un corpus en particular. La intuición era que estas representaciones nuevas debían contener información sobre patrones de contexto pero combinada con información sobre el significado de las palabras implicadas en la relación. Estas dos fuentes de información tenían que ser la base de una estrategia de generalización que ofreciera información incluso cuando las dos palabras no co-ocurrían. Así, proponemos dos representaciones nuevas que han mostrado resolver el problema de la falta de datos, como demuestra el hecho de que consiguen aumentar la cobertura en más de 20 puntos.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

We are experiencing a phenomenon of digitalised information explosion yielding huge amounts of data, which contain meaningful information expressed in human language. To exploit this information, systems must understand the human language. The task of human language understanding (NLU) is part of the Natural Language Processing (NLP) domain and with the Natural Language Generation (NLG) is the final aim of NLP.

The research in the NLP domain is taking small steps closer to NLU by creating task dedicated systems such as: information extraction [Agichtein and Gravano, 2000], information retrieval [Voorhees, 1998], word sense disambiguation [Leacock and Chodorow, 1998, Banerjee and Pedersen, 2002, McCarthy and Carroll, 2003], text classification [Scott and Matwin, 1998], semantic role labeling [Gildea and Jurafsky, 2002], semantic similarity [Lin, 1998a], automatic summarisation [Salton et al., 1997], among many others.

Many systems in NLP require human knowledge in form of lexical resources. These resources were created in form of taxomonies, such as WordNet [Miller, 1995] and its siblings EuroWordNet [Vossen et al., 1997], BalkaNet [Stamou et al., 2002] or CoreNet [Choi and Bae, 2004], EDR [Yokoi, 1995] and ontologies such as Cyc [Lenat et al., 1985], SIM-

PLE [Lenci et al., 2000], SUMO [Pease et al., 2002], ConceptNet [Liu and Singh, 2004], and contain semantic information and world knowledge information describing facts in the world.

WordNet is probably the most wide-spread used lexical resource in NLP. This project was developed at Princeton University and was motivated by the theories of the human semantic memory: humans organise their knowledge of concepts in an economic and hierarchical model. Psychological experiments showed that the retrieval time requested to access conceptual knowledge seems to be directly related with the distance in the hierarchy between the concepts. For instance, speakers are able to quickly verify *canary* - *sings* because *to sing* is recorded in memory as a direct property of a *canary*, but the time is larger to detect *canary* - *flies* because *to fly* is an inherited property from *bird*, a *canary* is a type of *bird*, and the subject has to go up in the hierarchy to the concept bird to retrieve this information. The response time was even larger to detect *canary* has *skin*, because the speaker has to go up in the hierarchy to the ancestor *animal*. Motivated by these observations, WordNet developers aimed to model all the concepts in a semantic memory modelled as a network created based on two key concepts: the relations of *synonymy* and *hypernymy* between concepts.

In WordNet, each word meaning is divided in senses, $word^{N}_{pos}$ referring to the N^{th} sense of the word *word* with the part-of-speech *pos*. Word senses that refer to the same concept are grouped in *synsets*, therefore a *synset* is a group of synonym items. Because of the polysemy of words, a word can have more than one sense. For instance, the Table 1.1 shows the synsets creating the senses of the word *dog*: { $\mathrm{dog}^{1}_{\mathrm{n}}$, domestic $\mathrm{dog}^{1}_{\mathrm{n}}$, Canis $\mathrm{familiaris}^{1}_{\mathrm{n}}$} is the synset representing the first sense of *dog*, referring at *a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds*. Each *synset* represents a concept of the language and therefore, each synset is a node of a network representing the semantic memory.

The network structure is based on the relation of hypernymy, connecting more specific concepts, *hyponyms*, to more general concepts, *hypernyms*. For instance, the synset *domestic_animal*$^{1}_{n}$ is a hypernym of the synset *dog*$^{1}_{n}$. The relation is bidirectional, with *dog*$^{1}_{n}$ being the hyponym of *domestic_animal*$^{1}_{n}$. Additionally, concepts are also related through the relation of *meronymy*, or *part-whole*, linking a *whole* synset to its *parts*. For instance, the *dog*$^{1}_{n}$ (whole) has a relation of *meronymy* with *flag*$^{7}_{n}$, *a conspicuously marked or shaped tail*, (part). The opposite relation is *holonymy*. Other relations in WordNet, but with fewer instances, are *antonymy* linking adjectives, *toponymy* linking verbs, *morpho-semantic* links linking words sharing a stem with the same meaning and semantic roles.

SENSE	SYNSET	GLOSS
dog^1_n	dog^1_n, $domesticdog^1_n$, $Canisfamiliaris^1_n$	(a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds)
dog^2_n	$frump^1_n$, dog^2_n	(a dull unattractive unpleasant girl or woman)
dog^3_n	dog^3_n	(informal term for a man)
dog^4_n	cad^1_n, $bounder^1_n$, $blackguard^1_n$, dog^4_n, $hound^2_n$, $heel^3_n$	(someone who is morally reprehensible)
dog^5_n	$frank^2_n$, $frankfurter^1_n$, $hotdog^3_n$, $hotdog^3_n$, dog^5_2, $wiener^2_n$, $wienerwurst^1_n$, $weenie^1_n$	a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll
dog^6_n	$pawl^1_n$, $detent^1_n$, $click^3_n$, dog^6_n	(a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
dog^7_n	$andiron^1_n$, $firedog^1_n$, dog^7_n, $dog - iron^1_n$	(metal supports for logs in a fireplace)

Table 1.1: WordNet information for dog_n in WordNet 3.1

Resources like WordNet typically contain words of the general domain. The systems that rely on WordNet information, and address NLP tasks in specific domains, need WordNet-like resource covering the target domain to achieve good results. Moreover, WordNet is language dependent and it has to be translated for each target language a system wants to cover. English WordNet development (currently with about 117,000 synsets) took almost 30 years and it still misses a huge amount of concepts and relation links. Therefore, the manual development of lexical-semantic resources for each language (note that Spanish WordNet only contains about 59,000 synsets) is very costly and very time-consuming [Agirre et al., 2012].

The aim of this thesis is to help the automatic development of these lexical resources, addressing the acquisition of relation instances: given a semantic relation, and a corpus of texts in a language, the system outputs word pairs holding the target relation. Automation of language resource production will contribute to solve the language and domain coverage problem that current systems face in real world applications.

1.1 Automatic Acquisition of Lexical-Semantic Relations

Over the years, many lexical resources were developed containing different semantic relation types, depending on the final target of the resource. For instance WordNet contains relations that describe concepts, like a *dog* is an *animal*, while Freebase [Bollacker et al., 2008], BabelNet [Navigli and Ponzetto, 2010] or Yago [Suchanek et al., 2008] contain relations created by experiences in the human life, such as *Eliade* ***was born*** *in Bucharest* or *Eliade* ***has occupation*** *author*. We classify the semantic relations holding between a word pair in two main classes depending on the type of terms linked: (1) relations that correspond to aspects of the meaning of lexical items, which represent the cognitive organisation of the human language, such as the WordNet relations; and (2) relations that represent world knowledge, created by external factors other than a human language, for instance the relations acquired from Wikipedia, to populate Freebase, BabelNet or Yago.

This present work focuses on the automatic acquisition of relations of the former type, relations that emerge from the semantic properties of the lexical items, and which are usually named **lexical-semantic relations** [Cruse, 1986]. The relations that belong to the lexical-semantic relation class are: synonyms, such as *animal* and *animate being*, hypernyms, such as *dog* and *animal*, meronyms such as *dog* and *tail*, co-hyponymy, such as *dog* and *cat*, selectional preferences such as *dog* and *bark* or *dog* and *furry*, among others (see Section 2.2).

The automatic acquisition of lexical-semantic relations, and consequently the automation of the development of lexical resources, is possible thanks to the availability of large amounts of digitalised texts. However, to use all these texts, the information contained in human language has to be represented in a computer understanding format. The meaning of words is represented thanks to the formulation of the *distributional hypothesis* [Harris, 1954]: the main tool to represent the meaning of a word is the set of words that may co-occur with it in the same context (see Section 2.1). The representation of the semantic relations between words has been influenced by Hearst's pioneering work [Hearst, 1992]. [Hearst, 1992] proposed an approach based on manually developed patterns of contexts that are indicative of semantic relations among word pairs, for instance **is a** for hypernymy cases, such as *cat **is an** animal*. Therefore, the distributional hypothesis derived in the *latent relational hypothesis*: the information provided by the contexts where two words co-occur assesses their relation. These two hypothesis are fundamental for our acquisition task, because they are the main pillars for semantic processing, allowing the representation of word meaning and of lexical-semantic relations (see Section 2.3).

Hearst's approach has an important drawback: patterns have to be developed for each relation and translated for each language. This approach can be extended, for instance in Section 5.1, we describe an automatic methodology based on a set of example instances for each target relation, which are used to mine reliable contexts indicative of that relation. These contexts are transformed into patterns by generalising over particular lexical items, and are then used to discover new word pair instances in the text corpus. This method also has an important drawback, though. Word pairs that do not occur both in the same sentence, are not considered as candidates of holding a semantic relation, because the patterns only capture co-occurring word pairs. Therefore, the acquisition capacity is upper-bound limited to the number of those actually co-occurring in the corpus. That is, word pairs that are instances of a lexical-semantic relation, but are not observed in the same sentence in the input corpus will never be acquired. Therefore, the main issue of these systems is that a word pair has to co-occur a sufficient amount of times to provide enough information for successful classification. Mainstream pattern-based systems are unable to associate any feature to word pairs that do not co-occur and, consequently, to classify them. To solve that limitation has been the central motivation of this work.

To enlarge the possibility of finding more co-occurring word pairs, initial approaches enlarge the size of the input corpus. However, more corpus is not always available (see Section 2.4) and, moreover, in any corpus of any size, there would be more words which might be semantically related, but that still do not co-occur and that will never be properly captured. The key point of our work is to what extent other information coming from word occurrences in the corpus could be used to breakthrough the co-occurrence limitation.

To solve the problem, our proposal relies on the assumption that lexical-semantic relations are strongly related with other distributional properties of the words holding the relation:

> *Because meaningful sentences are composed of meaningful words, any system that hopes to process natural languages as people do must have information about words and their meanings. The semantic relations into which a word enters determine the definition of that word [Cruse, 1986].*

These two information types have already been combined in an attempt to gain recall, but they were taken as two different, unrelated data, generating two different scores: the lexical similarity score between words and relational similarity score between word pairs, that are combined in a final score [Snow et al., 2004, Turney, 2008a, Ó Séaghdha and Copestake, 2009]. The results obtained by this approach did not prove any significant improvement in accuracy (see Section 3).

In this thesis we present novel approaches following the intuition that, indeed, the context where word pairs co-occur is highly important and gives the most valuable information to determine the semantic relation type. But to overcome the upper-bound limitation, word pair representations have to include also information about the semantic properties of their members, such as their distributional properties. The novelty of our approach is that these two sources of information are related and they have to be combined through a generalisation technique over the information in corpus. In this way, more information is squeezed from corpus, and consequently the dimensions of the input corpus are no longer a crucial issue.

We approach the task of lexical-semantic instances acquisition as a multi-classification task. A classifier is trained with labelled word pairs, representing positive and negative instances for a set of target relations, to learn how to determine the lexical-semantic relation holding between a novel word pair, if any.

1.2 Objectives

This thesis addresses the acquisition of lexical-semantic relation instances and our main goal is to overcome data sparsity, which hinders acquisition approaches based on the latent relational hypothesis and patterns of contexts. The objectives of the thesis are the following:

- To show the limits of systems that use only patterns of contexts, which represent the main motivation of this thesis;

- To find novel techniques that combine the information about the word co-occurrences with word distributional information in a way that they complete each other for reducing the lack of information;

- To create general systems based on this novel approach that are able to address various types of lexical-semantic relations;

- To determine the properties of the contexts that are meaningful for the acquisition of lexical-semantic relations;

1.3 Thesis Layout

The rest of the thesis is organised in the following chapters:

Chapter 2: Theoretical framework explains the relations that are targeted in this work, describes the general approach to convert the corpus in information useful for the acquisition task, and the problem that hinders this approach.

The automatisation of the acquisition of these types of relations is possible thanks to the large amounts of corpus available. In **Section Distributional Hypothesis** we introduce two hypotheses that motivate the acquisition based on contextual information, in form of distributional properties and patterns of contexts. **Section Lexical-Semantic Relations** introduces the subset of semantic relations we focused on, and how they differ from other semantic relations. In the **Section Contextual Information Acquisition**, we explain how this information is extracted from corpus and represented in vectorial spaces that model word meaning and lexical-semantic relations holding between words. Due to the data sparsity problem, the corpus lacks of sufficient patterns of context information, fact that motivates the present work, explained in **Section The Data Sparsity**

Problem. **Section Graph Theory** lists the most important notions about graphs that will be used in the following chapters.

Chapter 3: Related Work revises the major techniques developed for the acquisition of instances of the most important lexical-semantic relations. **Section Multi-relation Acquisition** focuses on general techniques developed for the acquisition of semantic relations, without targeting any specific relation. **Section Graph-based Representations**, introduces the approaches that used notions of graph theory for the same task, and **Section Predictive Vector Space Models** discuss the advancements in the Predictive Vector Space Models (VSM). Graph theory and Predictive VSM are key representations for the novel approaches that will be latter described.

Chapter 4: Approach and evaluation sets up the scenery for the experiments describer in the following chapters: acquisition is addressed with a supervised classification approach, and it is tested over two corpora and two datasets.

Chapter 5: A Pattern-based Model aims at highlighting the problems of systems that only use patterns of context, and a possible follow-up based on a part-of-speech-based generalisation technique. The results are limited due to the necessity of seeing two words co-occurring in the same sentence for being able to represent them.

Chapter 6: Overcoming Data Sparsity introduces two novel systems for the automatic acquisition of lexical-semantic relations, which are based on techniques of generalisation of the information in corpus. **Section Graph-based Representation Model** proposes a generalisation of the corpus information in a graph-representation of the corpus, in this way, patterns of contexts that are detected as informative for a target relation are combined with information coming from a network of words encoding word distributional properties. **Section Word Embeddings Representation Model** uses the vectorial representations created with word embeddings, which combines distributional information with information about semantically related words.

Chapter 7: Conclusions draws the main conclusions of this thesis and outlines future research lines to follow.

Chapter 2

THEORETICAL FRAMEWORK

2.1 Distributional Hypothesis

The theoretical framework of this thesis is based on the Distributional Hypothesis introduced by Harris (1954):

> *Words that occur in the same contexts tend to have the same meaning.*

This assumption was further popularised by Firth:

> *You shall know a word by the company it keeps! [Firth, 1957]*

Therefore, the words with which a particular word occurs together in a corpus of texts, i.e. co-occurs, represents its meaning:

> *The meaning of a lexical unit reveals itself through its contextual relations, without commitment to what the meaning "really is" [Cruse, 1986].*

We refer to the *distributional properties* or *distributional information* of a word, the lexical items that co-occur with it. These co-occurring lexi-

	Syntagmatic relations			
Paradigmatic relations	I	bought	red	apples.
	The boy	eats	green	vegetables.
	Mom	made	tasteful	cakes .

Table 2.1: The lexical relation representation.

cal items joint together represent the meaning of the target word. To compare the meanings of two words in the computational linguistic (CL), one has to compare their distributional properties.

Harris hypothesis is influenced by Saussure's structuralism theory. Ferdinand Saussure (1857-1913), considered the father of modern linguistics, focuses in his work on the structure of the language (*la langue*). His posthumously published collection of lecture notes *Cours de linguistique générale*, presents the language as a system of *signs*, ultimately lexical units. In the system, each sign has a value (*valeur*), roughly corresponding to the meaning of the lexical unit, and that can be described by its relations (intuitively meaningful *differences*) with the other signs of the system. These relations between signs can be studied along two different perspectives or "axes": paradigmatic and syntagmatic relations (vertical and horizontal axis respectively, see Table 2.1).

The *paradigmatic relation* is based on the substitution operation: it relates units that occur in the same context but not at the same time. Each item in a paradigmatic relation is significantly different and the use of one item, rather than another, shapes the preferred meaning of a text. Paradigmatically related words create sets that can be analysed in terms of a defining category, also referred as topical domain, like *humans* or *vegetables*, and may be further structured; for instance the class of *vegetables* contains *greens*, *root vegetables* and other similar lexical items; additionally *root vegetables* contains *carrots* and *potatoes*. A hierarchical structure can be derived from each set and it is the backbone of taxonomies, such as WordNet [Miller, 1995].

> *Paradigmatic relations, for the most part, reflect the way infinitely and continuously varied experienced reality is apprehended and controlled through being categorized, subcategorized or graded along specific dimensions of variation [Cruse, 1986].*

The *syntagmatic relation* is a relation based on linear sequences of lexical items in sentences, and relate units that occur together in the same context and at the same time. The final meaning of a text is given by the combination of the meaning of each lexical unit. These relations can be said to describe the semantic space of a lexical unit and together with taxonomic relations mentioned earlier serve to create ontologies, such as ConceptNet [Liu and Singh, 2004].

> *Syntagmatic aspects of lexical meaning [...] serve discourse cohesion adding necessary information redundancy to the message, at the same time controlling the semantic contribution of individual utterance elements through disambiguation [Cruse, 1986].*

While syntagmatic relations, being *in praesentia* relations, may be directly observed in a corpus of text, the paradigmatic relations, *in absentia* relations, usually are not directly observed in corpus, except when they occur in very specific linguistic patterns describing the semantic relation, such as *"a carrot **is a** vegetable"* or *"vegetables **such as** potatoes and tomatoes"*.

The structuralist theory presents the language as a system of signs connected by syntagmatic and paradigmatic semantic relations in which they enter. Therefore, the *meaning* or the *semantic space* of each sign is described by these relations. The present work focuses on lexical items as signs, more specifically words, and the syntagmatic and paradigmatic relations are divided in the lexical-semantic relations (see Section 2.2).

Following the distributional hypothesis, two lexical units that are found in similar contexts are considered *lexically similar*[1], and members of the

[1] also named *attributional similarity* in [Turney, 2006b]

same paradigmatic class. However, lexically similar words hold different types of relations between them. For instance, *dog* and *pet*, are lexically similar and both are instances of the same semantic class, but their relation is different of the relation between *dog* and *cat* or between *traffic* and *street*, which are also lexically similar words.

[Turney, 2006b] introduces the second type of similarity in language: two pairs of lexical items holding the same semantic relation are called *relationally similar*. For instance the relation between *(dog, pet)* is similar to the one holding for *(apple, fruit)* and *(bottle, container)* because *dog* ***is a*** *pet*, *apple* ***is a*** *fruit* and *bottle* ***is a*** *container*, therefore, all pairs are instances of the relation of *hypernymy-hyponymy*. To detect similar related word pairs, [Turney, 2008a] extends the distributional hypothesis for detecting the relational similarity between word pairs and creates the Latent Relation Hypothesis:

> *Latent Relation Hypothesis: Pairs of words that co-occur in similar patterns tend to have similar semantic relations.*

In summary, the distributional hypothesis is used to represent the meaning of a target word, while latent relation hypothesis is used to represent the semantic relation holding between two words. Both approaches rely on information extracted from context: the distributional hypothesis uses the *words* co-occurring with a target word, and, the latent relational hypothesis uses *patterns of contexts* where words co-occur. Therefore, in this work, we mean *distributional properties* or *distributional information* of a word when we refer to the other words co-occurring in the same context with it, and *patterns of context of co-occurrence* when we refer to the context where a word pair co-occurs.

2.2 Lexical-Semantic Relations

In the present work, we address the classification of lexical-semantic relations, relation that emerge from the meaning of the words involved in the relation and stand between common names. Such relations are: hypernymy and hypernymy, co-hyponymy, meronymy or those derived from selectional preferences.

We will not deal with relations between named-entities, for instance *(Mircea Eliade, Bucharest)*, as a result of factual statement such as *Mircea Eliade was born in Bucharest*. These types of relations have different lexical and contextual properties, and usually are detected only by the observations in corpus. For instance, the same pair of words can instantiate more than one semantic relation, like the pair (Obama, USA) which holds the relations *BornIn* and *PresidentOf* [Hoffmann et al., 2011, Surdeanu et al., 2012], and tend to be explicitly expressed in language data in sentences like *Mircea Eliade* ***was born in*** *Bucharest*, in which the explicit context *X* ***was born in*** *Y* expresses the relation *BornIn*. This is generally not the case for lexical relations, as these relations tend not to be explicitly expressed in language data, but they are inferred from the set of contexts in which their instances co-occur. For example, [Girju et al., 2003] extracted patterns of contexts where meronyms tend to co-occur: 92.15% of these patterns were general patterns with an implicit meaning, such as *the door of the house* where *the house* is the *part* and *the house* is the *whole*. These patterns often match contexts expressing more than one relation type, making the acquisition of lexical relation instances more complicated.

In the next lines we introduce the most common lexical-semantic relations addressed in the NLP area, and we classify them as the Saussure's relations, paradigmatic and syntagmatic. In this work, we focus on the relations contained by the BLESS dataset [Baroni and Lenci, 2011] described in Section 4.2.2: hypernymy, co-hyponymy, meronyms and relations defined by the selectional preferences of the words. For the sake of completeness we also describe the relations of synonymy and antonymy, although this work does not address these relations. Synonyms and antonyms are relations having the highest lexical similarity and usually are addressed comparing their distributional similarity.

2.2.1 Paradigmatic Relations

Paradigmatic related words, are words that tend to occur in the same context, therefore, they have similar semantic properties that make them similar in meaning.

Synonyms & Antonyms Two words that have an identical or similar meaning, such as *(car, automobile)* and *(couch, sofa)*, are named *synonyms*. Formally, this relation is defined: two words that can be replaced one for the other in a sentence without changing the true value of the sentence hold a relation of *synonymy*.

While synonyms are words with the same meaning, antonyms are words that have opposite meanings, such as *(long,short)* and *(friend,enemy)*. More formally, two words that share all the aspects of their meaning but one, which puts them at opposite ends of some scale. For instance, the words *friend* and *enemy* share a significant part of their meaning, but one property puts them on opposite ends of the scale *relationship between known persons*: one person can be either your friend or your enemy, but not both at the same time.

Because, synonyms and antonyms differ just by one semantic property it is very difficult to automatically distinguish these kind of relations using patterns of contexts, as they tend to occur at the same contexts, but usually in the same time.

Hyponymy One word is the *hyponym* of a second word if the former word is more specific, denoting a subclass of the latter one, such as *(apple, fruit)* or *(dog, animal)* The opposite relation is *hypernymy*: *fruit* is the hypernym of *apple*. The superordinate is named the *hypernym*, and the subclass the *hyponym*. Hypernymy relation has some important properties:

Inclusion The class denoted by the hypernyms includes the class denoted by the hyponyms. For instance the class of *animals* includes the class of *dogs*.

Entailment Given a word pair (A, B) instance of the hypernymy relation, A is the hyponym of B, and B is the hypernym of A. Moreover, A entails being a B and A inherits all the semantic properties B.

Transitivity If A is a hyponym of B and B is a hyponym of C, then A is a hyponym of C.

The relation of hyponymy is a general relation that can be applied over almost every concept in language.

Co-hyponymy Two words have a relation of co-hyponymy when they have the same hypernym, but they do not have a relation of hyponymy, hyperonymy or synonymy. For instance, *apple* and *pear*, both are hyponyms of *fruit*. Moreover, for a set of words to be co-hyponyms, they have to be part of the same semantic class and to have a relation of incompatibility. For instance, if a fruit is an *apple*, it cannot be also a *pear* or any other fruit. Differently, *queen* and *mother* are both hyponyms of *woman* but they are not co-hyponyms because they do not "respect" the incompatibility rule: a queen can be a mother, and they do not belong to the same lexical class, *queen* belongs to the *aristocracy* lexical class, while *mother* belongs to the *relatives* lexical class.

Meronymy Two words have a relation of *meronymy*, or *part-whole* relation, if one is part of the other. For instance, *wheel* is part of a *car*, where *wheel* is a meronym of *car*. The inverse relation is named *holonymy*, in this case, *car* is the holonym of *wheel*.

In linguistics, the meronymy relation is seen as a collection of relations. [Winston et al., 1987] determines six types of meronymy relations: (1) *Component - Integral Object*, such as (car, wheel); (2) *Member - Collection*, such as (tree, forest); (3) *Portion - Mass*, such as (slice, pie); (4) *Material - Object*, such as (alcohol, wine); (5) *Feature - Activity*, such as (chewing, eating); (6) *Place - Area*, such as (oasis, desert).

Here we do not distinct between different types of meronyms, although conceptually they define different relations, in context they occur in similar patterns [Ittoo and Bouma, 2010].

2.2.2 Syntagmatic Relation

The distributional hypothesis is based on the syntagmatic relations holding between words. The context within the words occur, i.e. the words with which it combines to create the meaning of the text, describe their semantic properties.

Selectional prefernces Syntagmatic relations are linear relationships holding between words. They indicate compatible combinations between words in contexts and create co-occurrence restrictions within syntax. Unlike the paradigmatic relationships, the syntagmatic relationships of a word are not only about meaning, but they are also about the lexical company that the word keeps.

[Cruse, 1986] describes the syntagmatic relations in terms of *selector* and *selectee*: in a syntactic construction such as head-modifier or head-complement construction, the members of the construction can be a *selector* or a *selectee*, and the selector proposes one or more semantic traits of the selectee. For instance, in the construction *drink X*, the verb *to drink*, the selector, co-occurs only with nouns having the semantic property of *liquids*, which are the selectees. In the construction *X water*, the selector, *water* demands a verb that can be done to the concept of *water* which is a *liquid*, such as *to drink* or *to pour*, but cannot be *to build*. Likewise, in the construction *pregnant X*, the word *pregnant* co-occurs with lexical units bearing the semantic trait *female*.

The acquisition of syntagmatic relations was addressed in the NLP area through the selectional preferences task: the inclination of a selector (also named *predicate*) to select possible and plausible filler selectees (also named *arguments*) for particular roles, such as the subject and the objects of a verb or the nouns and adjectives that may modify a target noun. By definition, when the selectional preference task means the selection of the predicate given the argument. The inverse task, the selection of the argument based on a given predicate, is named inverse selectional preference. The syntagmatic relations are classified according to the syntactic class of the selector in classes such as *verb-argument preferences* or *noun-modifier preferences*.

2.3 Contextual Information Acquisition

The assumption of the present work is that the acquisition of lexical relation instances rely on the acquisition and on the representation of information about the contexts where the words co-occur, and also on the meaning of the words. The former information is provided by patterns of contexts were words co-occur, and the latter is provided by the distributional properties of the words. This section aims at introducing the reader to the main techniques employed to acquire the desired information from an input corpus, and next to illustrate the meaning of words or the relation holding between word pairs.

2.3.1 Context

For the NLP systems, a corpus in a human language is a sequence of words without having an associated meaning. The meaning is induced from the corpus, on the basis of the distributional hypothesis, which states that word meaning is represented by the context of the word. In order to acquire the necessary contextual information from corpus, first we have to define which words in sentences have their meaning connected.

In literature there are two ways to consider the context of a word: as a bag-of-words or based on dependency relations holding between words, and the choice of one over the other influences the representation of lexical items.

In the *bag-of-words model*, the context is seen as a linear sequence of words, where the meaning of a word is represented by the words that are close to it in all the contexts where it occurs. Therefore, the distributional properties of a word are created with the set of words occurring in a window of *(-m,n)*, *-m* being the number of words before and *n* the number of words after the targeted word. In this model, the information about the contexts of co-occurrences of a word pair *(w_1,w_2)* are extracted using surface lexical patterns:

$$\text{[0 to } n_1 \text{ words] } x \text{ [0 to } n_2 \text{ words] } y \text{ [0 to } n_3 \text{ words]}$$

and the occurrences of the target word pair are delexicalized, i.e. substituted by a place-holder.

The *dependency-based model* is based on the dependency structure of each sentence. The distributional information is created only with the words entering into a dependency relation with the target word. The information about word co-occurrences is extracted based on the dependency paths connecting two words in the the dependency tree outputted by a dependency parser:

$$x \xrightarrow{d_1} w_1 \xrightarrow{d_2} \dots w_n \xrightarrow{d_n} y$$

The differences between these two models are presented with the following example. Given the sentence:

(S2.1) Apple$_1$ juice$_2$ and$_3$ apple$_4$ cider$_5$ are$_6$ both$_7$ fruit$_8$ beverages$_9$ made$_{10}$ from$_{11}$ apples$_{12}$.

For the bag-of-words model, the co-occurrences relations are directly extracted from the sentence. For the dependency-based model, instead, the sentence is parsed, the resulted dependency relations are shown in Table 2.2. Only these relations are considered word co-occurrences.

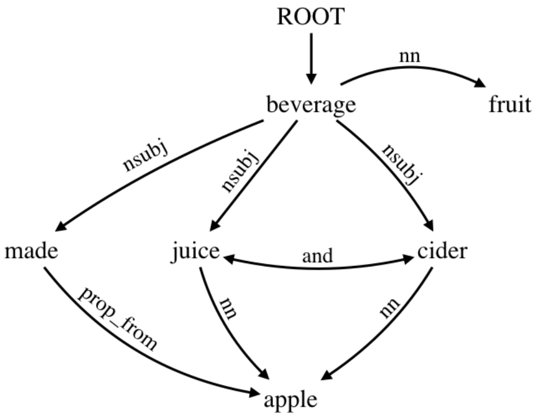


Figure 2.1: Parsing tree

Dependency relations
nn(juice$_2$, apple$_1$)
nsubj(beverages$_9$, juice$_2$)
cc(juice$_2$, and$_3$)
compound(cider$_5$, apple$_4$)
conj_and(juice$_2$, cider$_5$)
nsubj(beverages$_9$, cider$_5$)
cop(beverages$_9$, are$_6$)
det(beverages$_9$, both$_7$)
compound(beverages$_9$, fruit$_8$)
root(ROOT, beverages$_9$)
nsubj(beverages$_9$, made$_{10}$)
case(apples$_{12}$, from$_{11}$)
nmod_from(made$_{10}$, apples$_{12}$)

Table 2.2: Dependency relations acquired with the Stanford Parser in the collapsed format from the sentence S2.1.

The Table 2.3 presents the co-occurrences in which *apple* occurs obtained with the *bag-of-words* model within a window (-3,+3) and Table 2.4

the co-occurrences obtained with the *dependency-based* model. The first column indicates co-occurrences and the second column indicates the number of instances of the head word pair in the context. For instance *(apple, juice)* is seen two times in the bag-of-words approach in the given window (-3,3): *(apple$_1$, juice$_2$)* and *(juice$_2$, apple$_4$)*, the numbers indicate the position of the words in contexts, the word position in the pair is not considered; in the dependency-based model, the same pair is seen only once in the dependency relation *nn(juice,apple)*. For this example, we have also lemmatised the words.

Co-occurrences	#
(apple,juice)	2
(apple,cider)	1
(apple,and)	2
(apple,apple)	1
(apple,be)	1
(apple,both)	1
(apple,beverage)	1
(apple,make)	1
(apple,from)	1

Table 2.3: Co-occurrences of *apple* acquired with the bag-of-words approach from the sentence S2.1.

Co-occurrences	Dep. rels	#
(apple, juice)	nn(juice,apple)	1
(apple, cider)	nn(cider,apple)	1
(apple, made)	prep$_{\text{from}}$(made, apple)	1

Table 2.4: Co-occurrences of *apple* acquired with the dependency-based approach from the sentence S2.1.

Differences also appear in the patterns of context. For instance, the pattern of context containing *apple* and *juice* in the two models are presented in Table 2.5 and Table 2.6:

Context	Pattern of context	#
apple juice	X Y	1
juice and apple	X and Y	1

Table 2.5: Patterns of contexts holding the pair *(apple, juice)* acquired with the surface lexical pattern X [0 to 3 words] Y from the sentence S2.1.

Context	Pattern of context	#
$apple \xleftarrow{\text{nn}} juice$	$X \xleftarrow{\text{nn}} Y$	1
$apple \xleftrightarrow{\text{cc}} cider \xleftarrow{\text{nn}} juice$	$X \xleftrightarrow{\text{cc}} cider \xleftarrow{\text{nn}} Y$	1
$apple \xleftarrow{\text{cc}} made \xleftarrow{\text{nn}} beverage \xrightarrow{\text{nsubj}} juice$	$X \xleftarrow{\text{cc}} made \xleftarrow{\text{nn}} beverage \xrightarrow{\text{nsubj}} Y$	1

Table 2.6: Patterns of contexts holding the pair *(apple, juice)* acquired with the dependency based model with paths up to 3 edges from the sentence S2.1.

In this section we introduced the definition of a context and the method leveraged to acquire distributional information for words and patterns of contexts for word pairs. For a system to be able to use this information, it has to be represented in a way that reflects the meaning of a word and the lexical-semantic relation of a word pair, respectively.

2.3.2 Representations

The most common approach to represent the meaning of lexical items is based on the vectorial metaphor [Sahlgren, 2006], by which each lexical unit is represented as a point in a vectorial space. The vectors are positioned according to their semantic similarity, close vectors are expected to be semantic similar lexical units while distant vectors are lexical units that do not share semantic properties.

> *The geometric metaphor of meaning: Meanings are locations in a semantic space, and semantic similarity is proximity between the locations.*

> *[Sahlgren, 2006].*

This language model is called Vector Space Model (VSM) and is applied to represent the meaning of words, when encodes the distributional properties of words, and the lexical-semantic relations, when encodes patterns of context.

There are two mainstream approaches to transform the contextual information extracted from the input corpus into a vectorial space: the Traditional VSM, where the vectors are created directly by counting the co-occurrences observed in corpus as independent observations, and the Predictive VSM, where the co-occurrences observed in corpus are used as dependent observations to create vectorial representations to predict the co-occurring contexts for the target lexical unit. Both approaches are based on the word co-occurrences in context.

A large amount of work was dedicated to the representations of lexical units in a Traditional VSM. The creation of a Traditional VSM is directly motivated by the distributional hypothesis as it was defined by Firth and are created in the NLP field.

The Predictive VSM, named like this by [Baroni et al., 2014], contains vectorial representations computed using a neural network, also called *neural network models* or *word embeddings*. These representations have been relatively recently introduced in the NLP community, achieving state-of-the-art results for many tasks [Baroni et al., 2014], and were created by researchers in the machine learning field.

The method implemented in this work addresses the acquisition of lexical relations based on latent relational hypothesis and on distributional hypothesis, and therefore searches for reliable representation of two types of lexical items: words and word pairs. In the remaining of this section, we introduce the information encoded in the vectorial representation of each lexical items in each vectorial space, as well as the technique used to detect similarity between the lexical items.

Traditional Vector Space Models

The Traditional VSM represents each lexical item into a vectorial space where the dimensions of the vectors are context features c_i and the information encoded in the vectors, x_i, weights the importance of each context c_i for the $lexical\_unit$.

$$\begin{array}{rcccc} lexical\_unit = & (x_1, & x_2, \ldots, & x_n) \\ & \uparrow & \uparrow & \uparrow \\ & c_1 & c_2 \ \ldots & c_n \end{array}$$

Because of the zipfian distribution of words in any corpus, many possible co-occurrences are not observed [Bel, 2010] and consequently many vectorial representations in a Traditional VSM are very sparse.

To overcome this issue, the initial vector space may be reduced by applying a dimensionality narrowing technique such as Singular Value Decomposition [Golub and Kahan, 1965] or Random Indexing [Sahlgren, 2006].

Next, we introduce how the representations of various word pairs are created.

Word Representation The word representation starts from word co-occurrences in corpus. As previously explained, different types of contexts create different representations. [Lund and Burgess, 1996, Rapp, 2003, Sahlgren, 2006, Bullinaria and Levy, 2007] use co-occurrences within a certain distance from the target word, [Grefenstette, 1994, Lin, 1998a, Almuhareb and Poesio, 2004, Turney, 2006b, Padó and Lapata, 2007, Erk and Padó, 2008, Rothenhäusler and Schütze, 2009, Baroni et al., 2010] create vectorial representations based on dependency relations. [Baroni and Lenci, 2010] overpass the two-dimension representation and use higher tensors to represent tuples *word-link-word*. However, all these representations having high dimensions with sparse information, demand significant processing time and huge physical resources. Therefore, [Biemann and Riedl, 2013b] optimised the process using the Hadoop distributional system and proposed a novel representation based on lists of features created by the holing operation.

	$juice_N$	$apple_N$	$pear_N$	eat_V	$tart_N$	$cider_N$	$make_V$	jam_N
$apple_N$	37	29	54	43	17	11	10	1
$orange_N$	278	43	4	12	0	0	5	1
$quince_N$	0	1	0	0	1	0	0	1
$wall_N$	0	0	0	3	1	0	131	0

Table 2.7: Co-occurrence counts of the words $apple$ and $quince$, as nouns, within a set of manually defined context words extracted from BNC with the dependency-based approach.

Once the context is defined, these systems follow a similar methodology to transform co-occurrences in vectorial representations. Given a list of co-occurrences, these co-occurrences create a *word-context matrix* [Turney et al., 2010] counting how many times a word co-occurred within each particular context. Table 2.7 represents the co-occurrences counts of four headwords *apple*$_N$, *orange*$_N$, *quince*$_N$ and *wall*$_N$ in the BNC [bnc,] corpus with a set of vocabulary words, manually chosen as context for the sake of this example. The matrix rows represent the target words whose meaning is aimed to be represented as a vector; the columns represent vocabulary words and are used to represent the meaning in shape of dimensions of the vector.

Relying on the distributional hypothesis statement, these vectorial representations depict the distributional characteristics of each target word. The more similar two vectorial representations are, the higher their semantic similarities will be.

Formally, the semantic similarity of two words x_1 and x_2 is estimated by similarity scores of their vectorial representations v_{x1} and v_{x2} respectively:

$$sim_{sem} = score(v_{x1}, v_{x2})$$

[Curran, 2003, Weeds, 2003, Budanitsky and Hirst, 2006] give an extensive overview of existing vectorial similarity scores and their applications in various NLP tasks. However, this representation is hindered by the lack of information in corpus (see Section 2.4), not all the similar word pairs co-occur within the same context. For instance, *Apple*$_N$, *orange*$_N$ and *quince*$_N$ are all *fruits*, and are expected to have similar word representations. In Table 2.7, one can observe that *apple*$_N$ and *orange*$_N$ tend to share many co-occurring words, although with slightly different frequencies. Moreover, these representations are very different of the representation of a non-fruit noun, such as *wall*$_N$.

Instead, *quince*$_N$, has a very different representation of *apple*$_N$, *orange*$_N$: they share few contexts with very low frequencies. This is due to the fact that *apple*$_N$ and *orange*$_N$ occur more frequently in corpus, 1901 times and 945 times, respectively, while *quince* occurs only 17 times. To reduce the impact of the differences in co-occurrence frequencies, the word representations are created using a weighting scheme based on statistical measures that balance the co-occurrence frequency with the frequency of occurrences of each word. [Curran and Moens, 2002, Evert, 2005] carried out a broader survey of weighting measures in the context of Traditional VSM.

Word Pair Representation There are two mainstream research lines that addressed the word pair representation in Traditional VSM. The first approach is a compositional one: given the words *x* and *y*, they are represented in a Traditional VSM, creating *v(x)* and *v(y)*. The representation of the word pair *(x,y)* is created applying vector operations to *v(x)* and *v(y)*:

$$v((x, y)) = v(x) \oplus v(y)$$

Various vector operations were tested like addition, multiplication or tensor-product [Clark and Pulman, 2007, Widdows, 2008, Mitchell and Lapata, 2010, Guevara, 2010, Baroni and Zamparelli, 2010, Grefenstette and Sadrzadeh, 2011]. The results obtained differ across NLP tasks, raising many doubts about a general approach to represent the meaning of a word pair through vector composition in the Traditional VSM.

The second approach treats each word pair as a unit, and constructs its representation relying on the Latent Relation Hypothesis, as we will do in this work: the contexts where two words co-occur reflect properties of the lexical relation holding between them. To create word pair representations, the dimensions of the vectors are created with the contexts where the members of the pair occur together. As already stated, the contexts of co-occurrence are expressed through patterns of contexts, and depending on how the context is interpreted, as a bag-of-words or as a dependency-model, the patterns are acquired in form of surface patterns, or in form of dependency patterns.

The *word-context matrix* created for the word representations is modified into the *pair-pattern matrix* [Turney et al., 2010]: each row represents a word pair *(x,y)* and each column represents a pattern of context *p*, where *x* and *y* occur together. Each cell of the matrix contains weights that show the information provided by each pattern for the target lexical relation created based on the frequency of occurrence of the word pair within the pattern. [Turney, 2012] names this approach a *holistic approach* because word pairs are opaque wholes as their members do not have separate representations.

Two word pairs that hold similar lexical relations should have similar representations:

$$sim_{rel} = score(v(wp_1), v(wp_2))$$

Predictive Vector Space Models

The Predictive VSM [Bengio et al., 2003b, Bengio et al., 2003a] is the new generation of word vectorial representations. Different of the Traditional VSM, which shows within the vector components the importance of the context c_i observed in corpus for the target word *w*, the components of the

Predictive VSM is set to maximise the probability that a target word tends to appear within the context c_i. This way, semantically similar words have similar vectors.

A Predictive VSM is created using a neural network, which takes as inputs word co-occurrences and outputs vectorial representations of words in a continuous space, named also *word embeddings*. An important characteristic of these vectors is that their dimensions have no linguistic significance. Instead, they find patterns in the word co-occurrences, that are fold together in the vectorial representation.

In this work, we use an off-the-shelf system called `word2vec`[2] [Mikolov et al., 2013a] to create the word embeddings. The rest of the section is dedicated to explain how these representations interpret the contextual input to create representations of lexical items, without entering deeply into details in the mechanism used by the neural networks for learning the vectorial representations, which is behind the scope of this work.

Word Representation The `word2vec` system contains two architectures that differ in the way they interpret the input co-occurrences: the Continuous Bag-Of-Words model (CBOW) and the Skip Gram model. Both models are inspired by the Feedforward Neural Net Language Models, simplified to minimise the computational complexity, for training them on larger amounts of data. The CBOW model uses a log-linear classifier that, given a symmetric window, containing N words from the past and N words from the future, learns the classification of the word in the middle based on the sum of the representations of the other words in the window. The Skip-gram model has a different approach over the co-occurrences of words. Instead of learning the current word based on its context, this model uses each current word as input to a log-liner classifier, to predict words within a window of words.

Both models are very computational intensive, because they need to

[2] https://code.google.com/p/word2vec/

calculate the conditional probability of all words given a history. To reduce this computational cost, two statistical techniques are leveraged: the *hierarchical softmax* classifier and the *negative sampling* technique. The hierarchical softmax is a normalisation technique, which instead of calculating the probabilities for each node, it first creates a Huffman tree based on word frequencies. The final probability of a target words is estimated by mutiplying all the probabilities from the root to its position in the tree[3]. The negative sampling technique uses co-occurrences extracted from a corrupt corpus as negative examples when the neural network is trained to predict the vectorial representations. Another approach to reduce the computational time is reducing the number of occurrences of high-frequent words, which occurring very frequently in context and co-occurring with almost any word do not bring distinctive information about the meaning of a target word. To do so, a *subsampling technique* is used: the word occurrences in the training data are discarded with a probability that is proportional to their frequency.

The best results obtained in finding general semantic similar word pairs were obtained with the Skip-gram model, with negative sampling and the subsampling technique of frequent words, as explained in Section 3.3, and these are the parameters used in this work.

Word Pair Representation

The most remarkable property of the Predictive VSM, and the reason why they attracted a huge interest, is that: analogies between words seem to be encoded in the difference vectors between words. For instance $v(cars) - v(car) \approx v(apples) - v(apple)$, and $v(king) - v(man) \approx v(queen) - v(women)$. This observations show that they are able to preserve syntactic and semantic information: word pairs holding the same

[3] https://yinwenpeng.wordpress.com/2013/09/26/hierarchical-softmax-in-neural-network-language-model/

relations tend to be related by the same constant. Therefore, differently of the Traditional VSM, where two different spaces have to be created for the representation of words and word pairs, in the Predictive VSM the representation of the semantic relation of a word pair is done based on the vectorial representations of its members [Mikolov et al., 2013c]. Given two words x and y, with their embeddings *v(x)* and *v(y)*, the syntactic or semantic relation between them is reflected by their vector offset:

$$v((x, y)) = v(x) - v(y)$$

The intuition behind this observation is that although the dimensions of the vectors cannot be interpreted, the offset represents the changes in the dimensions of the vectors that are necessary to transform one word into another.

To similar offsets correspond similar relations [Zhila et al., 2013], therefore, the relational similarity between two word pairs $wp_1=(x_1,y_1)$ and $wp_2=(x_2,y_2)$ is estimated calculating the vectorial distance between theirs offsets.

$$sim_{rel} = score(v(wp_1), (wp_2)) = score(v(x_1) - v(y_1), v(x_2) - v(y_2))$$

2.4 The Data Sparsity Problem

Most of the research tackling the acquisition of lexical-semantic relations instances was done in the Traditional VSM using the holistic approach [Turney et al., 2005, Turney, 2006a, Turney, 2006b, Turney, 2008a, Turney, 2008b, Turney et al., 2010, Turney, 2012].

In Traditional VSM, the acquisition of lexical-semantic relations is done by calculating the vectorial similarity between word pair representations created using patterns of contexts. This approach achieves good precision, but it has an important drawback reflected in system's recall:

two words have to co-occur in an input corpus for assigning them a vectorial representations, and therefore, for detecting their relation. In any corpus of any size there will always be many semantically related word pairs that do not co-occur and consequently cannot be identified as instances of any semantic relation. Although the Web could provide additional information about the co-occurring contexts of a word pair, this approach can be applied only for the general domain, while for specific domains additional information may not be available. Consequently, the results obtained in the Traditional VSM, relying only on the word pair contexts of co-occurrences, are upper-bound limited by the amount of word pairs that co-occur in corpus.

Lexical-semantic relations holding between lexical items are an intrinsic part of the meaning of the words holding the relation.

> *It is assumed that the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts (Cruse, 1986).*

There is a strong connection between the lexical-semantic relation holding between a word pair and the semantic properties of its members, which are reflected in their contexts of co-occurrences, as stated by the distributional hypothesis in Section 2.1. We assume that in any corpus there is more valuable information, besides the contexts where two words co-occur, that can be used for identifying the lexical-semantic relation holding between two words. For instance, [Turney, 2006b] showed that including in the representation of a word pair also the information acquired with pairs created with the synonyms of the members improves the results, especially in recall.

In the previous section, we presented two vectorial spaces, Traditional VSM created directly from empirical evidences, and Predictive VSM, which are trained to predict the observed co-occurrences, for representing lexical units as vectors. Each vectorial space has its own way to represent words and word pairs. While the Traditional VSM is hindered by the lack of information, the initial experiments in the Predictive VSM although showing state-of-the-art results, the results obtained are far from the results obtained by human [Levy and Goldberg, 2014b]. This work searches for novel word pair representations for the automatic acquisition of lexical-semantic relations that fulfil our expectations: to combine information about the contexts where two words co-occur with the distributional information of each member in generalisation of the information in corpus. We propose a novel representation in the Traditional VSM based on a graph-based representation of the corpus and a novel system using a machine learning to discover information encoded in the Predictive VSM representations.

2.5 Graph Theory

To create a novel word pair representation in Traditional VSM, we use a graph-based corpus representation to generalise the information from the input corpus (see Section 6.1). Here, we introduce the basic notions from the graph theory that will be useful later in the acquisition of lexical information.

The graph theory is the study of graphs: a graph is a mathematical structure that represents a set of objects and their relation through a set of vertices and edges. Formally, a graph G, is a tuple (V, E), where $V = \{v_1, \ldots, v_n\}$ is a set of unordered elements, named vertices, and $E = \{(v_i, v_j)\}$ is a set of edges, unordered pairs of elements from $V \times V$.

In the context of language processing, vertices denote language units, whereas edges represent relations between them. We present now the main concepts related to graphs.

Vertices and Edges All the graph properties refer to the vertices, also named nodes, and the edges of the graph which are basic units in any graph structure. Given an edge $e = (v_i, v_j) \in E$, it is said that e connects v_i and v_j, and consequently that v_i and v_j are endpoints of e and v_i and v_j are two adjacent vertices. Equivalently, two edges that share the same vertices $e_1 = (v_i, v_j)$ and $e_2 = (v_i, v_k)$ are said to be adjacent edges. However, a vertex may exist in a graph and not belong to an edge, named isolated or singleton.

Dimension of the graph The dimension of the graph is expressed through the *order* of the graph which is the number of vertices, n=$|V|$, and through the *size* of the graph which is the number of edges, m=$|E|$. Therefore a graph G(n,m) denotes a graph G of order n and size m.

Graph Types Depending on the type of edges contained, the graphs are of various types. An *undirected graph* refers to its edges as unordered pairs of vertices, (v_i, v_j) and (v_j, v_i) represent the same edge. A *directed graph* has associated to the each edge a direction, (v_i, v_j) and (v_j, v_i) are two different edges. In this type of graph, the edges are named arcs, directed edges or arrows.

A graph, having more than one edge connecting the same tuple of vertices is named *multigraph*. A direct graph having multiple arcs connecting the same ordered pair of vertices is named *directed multigraph*.

Neighbourhoods A neighbourhood of a vertex $v \in V$ is the set of vertices that are adjacent to v in G: $neigh(v) = \{v_i | (v, v_i) \in E\}$.

Degrees The *degree* of a vertex v, $d(v)$, is the number of edges having v as an endpoint, which is the same as the number of neighbours of v: $d(v) = |neighbours(v)|$. A singleton vertex has the degree 0.

In a direct graph the degree is split in the *in-degree*, $d_{in}(v)$, which is the number of edges ending at v, and *out-degree*, $d_{out}(v)$, which is the number of edges starting at v.

Subgraph A subgraph G' of a graph G is a graph induced by a set of vertices $V' \subset V$ that contains all edges of G that connect any two nodes $v_i, v_j \in V'$.

Path or Walk A path, also named an open walk, in the graph G is a finite or infinite sequence of edges $(v_1, v_2), (v_2, v_3), \ldots, (v_{k-1}, v_k)$ which connect the sequence of vertices $v_1, v_2, \ldots, v_{k-1}, v_k$. In a directed graph, the sequence of arcs are directed in the same direction. If for every pair of vertices $v_i, v_j \in V$ exists a path from v_i to v_j, the graph is *connected.* A maximal connected subgraph of G is called a component of G.

Random Walk A random walk is a mathematical formalization of a path that consists of a succession of random steps. A random walk on a graph G is a path along its edges and the probability of walking along an edge is proportional to its weight. Given a random walker, the next step the walker does dependends only on the vertex where it is positioned at a certain time step and the probabilities of the weights of the edges are invariant of the specific time step.

Clustering A cluster of a graph G is a set of disjoint sets of vertices $\{C_1, C_2, ...C_n\}$ where $C_i \subset V$ and for all $i, j \in \{1 \ldots n\}, i \neq i : C_i \cap C_j = \emptyset$ and $\bigcup_{\{1..n\}} C_i = V$. Each vertices subset C_i is called *part* or *cluster*.

Connected component A connected component in a graph G is the maximal subgraph in which any two vertices are connected to each other by a path.

Strongly connected component A strongly connected component in a graph G is the maximal subgraph in which any vertex is reachable from every other vertex.

Weights A graph G is *edge-weighted* if each edge (v_i, v_j) has a weight w associated. Therefore, a function $f_e : E \rightarrow R^+$ exists that weights a given edge: $f_e((i, j)) = w_{(v_i, v_j)}$. A graph G is *vertex-wieghted* if each vertex v has a weight w associated. Therefore, a function $f_v : V \rightarrow R^+$ exists that weights a given vertex: $f_v(v) = w_v$. Unweighted graphs, also known as simple graphs, are a special case of weighted graphs with all vertex and edge weights set to 1.

Types A graph G is *edge-typed* if each edge (v_i, v_j) has a type t_e assigned. Therefore, given a set S_e of edge types there exist a function $t_e : E \rightarrow S$ that labels the given edge: $f_e((i, j)) = s$, where $s_e \in S$. A graph G is *vertex-typed* if each vertex v has an type t_v assigned. Therefore, given a set S_v of vertex types a function $t_v : E \rightarrow S$ exists that labels the given vertex: $t_v(v) = s_v$, where $s_v \in S$.

The entirety of vertex type assignments induces a partition of G: $V = \bigcup V_{1..n}$, for all $i, j \in \{1, \ldots n\}, i \neq j : V_i \cap V_j = \emptyset$ and for all $v_i \in V_k$ and $v_j \in V_k$, $t_v(v_i) = t_v(v_j)$.

Adjacent matrix The adjacency matrix of a graph G is a matrix A associated with G where $a_{ij} = 1$ if an edge exists between vertices v_i and v_j, $a_{ij} = 0$ otherwise. For edge-weighted graphs, $a_{ij} = w(i, j)$.

Chapter 3

RELATED WORK

This chapter presents the state-of-the-art in the acquisition of lexical-semantic relation instances. Previous works differ in the information used for the acquisition: patterns of contexts, distributional information of grammatical constructions, and in the approaches employed for detecting novel instances: bootstrapping or supervised classification of vectorial representations. However, their results, especially the recall, show that all are hindered by the data sparsity.

We assume the common shortcoming of these systems is the usage of information only directly observed in the corpus and their incapacity to generalise because of the chosen representations. Therefore, novel representation approaches of the information are presented in the final sections of this chapter: one based on a graph, as way to join the observed information towards a generalisation step, and another one based on Predictive VSM, the novel representation of words created with neural networks.

3.1 Lexical-semantic relation acquisition

The acquisition of lexical-semantic relations started with the seminal work of [Hearst, 1992] that introduced the patterns of contexts as information for the detection of the relation instances.

Hearst's work showed that word pairs having a relation of hyponymy may be identified in a corpus of text using a collection of manually defined lexico-syntactic patterns of contexts, presented in Table 3.1. Besides creating patterns that reflect the relation of hypernymy, Hearst defined a desiderata for creating reliable contextual patterns:

(i) They occur frequently and in many text genres.

(ii) They (almost) always indicate the relation of interest.

(iii) They can be identified with little or no pre-encoded knowledge.

(A) HYPERNYMS	
(H1)	NP$_H$ such as {NP$_h$,}* {and|or} NP$_h$
(H2)	such NP$_H$ as {NP$_h$,}* {and|or} NP$_h$
(H3)	NP$_h$ {,NP$_h$}* {,} or other NP$_H$
(H4)	NP$_h$ {,NP$_h$}* {,} and other NP$_H$
(H5)	NP$_H$ including {NP$_h$,}* NP$_h$ {and|or} NP$_h$
(H6)	NP$_H$ especially {NP$_h$,}* {and|or} NP$_h$

Table 3.1: Manually-developed lexico-syntactic patterns, where NP$_H$ is a noun phrase representing the *hypernym* and NP$_h$ the *hyponym*

The intuition passed on by this work is that the information used and the approach employed in an acquisition system depend on the properties of the relation targeted. In the following lines, we present the acquisition of lexical-semantic relations focusing on the types of relations addressed. We introduce approaches addressing the acquisition of the same relations as we do in this thesis: hypernymy, co-hyponymy, meronymy and selectional preferences of nouns. Because we focus on creating a general system for the acquisition of lexical-semantic relations, we also present approaches that were tailored for the acquisition of various lexical relations [Pantel and Pennacchiotti, 2006, Turney, 2008b].

3.1.1 Hypernyms Acquisition

The hypernymy acquisition started with Hearst's work. The Hearst patterns were an interesting approach for the acquisition of hypernyms but they were hindered by their scarce occurrence in text, which considerably limited the number of instances that could be detected. Moreover, due to the ambiguity and underspecification of the local context, some patterns did not fulfil the second point of the desiderata, and therefore the system was not accurate enough. For instance, the verb *to include* indicates the membership in lexical classes as well as the membership of groups such in *entire families including young children.*

The following works focused on improving Hearst's approach. A strategy to give a boost to the recall, thus, to find more hypernyms other than those co-occurring in Hearst patterns, was to leverage co-hyponyms of the more specific term of the relation (hyponyms): knowing that h_i is the hyponym of H, and h_i and h_j are co-hyponym, it is inferred that H is also the hypernym of h_j. This approach was used in [Caraballo, 1999, Pantel and Ravichandran, 2004]. They found clusters of similar words h_i, using similarity in their distributional properties, and Hearst patterns were used to find the common hypernym of the words from each cluster. [Cederberg and Widdows, 2003] initially extracted hypernyms (H, h) using Hearst patterns. Then, used coordinations in texts to find similar h_i with h, and finally inferred that H was hypernym of all the h_i.To reduce the errors caused by ambiguity or over-generalisation, the authors used the distributional similarity of words between H and each h_i, to filter out pairs whose members had a low similarity score. [Ritter et al., 2009] introduced machine learning techniques to learn general properties of the context of co-occurrence of a hyponym h with a hypernym H, and after, to discover co-hyponyms h_i for each hyponym h.

The aforementioned approaches relied on a small set of manually defined lexical patterns enhanced with distributional properties of words. These approaches were still hindered by the sparsity of the data in corpus and raised questions about their scalability for larger corpora [Pantel et al., 2004]. Therefore, the following work presented approaches that used the distant supervision approach to acquire a larger amount of surface patterns of context where hypernyms co-occur. [Pantel et al., 2004] proposed a generalization strategy based on part-of-speech information to combine them into more general patterns. Next, the general patterns were used to search in corpus novel instances in a bootstrapping approach. [Snow et al., 2004] extracted large amount of weak hypernym patterns of co-occurrence contexts created with dependency relations, and used their frequency in corpus to filter out unreliable patterns. Next, the patterns were used as features to create a traditional vector space that modeled the lexical relation holding between a pair of words as described in Section 2.3.2. Finally, a machine learning system was trained to weight the importance of each feature to detect hypernyms. This work reported 80% precision at 10% recall and 25 % precision at 30 % recall.

Other works used heuristic approaches for the acquisition of hypernyms based on these instances. [Yang and Callan, 2009] created a word pair representation using heterogenous features that described the patterns of contexts and the distributional properties of words. They achieved 79 points in recall over 50 WordNet datasets harvested from 12 topical domains. This recall was achieved with Wikipedia corpus and complemented with a dedicated corpus extracted from Google. [Hovy et al., 2009] also used the Web and a bootstrapping approach to extract hypernyms with doubly-anchored patterns in the backward direction [* *such as class_member*$_1$ *and class_member*$_2$]. [Navigli and Velardi, 2010] took advantage of the rigid structure of the definitional sentences and proposed a systems for the extraction of hypernyms and word-class lattices. [Boella and Di Caro, 2013] improved upon this work by learning syntactic properties of the hypernym and the hyponym and using a machine learning approach to weight the feature importance.

[Weeds et al., 2004, Zhitomirsky-Geffet and Dagan, 2005, Zhitomirsky-Geffet and Dagan, 2009, Lenci and Benotto, 2012] addressed the hypernymy acquisition under the assumption of the Distributional Inclusion Hypothesis: the more specific term appears in a subset of the distributional contexts in which the more general term appears. The approach is similar to the acquisition of co-hyponyms for which the similarity in the distributional properties is calculated. Being the hypernymy an asymmetric relation, the distributional approaches focused on discovering an inclusion score. Due to the data sparsity, the Distributional Inclusion Hypothesis did not stand for an offline corpus, forcing the authors to rely on an additional corpus acquired from the Web to prove the hypothesis.

A very recent approach used Predictive VSM to acquire hypernyms[Fu et al., 2014]. He showed that the relation of hypernymy is more complex than the semantic relations on which word embeddings have initially been tested, and the hypernymy relation cannot be represented based on the vector offset. Instead they learned linear projections that mapped the more specific term to a cluster of hypernym instances. For a Chinese corpus, the best results achieved 80 points in precision as well as recall, representing the best system in the hypernymy detection.

3.1.2 Co-hyponyms Acquisition

Co-hyponyms acquisition was initially addressed as a semantic lexicon acquisition task: the aim was to create a dictionary of words belonging to a given semantic class. For instance, given the topical domain of *fruits*, the objective was to extract from a corpus items belonging to this domain, like *apple* or *oranges*. However, the specificity of the items was not verified and even more specific instances of the domain, such as *pink lady apple*, were accepted. Therefore, this task extracted any term that belonged to the target semantic class, and this could include: co-hyponyms, hypernyms, hyponyms, synonyms or antonyms. Initial works relied on the assumption that words from the same semantic class tended to co-occur in grammatical constructions, and used bootstrapping approaches for identifying candidate word pairs [Riloff and Shepherd, 1997, Roark and Charniak, 1998, Phillips and Riloff, 2002]. This approach had very low coverage because only a few words co-occured in such constructions. To improve it, [Riloff and Jones, 1999] introduced the *mutual bootstrapping* technique. The initial set of grammatical constructions was replaced by a system that automatically identified the lexical patterns that produced candidate words. Mutual bootstrapping assumed that all the words extracted with each pattern belonged to the target semantic class, which was often untrue and incorrect terms were allowed in the lexicon, provoking a semantic drift. Therefore, next systems focused on filtering techniques for patterns and instances. [Curran et al., 2007] proposed to acquire only words that occured in only one semantic class. [Thelen and Riloff, 2002] evaluated the relatedness of each extraction pattern and of each candidate word against the target semantic class before being accepted as a positive instance of the class. [Igo and Riloff, 2009] used the Web to check for more reliable co-occurrences.

The bootstrapping technique was very sensitive to semantic drift. Therefore, other approaches used a graph representation of acquired candidate word pairs with a few patterns of contexts. The graph was created with each candidate word as a node and two nodes were connected if they were co-occurring in corpus. Regions of graphs that were more connected internally than with the other nodes in the graph represented a lexical class. A better explanation about the techniques used to analyse the graph structure will be detailed in the following Section 3.2 dedicated to the graph-based approaches in the acquisition task. Here we focus on the type of patterns used in the acquisition. The first approach using this methodology was [Widdows and Dorow, 2002, Dorow et al., 2004], used only the pattern [X and/or Y] to acquire possible co-hyponyms that were next represented in the graph. Using only one pattern to find possible candidates, resulted in a very low coverage of the corpus. Therefore, [Davidov and Rappoport, 2006] assumed that co-hyponyms tended to co-occur in symmetric patterns created with high frequency words and content words. Therefore, this work created a heuristic approach to automatically acquire symmetric patterns. [Kozareva et al., 2008] expanded the Hearst patterns and created double anchored hyponym pattern: *[class$_{name}$ such as 'seed' and *].* These patterns were used in a bootstrapping process to extract from the Web a set of words that were possible instances of the semantic class $class_{name}$.

Co-hyponyms were acquired also with distributional approaches based only on distributional properties. Distributional approaches were of two types, supervised, addressing this task as a classification problem: given a set of example instances for each lexical class, candidate instances were acquired based on the similarity of their distributional properties with the given examples. These approaches differed in the information encoded in the vectorial representation and the similarity scheme used [Hindle, 1990, Lin, 1998b, Kilgarriff and Yallop, 2000, Weeds and Weir, 2003, Curran, 2003, Gorman and Curran, 2006, Turney et al., 2010]

To avoid the need of annotated examples unsupervised approaches were introduced based on clustering techniques over the word representation. [Pantel and Lin, 2002, Lin and Pantel, 2002] created Clustering by Committees (CBC), which grouped similar words in committees, and each committee was considered a different lexical class. [Baroni et al., 2010] clustered the words based on the most prototypical properties of the word and [Baroni and Lenci, 2010] used the vectorial representation provided by the Distributional Memory to cluster the words. [Fountain and Lapata, 2012] used hierarchical clusters of words to deal with the granularity of the clusters. Similarities in the vectorial space model were used to create a semantic network, or a graph of words, reflecting the similarity scores between words. This network is partitioned in an initial set of word clusters and the Hierarchical Random Graph algorithm [Clauset et al., 2008] was leveraged to create hierarchical clusters, clusters that could be seen as a taxonomy without labels for the non-leaf nodes. The unsupervised approaches presented an important shortfall, the clusters provided were not labeled with the name of the represented lexical class.

Distributional approaches were able to acquire co-hyponyms that did not co-occur in corpus. Unfortunately, they extracted words holding different lexico-syntactic relations or topical relations. For this reason, the most meaningful approach to label lexical-semantic instances were approaches based on patterns of contexts [Weeds et al., 2014].

3.1.3 Meronyms Acquisition

While the acquisition of co-hyponyms and hypernyms may be addressed using both patterns of contexts and distributional properties of the members, the acquisition of meronyms was done only based on contextual patterns. [Berland and Charniak, 1999, Poesio et al., 2002], in parallel, followed Hearst's desiderata and developed a set of five patterns, M1-M5 in Table 3.2 for the acquisition of meronyms. However these patterns showed to be very ambiguous, and therefore, [Berland and Charniak, 1999] used only the first two patterns to acquire candidate meronyms which were filtered out based on a reliability scheme.

(B) MERONYMS	
(M1)	NN[S]_w 's NN[S]_p
(M2)	NN[S]_p of {the|a} [JJ|NN]* NN_w
(M3)	NN_p in {the|a} [JJ|NN]* NN_w
(M4)	NN[S]_p of NN[S]_w
(M5)	NN[S]_p in NN[S]_w
(M6)	NN_w consists of NN_p
(M7)	NN_w is made of NN_p
(M8)	NN_p member of NN_w

Table 3.2: Manually-developed lexico-syntactic patterns: NN_p is a noun representing the *part* and NN_w the *whole*.

[Girju et al., 2003] ran a broader analysis of the co-occurring contexts of meronym instances, and split the patterns in two classes: explicit patterns, M6-M8 in Table 3.2, and implicit patterns, all the others. While explicit patterns were accurate but occured very few times in corpus, the majority of the meronyms co-occured within implicit patterns, which were very ambiguous patterns. [Girju et al., 2003, Girju et al., 2006] disambiguated the patterns using the semantic class of the words as provided by WordNet and a supervised classifier learned the semantic features of the constituents for each pattern. They reported 80.95% precision and 75.91% recall, paid with a considerable amount of manual work to label each word with its semantic class. This work showed the challenge of direct acquisition of meronyms from patterns of contexts. Therefore, systems addressing the meronymy acquisition needed additional information to achieve satisfactory results. Therefore, domain dedicated systems were developed based on heuristic techniques. For instance, [Van Hage et al., 2006] aimed at finding *wholes* for a set of target *parts*. They searched the Web for words that represented the *whole* for each *part* word, and to avoid the topic drift and assure reasonable precision, this work used a dictionary of domain vocabulary from which the *whole* had to be chosen. [Ling et al., 2013] approach used a distant supervision approach combined with the multi-instance learning approach [Bunescu and Mooney, 2007] for the meronymy acquisition in the biology domain. They reported 78.6% precision and 79.1% recall, and an improvement of almost 10 point in F-measure compared with the approach of [Mintz et al., 2009], due to the multi-instance learning. However, they used a different supervised classifier than [Mintz et al., 2009] which also could influence the final results.

[Ittoo and Bouma, 2010] tested if the problematic of the meronymy relation came from the various sub-relations that created the meronymy relation as explained in Section 2.2. However, even when addressing each sub-relation the precision improved only 2% [Ittoo and Bouma, 2010] and it was observed that the instances acquired were spread over various meronym sub-types.

This section tried to highlight the challenge of the systems addressing the meronymy acquisition.

3.1.4 Selectional Preferences

The selectional preferences task means finding valid arguments for a predicate in a target role. The most straightforward approach is to extract from a corpus the words co-occurring with the predicate in that target role. In any corpus there will be *predicate-argument* that will not co-occur and consequently, this approach results in a very scarce list of words, especially for very infrequent predicates. For instance, in the BNC corpus, the verb *eat* occurs with *cabbage* but it does not occur with *radish*, *cauliflower* or *beet*, which are also *vegetables* and valid predicates of the verb. Therefore, automatic systems are necessary to generalise beyond seen co-occurrence frequencies and to infer a semantically coherent set of candidates that can fill a given argument position.

The selectional preference task consists of two steps, the extraction of the seen co-occurrences of a predicate and its arguments and the generalisation of the arguments in a class representing valid fillers of the predicate. For instance, seeing in the corpus that the verb *to eat* co-occurs with *tomato* and *potato*, the model should be able to detect that all *vegetables* are valid fillers of the predicate *eat*. While the extraction step is done based on the dependency relations connecting a predicate with its arguments, the generalisation task is the key issue that splits the approach in two classes: approaches that are based on hand-crafted resources and approaches that rely only on the distributional properties of words in corpus.

Approaches that use manually created resources, such as WordNet, group the observed arguments in clusters of similar words, and then search in the given resource for the best generalisation of each cluster. The hypernym found represents a descriptor of the the valid fillers of that predicate. Systems leveraging this technique differ in the methodology used to cluster the words and to find the WordNet descriptor: the Kullback-Leibler divergence [Resnik, 1996], the minimum description length [Li and Abe, 1998], the chi-square test [Clark and Weir, 2001], and the Latent Dirichlet Allocation probabilistic model [Ó Séaghdha and Korhonen, 2012].

Distributional approaches acknowledged the shortcomings of relying on manually-built resources and focused on purely distributional methods to learn the generalisation of arguments from corpus observations alone. Novel arguments for a given predicate and argument position, were discovered based on their distributional similarity in corpus with the seen arguments. [Rooth et al., 1999] created clusters of similar agruments seen in corpus with the expectation-maximization algorithm. Each cluster was used to find novel fillers. [Erk, 2007, Padó et al., 2007, Erk et al., 2010] used vectorial similarity scores between the vectorial representations words, observed arguments and possible fillers, created with their co-occurring predicates in corpus. [Bergsma et al., 2008] automatically acquired positive and negative instances and trained a discriminative model for learning selectional preferences from unlabeled text. In this work, the vectorial representation used went beyond the co-occurrences in text, and was created with co-occurring verbs, string shallow features and clusters created with CBC algorithm [Pantel and Lin, 2002]. [Van de Cruys, 2010] presented a multi-way selectional preference induction, for more than one argument role, based on tensor factorization. The co-occurrences of subjects, verbs, and objects were represented as a tensor of dimension three, and a latent factorisation model was applied to generalise over unseen arguments. [Thater et al., 2010] modeled the predicate argument relations as a combination of first-order and second-order dependency vectors and used vectorial similarity scores to find similar predicates. Therefore, the selectional preference task was reduced at finding classes of similar words.

A very recent approach applied a Predictive VSM for the selectional preferences in [Van de Cruys, 2014]. They represented the predicate-argument tuple as the concatenation of the vectorial representations of the predicate and of the argument, and a feed forward neural network was trained to recognise valid predicate-arguments tuples. Despite the impressive results in other tasks, this approach based on Predictive VSM achieved a lower accuracy than [Erk et al., 2010] to find objects for target verbs.

3.1.5 Multi-relation Acquisition

The previous section presented systems developed for acquiring just one target lexical relation. The systems relied on the patterns of context of co-occurrences of words and, some of them, on properties of the instances of the target relation, reflected in their distributional properties. However, developing an acquisition system for each semantic relation is a laborious work, thus, other works created general approaches based on patterns of contexts for the classification of multiple relations simultaneously.

For each target relation, the mainstream research line leverages a set of word pairs, instances of this relation. These word pairs are used to automatically acquire contexts where they co-occur and to transform them into patterns of contexts. These patterns are used in two different approaches: in a bootstrapping approach or in a classification approach.

Illustrating the former line of work, [Pantel and Pennacchiotti, 2006] created the *Espresso* system, using a bootstrapping approach to the acquisition of any lexical relation. To find a balance between accuracy and coverage across different relation types, *Espresso* used a small set of high precise patterns combined with a set of generic patters, having a lower precision. They leveraged on Web queries to detect reliable instances for the target relation. The system was tested across various semantic relations, between them meronyms and hypernyms. Interestingly enough was their observation that the meronymy acquisition improved when generic patterns were used, instead, the hypernymy acquisition was hindered by this type of patterns, showing the different lexical properties of the two relations.

A classification approach was the Latent Relation Analysis (LRA) system [Turney et al., 2005, Turney, 2006b]. It focused on creating word pair vectorial representations producing the signature of their semantic relation. Therefore, word pairs instances of the same relation should have similar representations. They proposed a representation as described in Section 2.3.2. The components of the vectors were patterns of contexts of co-occurrences, that were filled in with the observed co-occurrences. The resulted vectors were very sparse, therefore, they used the Singular Value Decomposition to smooth the vectors, and lists of synonyms to improve the coverage of the system. Although this approach achieved state-of-the-art results for the analogy solving problem [SAT,], it was very inefficient, requiring nine days to run. In the same research line there were the works of [Nakov and Hearst, 2008] and [Bollegala et al., 2009]. Both works used text snippets returned by a web search engine as contexts of co-occurrence for a target word pair. The former approach improved the accuracy of the system, calculating the relational similarity between word pairs. The latter approach aimed at reducing the data sparsity of the vectors by grouping the patterns of contexts in clusters of similar patterns. Next, the obtained clusters were used to create the vectorial representation of word pairs.

The approaches presented before, are based only on patterns of contexts and are able to represent only word pairs that co-occur in corpus, treating them as a unit. For this reason, they are named *holistic* or *non-compositional* approaches. Their main drawback is that they cannot scale up for novel word pairs that are not observed in corpus at the moment of training [Turney, 2013].

Two mainstream research lines can be used to overcome this limitation of pattern-based approaches: combining information about the patterns of contexts of co-occurrences of two words, with the distributional information of each word, and compositional approaches, based on the vectorial representations of the members of the pair.

The former approach assumed that the members of the instances of a semantic relation were pairwise semantic similar. This way, the systems take into consideration, besides the relational similarity, the lexical similarity between pairwise words. Such systems were [Turney, 2006b, Turney, 2008a], which proposed to estimate the lexical relation holding between two words as a linear combination of the lexical similarity of the pairwise members. However, their findings showed that the relational similarity could not be reduced just to a linear combination of the lexical similarity. [Herdağdelen and Baroni, 2009] concatenated information about the contexts in which the two words occur independently, and the contexts in which they co-occur represented as a bag-of-words, therefore, eliminating the structure of the context. The results showed that this representation was not very reliable, achieving low results on the analogy task but competitive results on the selectional preference task. [Ó Séaghdha and Copestake, 2009] created a word pair representation that combined lexical and relational similarity between two word pairs using a kernel-based framework for comparing sets of contexts using a feature embedding function associated with a string kernel. [Guevara, 2010, Baroni and Zamparelli, 2010] generated a few holistic vectors to see word pairs and then a supervised regression model was trained to predict a representation for a novel word pair whose members were observed independently in the corpus.

The compositional approach aimed at finding a representation for the semantic relation holding between a word pair, by combining the vectorial representations of its members [Mitchell and Lapata, 2008, Mitchell and Lapata, 2010, Clark and Pulman, 2007, Erk et al., 2010]. In the lexical acquisition task, [Turney, 2012] proposed a compositional approach based on the domain similarity and the functional similarity of the members. The domain similarity was calculated in a domain space created with the nouns that occur near a target noun. The function similarity was calculated based on a function space created with the verbs occurring near the target noun. The final combination of these two similarities reflected the semantic relatedness between two words. On a very similar approach, [Turney, 2013] trained a supervised system to learn the feature importance for semantic relatedness. The results obtained in the task of solving analogies and in the task of noun-noun modifiers showed that the compositional approach achieves lower results than the non-compositional approach.

Research in the Traditional VSM cannot find a suitable representation of the relation holding between the word pairs: the holistic approach is hindered by the lack of information in corpus, and previous work could not find a way to compose the vectorial representations of words to represent their semantic relations. Predictive VSM stands out showing that for some general relations [Jurgens et al., 2012], the semantic relation of a pair of words is reflected in the vector offset and the similarity of the semantic relations is identified using the cosine similarity. [Necsulescu et al., 2015] applied the same approach to the acquisition of lexical-semantic relation instances and the results showed that methods based on compositional approaches over predictive representations did not work for this type of relations. [Levy et al., 2015b, Necsulescu et al., 2015] replaced the vectorial similarity with a supervised approach that was trained to learn distinctive features of the word pair representation that indicated the lexical-semantic relations, [Necsulescu et al., 2015] obtaining very good results. However, doubts were raised by [Levy et al., 2015b], which stated that the good results achieved by the supervised classifiers were due to the lexical memorization: “the phenomenon in which the classifier learns that a specific word in a specific slot is a strong indicator of the label”.

3.2 Graph-based Representations

Graph-based representations have been largely used to model the information in various NLP tasks: lexical acquisition [Widdows and Dorow, 2002], word sense disambiguation [Navigli and Velardi, 2005, Hughes and Ramage, 2007, Navigli and Lapata, 2010], language representation such as in WordNet, Roget’s Thesaurus, or ontology representation [Steyvers and Tenenbaum, 2005], text comparison [Mihalcea and Tarau, 2004, Tsang and Stevenson, 2010], among many others. graph-based approaches model relations between words in the structure of a graph, to create a generalization of the corpus for a target task.

Here, we analyse works that used the graphs as a reliable representation for the lexical information acquisition task, and as a possible representation of the structure of the corpus.

In the task of acquiring classes of similar words, the edges connect word pairs that possibly belong to the same class of words, and clustering algorithms are used to detect groups of similar words. Therefore, these approaches differ in the information joined in the graph and on the algorithm that analyses its structure. [Widdows and Dorow, 2002] created a graph representation based on all the nouns co-occurring in conjunctions in a given corpus. Words from the same class are detected using Markov Clustering Algorithm. [Davidov and Rappoport, 2006] used symetric patterns to acquire a larger amount of co-occurring words (see Section 3.1.2) and words from the strongly connected components were detected as words from the same lexical-class. On the same graph [Schwartz et al., 2014] used an iterative variant of k-Nearest Neighbour clustering algorithm to propagate the semantic class of words to its neighbours. [Biemann, 2006] created a graph reflecting the number of co-occurring words shared by two words. The edges were weighted with the number of common neighbours and the Chinese Whispering algorithm was run to detect classes words. [Schütze and Walsh, 2008] created a graph-theoretic model of the acquisition of lexical syntactic representations. The corpus was represented as a complete bipartite multi-graph, and the edges were weighted to show the probability that a word w_i was the left neighbour of w_j. A random walk was run over the graph to compute the probability of any two words to be in a syntagmatic and paradigmatic relation. [Biemann and Riedl, 2013a] used graph structures to represent the co-occurrence information extracted by the holing operation [Biemann and Riedl, 2013b] and heuristic techniques to detect the most accurate similar words based on the contexts.

For taxonomy acquisition the edges in graphs link possible candidates for the relation of hypernymy and the final objective is to transform the graph into a tree representing a taxonomy. [Kozareva et al., 2008] used the double anchored hyponym patterns to create *hypernym pattern linkage graph*. The graph was directed and each arc *(n_1,n_2)* $\in E$ signified that the node *n_1* was used in a pattern to discover the node *n_2*. The out-degree of the nodes was used to identify instances of the same semantic class and the principle of the longest path in the graph was used to discover valid hyponyms-hypernyms pairs [Kozareva and Hovy, 2010]. [Navigli et al., 2011] used a graph to represent all the word pairs acquired as possible hypernyms and used a heuristic technique to eliminate unreliable edges. The edges were weighted to leverage larger paths, and the graph was pruned into an acyclic graph, representing a taxonomy.

Graph models were also used as a representation technique to model the interaction of words in corpus based on their dependency relations. Such a graph is the dependency graph generated by a dependency parser like Stanford Parser [De Marneffe et al., 2006]. A dependency graph represents the dependency structure of each sentence: the vertices are tokens and the edges represent dependency relations. To find semantic relations between target words, [Nastase and Szpakowicz, 2006] matched word subgraphs representing their dependency relations. Other works aggregated the graph dependency of each sentence in a single corpus representation. For instance, [Szpektor et al., 2004, Tanev and Magnini, 2008] represented the isomorphic structures from different sentence graphs using only one edge in the corpus graph. In the [Minkov and Cohen, 2008, Minkov and Cohen, 2012] graph-based corpus representation, the node set was created with all the word tokens and their type from corpus, and the edges represented the syntactic relations between them or *mention* edges, which connected each token to its type. The edges and the vertices were weighted based on path-constrained graph walk optimized for the target task. [Navigli and Velardi, 2010] created word lattices to identify definitional sentences and to extract hypernyms of target terms. A word lattice is a direct acyclic graph, with a single starting point, that are used to represent any finite set of strings. Given a set of definitional sentences, each sentence is generalized based on part-of-speech information and they are combined into an untyped and unweighted word-class lattice. The previous presented graphs were designed for the acquisition of paradigmatic related words. Differently, [Toutanova et al., 2004] crated a graph model to estimate word dependency probabilities between two words. This task was considerably hindered by the data sparsity, therefore, various edge types were combined: syntactic dependencies, morphological links, edges between words that appeared as synonyms in WordNet and statistical edges. A random walk model was learned, whose stationary distribution represented the co-occurring probability of two words.

As observed, the graph represents an important technique to display information about the context which helps detecting semantic relations.

3.3 Predictive Vector Space Models

As already introduced in Section 2.3.2, the Predictive VSM are vectorial representations of words created with a neural network architecture. In the last years many types of Predictive VSM were created [Bengio et al., 2003a, Collobert et al., 2011, Huang et al., 2012, Mikolov et al., 2013a] achieving interesting results. In the present work, we use the vectorial representations created with the off-the-shelf system named `word2vec` [Mikolov et al., 2013a, Mikolov et al., 2013b, Mikolov et al., 2013c], which achieved high interest when it was shown that various syntactic and semantic regularities holding between words are captured in their vector offset. For instance, they captured masculine-feminine pairs, like *(man,woman)* and *(king,queen)*, language-spoken-in, like *(Spain,spanish)* and *(France,french)*, verb-past-tense like *(go,went)* and *(do,did)*. In what follows, we used interchangeably predictive vectorial representations and word embeddings to refer to the vectorial representations of words in the Predictive VSM.

Word embeddings were used to solve the analogy task, *"a is to b as c is to __"*, by searching the most suitable word w, which vectorial representation optimise the equation:

$$w^* = argmax_w(cos(w, v(b) - v(a) + v(c))) \tag{3.1}$$

where $v(a)$, $v(b)$ and $v(c)$ are the embeddings of a, b, and c. In other words, this equation finds the word w which is the most similar with b and c, and the most dissimilar to a. Word embeddings were also used to measures word pair similarity within a particular semantic relation (i.e., which pairs are most prototypical of a semantic relation) in the SemEval 2012 Task 2: Measuring Relation Similarity [Jurgens et al., 2012]. The relation holding between a word pair (w_1, w_2) is represented as a directional offset vector: $v((w_1, w_2)) = v(w_2) - v(w_1)$. The relation similarity problem is reduced to calculating similarities between offset vectors [Zhila et al., 2013]:

$$sim((w_1, w_2)), v((w_3, w_4)) = cos(v((w_1, w_2)), v((w_3, w_4))) \tag{3.2}$$

The Equation 3.1 was shown to be more suitable to discover syntactic related words, while the Equation 3.2 achieved the best results for labelling semantic relations instances, putting less emphasis on the similarity between the members of the pair. Moreover, [Levy and Goldberg, 2014b] stated that the analogy of syntactic related words is better reflected using vector multiplication rather than vector addition.

Regarding the construction of Predictive VSM, as explained in Section 2.3.2, there are two architectures included in the `word2vec` system: the CBOW and the Skip-gram model. [Mikolov et al., 2013a] stated that both models achieve similar results for detecting syntactic relations, while the Skip-gram model preformed better for detecting the semantic relations. However, the CBOW model was faster on training, and therefore it was more suitable for large datasets. To improve the word representations and to speed-up the training time, [Mikolov et al., 2013b] combined the Skip-gram model with negative sampling technique.

In the Traditional VSM, vectorial representations created with dependency relations were more reliable, especially for smaller corpora. Thus, [Levy and Goldberg, 2014a] created a dependency based word embeddings based on the original Skip-gram model. When compared with the original model, the dependency based model yielded more functional similarities of a co-hyponym nature, while the original model found broad topical similarities. However, on relation similarity task the dependency based approach performed worse than the original system based on bag-of-words.

Word embeddings were embraced by many researchers as a possible solution for tasks where the Traditional VSM reach their limits. [Baroni et al., 2014] conducted a set of experiments to test in which tasks the embeddings outperform the traditional representations. Surprisingly, the results confirmed that predictive models were the most suitable representations for all the tasks.

However, due to the compression of the corpus observations in a predictive space, these novel and efficient representations had a considerable drawback: their results cannot be linguistically explained. As a consequence, an important research work focused on replicating their results using traditional models, because their dimensions can be interpreted. [Levy and Goldberg, 2014b] showed that comparable results with those obtained by the Skip-gram model enhanced with the negative sampling, were achieved with a traditional model created with a Positive PMI, and [Levy and Goldberg, 2014c] discovered that the same embeddings are a factorization of a word-context PPMI matrix. Therefore "this result implies that the neural embedding process is not discovering novel patterns, but rather is doing a remarkable job at preserving the patterns inherent in the word context co-occurrence matrix" [Levy and Goldberg, 2014b].

[Pennington et al., 2014] created a novel embedding structures, named GloVe, aiming at improving the local-contextual information leveraged in the Skip-gram model, by combining it with the global information acquired by matrix factorization methods. The authors stated that GloVe outperforms `word2vec` on word analogy, word similarity and named-entity recognition. These results are inconsistent with the findings from [Levy et al., 2015a], which compared four different word representations and analysis the importance of their parameters: two predictive models, `word2vec` and GloVe, and two traditional models, one based on the PPMI matrix and the other on the SVD factorization of the said matrix. The results are of high importance in the actual battle between embedding representations and traditional models. Skip-gram model emerged over the other systems achieving higher or comparable results, a significant reduction in training time and disk space, and scalability for very large corpora. Despite all the skepticism emerged around them, it seems that these representations filled in a gap in the NLP research, the representation of very general syntactic and semantic relations.

The Predictive VSM provides reliable representations that somehow combine information about lexical and relational information, which makes them viable representations for the acquisition of lexical-semantic relations instances.

3.4 Conclusions

Methods to acquire lexical-semantic relations rely on the patterns of contexts of co-occurrences and/or distributional hypothesis. [Turney, 2012] stated that composition over the individual representations of words is not able to represent the semantic relation holding between them. Therefore, compositional approaches in the Traditional VSM are out of the scope of this work. Here we address the holistic approaches that rely on patterns of contexts, also named pattern-based approaches. To show the achievements of this type of approaches, we create a system that uses only patterns of contexts presented in Chapter 5. The straightforward combination with information about the lexical similarity between the pairwise members was shown not to increase the similarity score between similar word pairs [Turney et al., 2010].

Therefore, we search for strategies that combine distributional information of words with pattern-based systems for the recall improvement. The assumption behind this work is that patterns of contexts and distributional information of words must be embedded indistinctly in the structure of the word pair representations. Therefore, it focuses on a novel technique that combines pattern based approaches with distributional information, for generalising over attested pairs and investigates Predictive VSM as possible efficient representations for acquiring lexical-semantic relations instances.

Chapter 4

APPROACH AND EVALUATION

The objective of this work is to find a novel vectorial representation based on a generalisation of the corpus for detecting lexical related pairs of words even when these pairs do not co-occur in the same sentence in corpus. The present chapter presents the leveraged approach, and sets up the context for the future experiments aimed to test the capabilities of the proposed systems: the corpus, the dataset and the evaluation methodology.

4.1 Approach

The acquisition of lexical-semantic relation instances is the task of finding the lexical relation holding between the members of a target word pair. Formally, the task addressed is defined: given a set of target semantic relations $R = \{r_1, \ldots, r_n\}$, and a set of word pairs $W = \{(x, y)_1, \ldots, (x, y)_n\}$, the task consists in labelling each word pair $(x, y)_i$ with the relation $r_j \in R$ holding between its members and outputting a set of tuples $((x, y)_i, r_j)$.

This work focuses on finding a general system for the acquisition of lexical-semantic relation instances. As presented in Section 3.1.5 the mainstream approach is to use the contexts where word pairs co-occur as main source of information to detect the semantic relation holding between their members. The results of this kind of systems are very limited, as it will be shown in the next chapter where we created a system based only on patterns of co-occurrences. This system proposes to generalise the patterns of context using part-of-speech tags, this way similar patterns but slightly differently lexicalized are represented as one feature. The results show that using only contexts where words co-occur, the recall achieved is very low.

Our intuition is that for deciding the semantic relation holding between words two sources of information are necessary, information about co-occurrences of a word pair as well as information about the distributional properties of each member of the pair. These two types of information should be combined so both information complement each other, being able to overcome the lack on information which hinders the pattern-based systems. This information should be generalised in order to infer information about words that do not co-occur in corpus. Therefore, the system recall is not upper-bound limited by the number of word pairs co-occurring, but by the number of word pairs having both members occurring in corpus. This work focuses at proposing novel representations, as explained in Chapter 6.

All the systems tested in the next two chapters use a supervised technique to discover the importance of the proposed feature for each target relation. Given a set of lexical-semantic relations R and a set of tuples $E = ((x, y)_i, r_i)$ of example instances for r_i, to set the importance of each dimension of the space for each target relation $r_i \in R$, a machine learning system, a support vector machine (SVM) multi-class classifier with a radial basis function kernel [Platt, 1999] with the SMO algorithm [Platt, 1999] from WEKA [Hall et al., 2009], is trained. The SVM system is run using all the default parameters but the complexity constant C, which was set to $C = 20$. Then, given a new unlabeled pair $(x; y)_u$, the classifier decides if the relation r is held between x and y. The trained SVM classifier generates a distribution over relation labels and then the highest weighted label is selected as the relation holding between the members of the word pair.

4.2 Data Sets and Evaluation Methodology

The following corpora, datasets and evaluation strategy are chosen to show the capabilities of the systems to deal with corpora of different size and with datasets having different granularity.

4.2.1 Corpora

As explained in the previous chapters, the results of the pattern-based systems depend on the number of the co-occurring word pairs in the input corpus. Many pattern-based systems increase the size of the input corpus in an attempt to overcome the lack of the information in corpus and to achieve a better recall. To test the impact of the corpus size over the results, this work uses two corpora of different sizes: the British National Corpus (BNC), a 100 million-word corpus, and a Wikipedia dump created from 5 million pages and containing 1.5 billion words. Figure 4.1 shows that both corpora have a power law distribution, which hinders the classification of information based on patterns of contexts. The size difference allows us to measure the potential impact of increased word co-occurrence on recall. Both corpora were initially parsed with the Stanford dependency parser in the collapsed dependency format [De Marneffe et al., 2006]. Besides the dependency relations provided by the parser, a dependency relation vb_V was added connecting the subject with the direct object of the verb V in a given sentence.

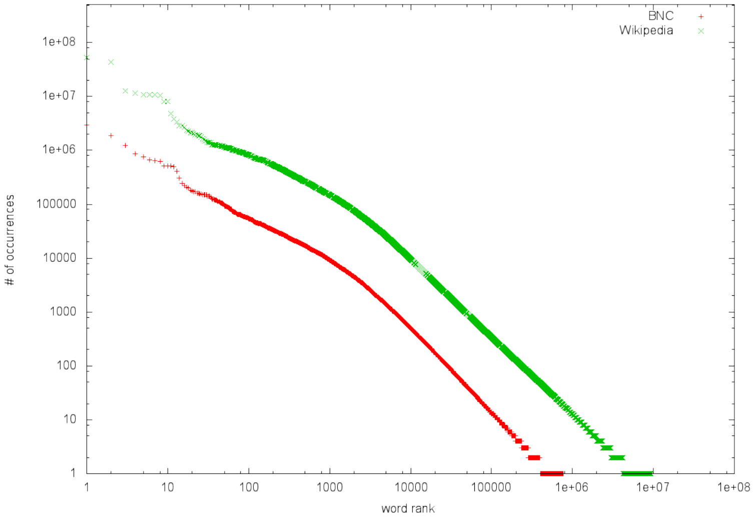


Figure 4.1: The plot of the words frequencies in BNC and Wikipedia; the words are previously lemmatised and desambiaguated by their PoS tag.

4.2.2 Data sets

The present work addresses the acquisition of lexical-semantic relations out-of-context. We chose two datasets for testing the ability of the novel systems to classify these types of relations: BLESS dataset [Baroni and Lenci, 2011] and K&H dataset [Kozareva and Hovy, 2010]. These datasets were chosen because their instances group different topical domain, allowing to test the importance of this type of information in the acquisition of lexical-semantic relation instances.

The BLESS dataset [Baroni and Lenci, 2011] spans 17 topical domains and includes five relation types: three paradigmatic relations: hypernymy, co-hyponymy and meronymy, attributes of concepts, and two syntagmatic relations: a relation holding between nouns and adjectives, and a relation holding between nouns and verbs showing the actions performed by/to concepts. In total, the BLESS dataset contains 14400 positive instances and an equal number of negative instances. The distribution across relations and across topical domain is presented in Table 4.1. This dataset allows to test the capabilities of the systems for lexical-semantic relation types other than the usual taxonomic relations and also to measure the generalisability of each system across various relations.

Domain	#Attr	#Co-hypo	#Act	#Hyper	#Mero
Amphibian reptile	73	32	83	36	52
Appliance	171	134	224	86	182
Bird	119	229	236	97	167
Building	186	103	276	72	304
Clothing	265	351	265	137	172
Container	111	67	131	43	121
Fruit	196	231	118	50	76
Furniture	105	81	207	45	112
Ground mammal	315	533	400	176	321
Insect	53	120	102	45	49
Musical instrument	93	174	155	44	96
Tool	213	397	306	107	124
Tree	71	75	63	28	119
Vegetable	126	284	174	76	56
Vehicle	316	371	594	123	648
Water animal	134	154	150	76	115
Weapon	184	229	340	96	229

Table 4.1: Distribution of BLESS instances across topical domains and across lexical relations.

WordNet is a state-of-the-art lexical resource and many works in the acquisition of lexical resources are carried out to automatize its development. To show the capabilities of our systems to acquire WordNet relations, we test the systems with the K&H dataset. [Kozareva and Hovy, 2010] collected a dataset of hyponym-hypernym instances from WordNet [Miller, 1995] spanning three topical domains: *animals*, *plants* and *vehicles*. However, the objective of this work is to detect other lexical-semantic relations than the relation of hypernymy. Therefore, this dataset is enhanced with instances of two more relation types: co-hyponymy and meronymy. Co-hyponyms are extracted directly from the K&H dataset: two words are co-hyponyms if they have the same direct ancestor[1]. To avoid including generic nouns, such as "migrator" in the "animal" domain, only leaf nodes are considered. The meronym instances are extracted directly from WordNet. The final dataset excludes multi-word expressions, which were not easily handled by any of the tested systems. The total number of instances considered in our experiments is presented in Table 4.2.

Domain	#Co-hypo	#Hyper	#Mero
Animals	8038 (92.4%)	3039 (97.2%)	386 (89.1%)
Plants	18972 (95.5%)	1185 (97.4%)	330 (82.4%)
Vehicles	530 (82.6%)	189 (97.9%)	455 (100%)

Table 4.2: Distribution of K&H dataset, with the % of instances which occur in the corpora.

4.2.3 Evaluation

We use two setups for evaluating the novel systems proposed in this thesis:

in-domain setup – only instances from one domain are used for training and testing in a 10-fold cross-validation technique. This setup allows to test the capabilities of the system to automatically classify word pairs as instances of a target relation.

out-of-domain setup – one domain is left out from the training set and used for testing. This setup allows us to see the importance of the topical domain information in the classification.

[1] y is a direct ancestor of x if there is no other word z which is hypernym of x and hyponym of y.

The system performances are compared based on the scores obtained for precision (P), recall (R) and F1-measure (F). Precision is defined as the percentage of correct relation classifications of those made by a system; recall is defined as the percentage of relation instances in the dataset correctly classified by a system. The F1 measure is the harmonic mean of precision and recall.

Chapter 5

A PATTERN-BASED MODEL

Approaches based on patterns of contexts in Traditional VSM represent the semantic relation holding between a word pair based on its members co-occurrences in context. Due to the ubiquitous phenomenon of data sparsity in corpus, they are characterised by low recall scores. The work presented in this chapter aims at showing the limitations of these type of systems.

Two pattern-based systems were considered as reference works for the acquisition of semantic relations instances: [Snow et al., 2004] introducing the dependency-based patterns for the hypernymy acquisition, and the work of [Turney, 2006b, Turney, 2008b] introducing generalised patterns of contexts created with wildcards and contextual words in an uniform approach for the acquisition of various lexical-semantic relations. This section introduces **Pa**ttern-based **C**lassification **M**od**E**l (PaCE), a system that combines the works of [Snow et al., 2004] and [Turney, 2006b, Turney, 2008b] in an attempt to acquire more information from corpus for improving the recall scores of pattern-based classification systems.

5.1 Pattern-based Representation Model

The present section introduces PaCE, a system that combines patterns of context created with dependency relations, with a generalisation technique based on part-of-speech information. [Snow et al., 2004, Mintz et al., 2009, Wu and Weld, 2010] also used dependency-based patterns of contexts, but they leveraged only the shortest dependency paths connecting the two target words. The shortest paths might be more precise but the set of paths discarded could provide valuable information for the increase of the recall. One of the strategies tested here for increasing the recall is to leverage all the dependency paths up to a maximum number of edges instead of only the shortest ones to create patterns of contexts. Additionally, to find similarities between slightly varying contexts of co-occurrence, and consequently increase the coverage of the automatically acquired patterns, a generalisation strategy based on part-of-speech information is also tested for creating more general patterns.

PaCE is based on the distant supervision approach: starting from a set of example word pairs, instances of a set of target relations, meaningful patterns of contexts are automatically acquired. To create a vectorial representation for given word pairs, the patterns are weighted according to the importance of each pattern for the detection of each target relation. The weights are learned using the SVM system introduced in the previous chapter, and the set of example instances*E*. Finally, new word pairs are classified based on their co-occurrences with the previously acquired contexts.

Pattern Acquisition PaCE uses patterns of context defined on dependency relations between words as collected from corpus data. As in the distant supervision approach, it is assumed that all the contexts in which the members of any word pair *(x,y)* co-occur are likely to provide information regarding the relation held between them. Therefore, for each pair of words $(x,y)_i$ that appears in the initial set of examples *E*, all sentences containing x and y are extracted. These sentences are then individually used to collect patterns of contexts as follows: all the dependency paths between x and y up to l_P edges and containing only nouns, adjectives, verbs and adverbs are harvested from the dependency graph of each sentence; neighbouring "window" nodes that are not included in the path, but are connected to one of the members of the word pair, are added to the path; each dependency path is transformed into a pattern of context by unlexicalizing x and y, i.e. x and y are replaced with a slot which can be filled in by any word within the same part-of-speech.

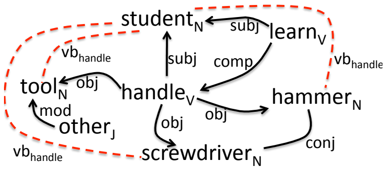


Figure 5.1: Dependency graph of the sentence "The students learned how to handle screwdrivers, hammers and other tools.". As mentioned, only words from the main part-of-speech are considered.

Pattern Generalisation For finding similarities between patterns of contexts slightly differently lexicalized, each pattern is generalised using part-of-speech information: $2^{(n-2)}$ patterns are generated, n being the number of words on the initial dependency path, by iteratively replacing each word in the pattern with its part-of-speech. Window nodes are not submitted to this generalisation procedure.

The patterns generalisation process is exemplified with the following sentence *"The students learned how to handle screwdrivers, hammers and other tools"*. From its the dependency graph shown in Figure 5.1, the path between *tool* and *hammer* generates the patterns shown in Table 5.1. The dependency relations inserted between the subject and the object of a verb in a sentence are represented using dashed red lines.

$screwdriver \xrightarrow{\text{conj}} X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N \xleftarrow{\text{mod}} other_J$
$screwdriver \xrightarrow{\text{conj}} X_N \xleftarrow{\text{obj}} V \xrightarrow{\text{obj}} Y_N \xleftarrow{\text{mod}} other_J$
$screwdriver \xrightarrow{\text{conj}} X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N$
$screwdriver \xrightarrow{\text{conj}} X_N \xleftarrow{\text{obj}} V \xrightarrow{\text{obj}} Y_N$
$X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N \xleftarrow{\text{mod}} other_J$
$X_N \xleftarrow{\text{obj}} V \xrightarrow{\text{obj}} Y_N \xleftarrow{\text{mod}} other_J$
$X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N$
$X_N \xleftarrow{\text{obj}} V \xrightarrow{\text{obj}} Y_N$

Table 5.1: Examples of dependency patterns generated from the dependency path $screwdriver \xrightarrow{\text{conj}} hammer_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} tool_N \xleftarrow{\text{mod}} other_J$

Feature Selection Table 5.2 shows the total number of patterns acquired from each corpus. As observed, the number of patterns vary across domains and increase with the size of the corpus. Therefore, only the most meaningful patterns must be acquired in $\mathbb{P}_P$ to create the vectorial space. To select the most meaningful patterns, they are ranked based on the amount of information provided for the relations from *R*, and the top t_P number of patterns are included in the final set of patterns $\mathbb{P}_P$. To avoid overfitting, the patterns that do not co-occur with at least two positive instances are eliminated.

In order to rank the patterns according to their importance for each relationr, two scores are tested: the number of unique word pairs co-occurring in the pattern in the corpus *ufr score*, and the *tf-idf score*. In the former scoring scheme the most meaningful patterns are those that co-occur more often with example instances from *E*. The latter score scheme is calculated for each pattern regarding each lexical-semantic relation. Only the highest *tf-idf score* obtained is associated to each pattern.

$$tfidf(p_i) = max_j(\frac{log(uniq(p_i, r_j) + 1) * |R|}{|R_p|})$$

where p_i is a pattern of context, *uniq(*p_i*,* r_j*)* is the number of unique in-

	BNC		WIKI	
	$l_E = 2$	$l_E = 3$	$l_E = 2$	$l_E = 3$
AMPH_REPT	71	199	5261	9534
APPLIANCE	733	1425	16726	24663
BIRD	246	668	36179	485189
BUILDING	1686	3141	53320	79253
CLOTHING	2266	4255	23718	37918
CONTAINER	728	943	4591	4923
FRUIT	756	1682	18223	46234
FURNITURE	1144	2179	4875	6987
GR_MAM	1934	5073	76152	235300
INSECT	178	329	6374	16612
MUS_INSTR	343	652	54197	126520
TOOL	405	662	4269	5510
TREE	145	247	8398	14195
VEG	1159	2561	18282	44503
VEHICLE	2477	4168	77722	93377
WATER_ANIM	213	454	4885	8977
WEAPON	507	977	32165	42505

Table 5.2: The number of dependency patterns used in PaCE for $l_P = 2$ and $l_P = 3$ for each corpora.

stances of the relation r_j occurring in the pattern p_i and $|R_p|$ is the number of relations r_j whose example instances are seen occurring in the pattern p_i.

Finally, the set $\mathbb{P}_\mathrm{P}$ forms a vectorial space, with each pattern corresponding to a separate dimension.

Word Pair Representation To classify a word pair *(x,y)* as an instance of a lexical-semantic relation, the word pair is first represented as a vector of features. Each pattern previously acquired represents a dimension and the total set of patterns creates a vector space. The semantic relation holding between each word pair is represented by a vector encoding on each position *i* the logarithm of the frequency of co-occurrence in corpus of *x* and *y* in the i^{th} pattern considered.

As mentioned earlier, our approach relies on the assumption that any sentence in which a pair of words co-occurs is likely to provide information regarding the lexical-semantic relation holding between them. Naturally, this does not necessarily hold for all the patterns of context acquired, as some of the features may not express any semantic relation of interest. For instance, the sentence *I can feel my fingers and close my hand.* contains the relation instance *((hand,finger),meronymy)* but the context does not provide any information regarding the relation holding between *hand* and *finger*. The machine learning algorithm is able to discover which features are noisy, i.e. not informative for the task at hand, and associates corresponding weights for minimising errors in classification results.

5.1.1 Parameter Selection

The results of the PaCE system, like other pattern-based systems, depend on a series of parameters, whose value affects the word pair representation: the length of the dependency paths, l_E, the number of patterns used to create the vectorial space, t_P, and the selection scheme of the most meaningful patterns, *tfidf* or *ufr*. To show the effect of each parameter over the classification results we test the following values: $\mathrm{l_P}$ = 2 and $\mathrm{l_P}$=3, and $\mathrm{t_P}$=5000, $\mathrm{t_P}$=8000 and $\mathrm{t_P}$=10 000.

Table 5.3 shows the results of the PaCE system with different parameter values. Over BNC, only the importance of l_P is tested because as it is showed in the Table 5.2, the number of acquired paths from almost all the topical domains does not overpass 5000 patterns. Over Wikipedia, all the parameters are tested.

Although we expected $\mathrm{l_P}$ = 3 to obtain significantly better results in recall compared with $\mathrm{l_P}$=2, the results show that there is no statistical sig-

		BNC	Wikipedia					
			fr			$tfidf$		
		5000	5000	8000	10000	5000	8000	10000
	P	78.1	74.6	72.9	71.7	73.7	72.4	71.9
l_E=2	R	50.5	68.1	67.4	66.2	66.4	65.7	65.6
	F	61.4	71.2	70	68.8	69.9	68.9	68.6
	P	77.1	72.8	71.1	70.1	70.7	70.4	70.4
l_E=3	R	51.1	67.5	67	66.5	63.6	63.3	63.3
	F	61.4	70.1	69	68.2	67	66.7	66.7

Table 5.3: Results of the PaCE system when initialised with different values for its three parameters: l_E, t_P and the selection scheme, ufr or $tfidf$.

nificant difference[1] between these two results. Therefore, the most meaningful paths connecting the members of lexical-semantic instances have two edges or less. We suppose these findings are due to the optimisations applied over the results of the parser: the propagation of the information on the members of a conjunction, and the introduction of the edges vb_V that connect the subject with the object of the verb *V*. This way, words that may be semantically related are directly connected. Comparing the results obtained for various values for t_P, it can be observed a decrease in the results with the increase of the number of features. Patterns that occur less frequently are not representative for any of the target relations. The same outcome is observed when comparing the ranking schemes. The most effective scheme, which highlights the most reliable patterns for the lexical-semantic acquisition, is *ufr*, showing that the most reliable patterns tend to co-occur with many instances.

[1] Statistical significance was calculated using Student's t-test with a 95-percent confidence interval.

5.1.2 Dependency Information and PoS-Generalization

PaCE was created to show the precision and recall of the pattern-based approaches for the acquisition task. Two strategies that may potentially improve the recall of patterns systems are encapsulated in PaCE: the use of all the dependency paths between two words up to l_P edges instead of only the shortest ones; and the incorporation of a generalisation strategy of the features used in classification based on part-of-speech information. To provide more place for generalisation of the patterns, in this experiment we use $l_P = 3$, despite it achieves slightly lower results over PaCE, $t_P = 5000$ and *ufr* as the selection scheme. To shed light on the impact of these strategies combined in the PaCE classification system, it is compared with five other systems described below.

PairClass We reimplemented the PairClass algorithm [Turney, 2008b], which provides a state-of-the-art pattern-based approach for classifying the relationship between word pairs and has performed well for many relation types. Using a set of seed pairs *(x, y)* for each relation, PairClass acquires a set of lexical patterns using the template:

$$[0 \text{ to } 1 \text{ words}]\ x\ [0 \text{ to } 3 \text{ words}]\ y\ [0 \text{ to } 1 \text{ words}]$$

From the initial set of lexical patterns extracted from a corpus, additional patterns are generated by basically generalising each word to its part-of-speech. For *N* seed pairs, the most frequent *kN* patterns are retained. We follow [Turney, 2008b] and set *k=20*. The patterns retained are then used as features to train an SVM classifier over the set of possible relation types. Finally, we underline that in the original experiments, PairClass was trained using a corpus of 5×10^{10} words, which is three orders of magnitude larger than the BNC corpus used in our experiments.

ShortestPaths The ShortestPaths system is a similar system to the one developed by [Snow et al., 2004]. This approach uses the shortest paths connecting two target words in the dependency graph of a sentence as features for the classification of word pairs. As for the PaCE system, for each target word in the path, a neighbouring window node can be added to the dependency path. For a correct comparison, the ShortestPaths system uses the same classifier as the PaCE system.

ShortestPaths$^{\mathbf{GEN}}$ For highlighting the importance of the PoS-generalization strategy, we created a variant of the ShortestPaths system that uses the same generalisation strategy, but relies on the features as the ShortestPaths system.

PaCE$^{\mathbf{LEX}}$ For evaluating the specific contribution of considering all the dependency paths between two words instead of using only the shortest ones, we create a variant of the PaCE system, PaCE$^{\mathrm{LEX}}$, that uses as features only lexicalized dependency paths up to l_E edges.

Patterns For completeness, we also constructed a composite system using Hearst patterns [Hearst, 1992] for detecting hypernyms and co-hyponyms, and Bearland patterns [Berland and Charniak, 1999] for detecting meronyms. Classification is performed by measuring the frequency in corpus data of each relation pattern and then selecting the relation whose patterns occur more frequently. This system classifies only hypernyms, co-hyponyms and meronyms because no standard manually-designed patterns exist in the literature for the two remaining lexical-semantic relations considered in our experiments.

For a better understanding of the information used by each systems, an outline of the compared systems is shown in Table 5.4.

	Dependency Info.	Generalisation Tech.
PairClass		X
ShortestPaths	Shorthest	
ShortestPaths$^{\mathrm{GEN}}$	Shortest	X
PaCE$^{\mathrm{LEX}}$	All	
PaCE	All	X

Table 5.4: The outline of the information contained in all the systems compared in this section that use automatically acquired patterns of contexts: the type of dependency information used, only the shortest paths or all the paths, and if it uses a generalization technique to find similar patterns.

Evaluation

The overall performance of each system assessed for each relation included in the BLESS dataset is shown in Table 5.5 and Table 5.6. All the systems relying on dependency information achieve better results than the PairClass system, which does not use this type of information. This comparison shows the impact this type of information has both on precision and recall across all the relations tested. The 8.1 point increase in precision over BNC and 19 points over Wikipedia achieved by the ShortestPaths mirrors the greater reliability of patterns based on dependency information when compared against surface patterns. Additionally, the ShortestPaths scores 12 and 14 points higher in recall, respectively, a difference that is attributable to the fact that these patterns go beyond the three word window used by PairClass, and therefore provide a larger amount of the data available in corpus to the classification system.

Comparing the PaCE$^{\mathrm{LEX}}$ system that uses all the dependency paths up to three edges with the ShortestPaths system that uses only the shortest paths no statistically significant differences are observed in the results over both corpora. Therefore, even when only the shortest dependency paths are considered, the most important patterns of co-occurrence are already gathered.

		Attr	Coord	Act	Hypo	Mero	ALL
Patterns	P		98.9		58.1	70.2	83.8
	R		31.8		10.8	18.9	23.3
	F		48.1		18.2	29.7	36.5
PairClass	P	80.8	76.3	73.5	25.4	67.4	66.8
	R	36.8	29.8	46.4	21.9	33.5	35.6
	F	50.6	42.9	56.9	23.5	44.7	46.4
ShortestPaths	P	83.9	78	78.7	88	75	78.9
	R	47.5	45.5	64.7	13.2	43.7	47.6
	F	60.6	57.4	71	22.9	55.3	59.4
ShortestPaths$^{\text{GEN}}$	P	82.1	78.4	79.1	85.9	73	78.4
	R	50.3	47.5	65	13.7	46.2	49.3
	F	62.4	59.2	71.3	23.6	56.6	60.5
PaCE$^{\text{LEX}}$	P	84.3	80.5	78.3	88.4	74.8	79.4
	R	46.9	46.9	64.6	14.2	44	48
	F	60.3	59.3	70.8	24.5	55.4	59.8
PaCE	P	79.2	82	78.2	73	69.3	77.1
	R	55.5	47.4	65.3	17	48.3	51.1
	F	65.3	60.1	71.2	27.5	56.9	61.4
UL-Recall	R	74	64.7	83	60	73.9	72.8

Table 5.5: The results obtained by the systems considered, across all relation types on BNC. The last line presents the upper-limit for recall (UL-Recall) (see Section 7).

Comparing ShortestPaths$^{\text{GEN}}$ with the ShortestPaths system, the importance of this generalisation strategy is highlighted: over BNC the recall results improve by 1.7 points, while over Wikipedia, no improvements are observed.

PaCE system mades apparent the impact of using all dependency paths up to three edges in combination with the PoS-generalisation strategy in the recall scores. Over BNC, PaCE achieves 3.5 points increase in recall, although the features used are apparently less reliable as the precision drops 2.3 points. Over Wikipedia, both the precision and the recall decrease with 1.6 points and 1.2 points, respectively.

		Attr	Coord	Act	Hypo	Mero	ALL
Patterns	P		98.4		76.4	57.5	81.1
	R		58		34.4	27.5	42.5
	F		73		47.4	37.2	55.8
PairClass	P	82.3	80.9	70.5	21.4	65	60.9
	R	45.4	57.1	63.4	50.1	48.5	54.1
	F	58.5	66.9	66.8	29.9	55.6	57.3
ShortestPaths	P	78.1	82	77.7	89.3	79.8	79.9
	R	71	74.5	72.8	42.4	63.1	68.1
	F	74.4	78.1	75.1	57.5	70.5	73.5
ShortestPaths^{GEN}	P	79.4	84.1	76.8	89.6	81.1	80.7
	R	70.9	75.9	73.5	42.7	63.3	68.6
	F	74.9	79.8	75.1	57.9	71.1	74.2
PaCE^{LEX}	P	78.8	86.4	77.7	91.5	80.2	81.3
	R	68.5	75.4	73.7	43.2	64.1	68.4
	F	73.3	80.5	75.6	58.7	71.2	74.3
PaCE	P	78.5	82.2	77.4	86.9	79	79.7
	R	66.9	76.1	72.4	37.8	63.2	67.2
	F	72.3	79	74.8	52.7	70.2	72.9
UL-Recall	R	84.4	85.9	92.9	79.1	90.8	87.8

Table 5.6: The results obtained by the systems considered, across all relation types using Wikipedia corpus. The last line presents the upper-limit for recall (UL-Recall) (see Section 7).

This experiment concludes that when the corpus is smaller and cannot provide a sufficient amount of information, like BNC, the two techniques combined, the PoS-based generalisation technique with the usage of all the dependency paths, are useful. Instead, over Wikipedia, a corpus 15 times larger, no system stands out by its results indicating that when combined these two techniques, they result in a less accurate system.

These experiments show the motivations of this work: using only patterns of context of co-occurrences the results are limited in recall. Adding more patterns from corpus or generalising them using part-of-speech the lack of information is not overcome, and therefore, for these techniques the only way to increase the results is by providing more corpus.

5.2 Error Analysis

A major point of discussion raised by the experiments conducted in this section is the main limitation of pattern-based systems: word pairs are classified based on evidence based on their co-occurrence in the same sentence. This way, the number of word pairs occurring in the same sentence in the input corpus constitutes the real upper-limit in recall (UL-Recall) of these approaches (see the last line in Table 5.5 and Table 5.6) in terms of recall. In our experiments, only 72.8% of the relation instances included in BLESS co-occur at least once in the same sentence in the BNC corpus and 87.8% in the Wikipedia. Being so, all the remaining candidate pairs cannot be classified due to a total lack of information regarding their semantic relations. Moreover, this is not a corpus-specific limitation, since due to the zipfian distribution of words, in any corpus of any size there will always be word pairs that will not co-occur in corpus to provide enough information to classifiers or to any automatic system pattern-based. For instance, 35% of the word pairs that are misclassified by PaCE using BNC corpus co-occur less than 5 times in the same sentence.

This problem has been addressed in the literature by combining approaches based on patterns extracted from co-occurrence contexts with semantic similarity between words [Turney, 2006b, Herdağdelen and Baroni, 2009]. The results obtained in solving analogies from the SAT test[2] show that the best performances were achieved by systems highly dependent on a very large corpus of ~50Gb [Turney, 2006b, Turney, 2013, Turney, 2006a, Turney, 2008b].

In order to scale down the dimensions of the input corpus, new techniques are necessary to go beyond the aforementioned upper-limit imposed

[2] http://aclweb.org/aclwiki/index.php?title=SAT_Analogy_ Questions_(State_of_the_art)

by the amount of co-occurrences in the same sentence observed in a given corpus, and to acquire more information to identify the relation holding between pairs of words that co-occur very infrequently or not at all in the same sentence in a given corpus, although they are related.

In order to get some insight on possible strategies for tackling this limitation of pattern-based classification systems for relation instances, an empirical error analysis of PaCE results over BNC was run, in which it is observed that 13% of the dataset does not co-occur in the same sentence. Yet, the members of these word pairs share a significant number of co-occurring words in the same type of dependency relation. In Figure 5.2 we present instances of co-hyponymy, hypernymy and attribute relations from our dataset that do not co-occur in the corpus, although they are densely linked by shared distributional properties, that we will call "bridging words".

Beet$_\mathrm{N}$ and cucumber$_\mathrm{N}$ are two co-hyponyms, both hyponyms of vegetable, which do not co-occur in the same sentence in our corpus. However, both occur in a relation of coordination with other hyponyms of vegetable such as potato$_\mathrm{N}$, carrot$_\mathrm{N}$, lettuce$_\mathrm{N}$ and pea$_\mathrm{N}$. Additionally, both target nouns occur as the direct object of the verb grow$_\mathrm{V}$, while only beet$_\mathrm{N}$ occurs as direct object of the verb sprout$_\mathrm{V}$. However, salad$_\mathrm{N}$, onion$_\mathrm{N}$ and raddish$_\mathrm{N}$ do occur as a direct object of the verb sprout$_\mathrm{V}$ and they occur in coordination with cucumber$_\mathrm{N}$.

Regarding the relation of hypernymy, cow$_\mathrm{N}$ and herbivore$_\mathrm{N}$ are an example of a word pair from the domain of *ground mammals* which instantiate this relation but does not co-occur in the same sentence in BNC, although both nouns occur in conjunction with other herbivores such as antelope$_\mathrm{N}$, goat$_\mathrm{N}$, sheep$_\mathrm{N}$, donkey$_\mathrm{N}$, horse$_\mathrm{N}$ and cattle$_\mathrm{N}$. This set of words co-occurs with herbivore$_\mathrm{N}$ in the pattern *herbivore such as X* or *herbivore is a X*. Also, both *cow* and *herbivore* occur with phrases like *group of X*, *number of X*, *population of X* and *herd of X*, as well as objects of the verb graze$_\mathrm{V}$.

Finally, in our dataset, fancy$_\mathrm{J}$ is an attribute of glove$_\mathrm{N}$ but these words do not co-occur in the same sentence in BNC. In our corpus, however, fancy$_\mathrm{J}$ is the modifier of boot$_\mathrm{N}$, coat$_\mathrm{N}$, dress$_\mathrm{N}$, hat$_\mathrm{N}$, jacket$_\mathrm{N}$ and tie$_\mathrm{N}$, among many others, which are words that co-occur in conjunction with the word glove$_\mathrm{N}$. Additionally, all of these words are objects of the verb to wear$_\mathrm{V}$ and they occur in the phrase *dressed in X*, just like glove$_\mathrm{N}$ does. Although fancy$_\mathrm{J}$ and glove$_\mathrm{N}$ do not co-occur, fancy$_\mathrm{J}$ co-occurs with colorful$_\mathrm{J}$, which is an adjectival modifier of jacket$_\mathrm{N}$ and dress$_\mathrm{N}$, two words that occur in a coordination relation with glove$_\mathrm{N}$.

Therefore, for detecting new relation instances, valuable information can be extracted from parallel sentences separately containing occurrences of each individual member of a candidate pair and where the members of the pair share other co-occurring words. The vectorial representation of each word in Traditional VSM captures this information but it is unclear how to use it to assess whether a given lexical-semantic relation holds or not for a candidate word pair. And yet, as made apparent by the examples above, dependency relations between individual target words and shared third party co-occurring words can be indeed a valuable indication of the relation holding between a word pair.

5.3 Conclusions

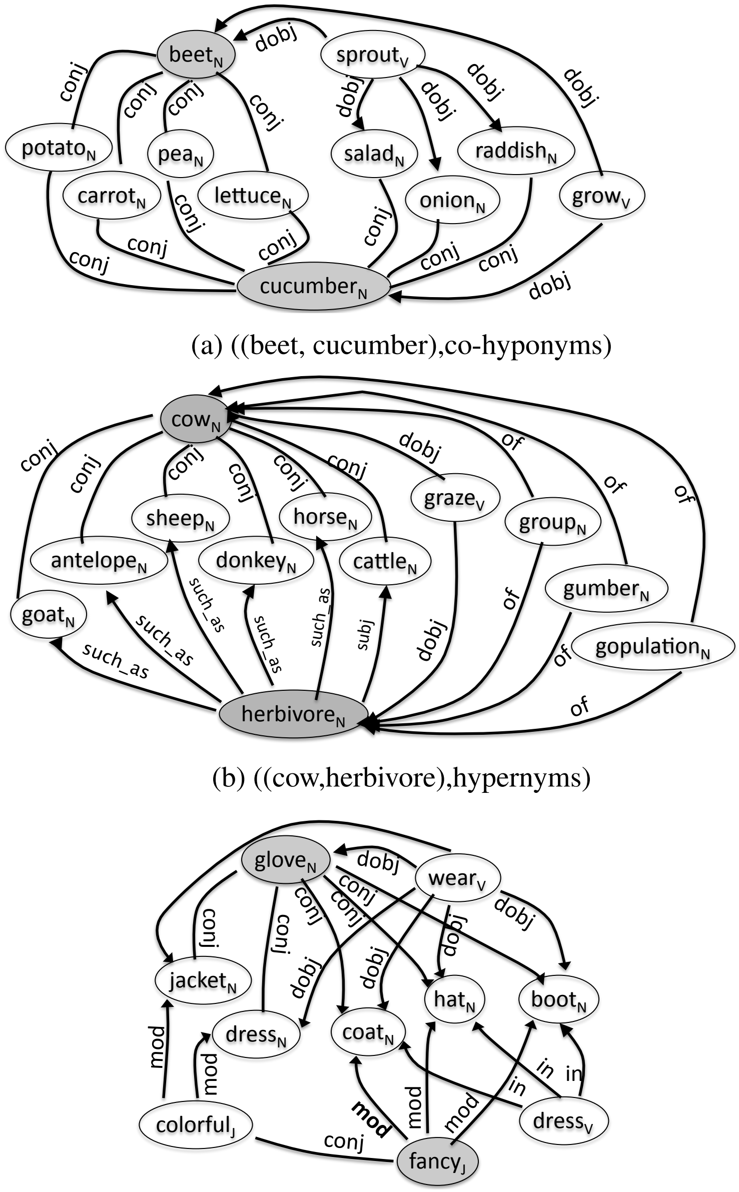
The present chapter shows the limits of pattern-based systems in the classification of instances of lexical-semantic relations. Moreover, it addressed the impact of using different information, in particular all the dependency paths, instead of only the shortest one, in combination with a generalisation strategy using part-of-speech information in the classification of instances of lexical-semantic relations.

The isolated use of all the dependency paths up to three edges has not resulted in any statistically significant improvement, when compared with using only the shortest paths.

On a smaller corpus, like BNC, PoS-generalized patterns of co-occurrence allow for correctly classifying more candidate word pairs and when combining these two strategies the recall further increases, but the precision slightly drops. On Wikipedia, the ShortestPath system obtained the best results, with a recall of almost 70%. Therefore, this strategy was found not useful for improving the acquisition recall. From this results, it is proved that the most important variable that affects recall in this approach is the corpus size.

However, note that comparing the best recall achieved over BNC, 51.1%, with the best recall achieved over Wikipedia, 68.6%, one observes that the 17.5 point increase in recall is obtained using a corpus 15 times larger.

The results obtained from the error analysis made apparent the need of overcoming the lack of enough patterns of contexts to fill in the vector that is handled by the classifier. As explained in Section 2.4, the sparse data problem is an important challenge for vector-based word representations. Our proposal is to gather information about the members of the word pair, in particular distributional information represented in terms of dependency relations. In this way, shared co-occurring contexts can be the basis of a generalisation technique that bridges information coming from different sentences, for improving the classification recall. This motivates the main goal of this thesis, to find novel word pair representations that embed information about word co-occurrences with the distributional information of words.



(a) ((beet, cucumber),co-hyponyms)

(b) ((cow,herbivore),hypernyms)

(c) ((glove,fancy),attribute)

Figure 5.2: Local networks of two elements unrelated in corpus

Chapter 6

OVERCOMING DATA SPARSITY

While the limitations of traditional pattern-based systems were presented in the previous chapter, this chapter introduces two new approaches for representing word pairs in order to accurately classify them as instances of lexical-semantic relations – even when the pair members do not co-occur. Our assumption is that the novel representation has to combine distributional information of words as well as information about the contexts of co-occurrences of word pairs.

The first approach creates a word pair representation based on a generalisation of the corpus information into a graph representation. It acts as a network of information where the words are connected in various dimensions, also interconnected, contrary to the dimensions of the Traditional VSM that act as independent features. The graph encodes the distributional behaviour of each word in the pair and consequently, patterns of co-occurrence expressing each target relation are extracted from it as relational information.

The second approach uses vectorial representations in Predictive VSM. These representations have been shown to preserve linear regularities among words and pairs of words, therefore, encoding lexical and relational similarities [Baroni et al., 2014]. Here, we test if their compositionality can represent also lexical-sematic relations instances.

6.1 Graph-based Representation Model

The present section introduces a novel word pair representation model based on patterns of contexts, and on a graph-based representation of the corpus. A word pair is represented as a vector of features set up with the most meaningful patterns of context and filled in with information extracted from the graph representation of the dependency parsed corpus. We refer to systems trained with these graph-based representations as **Gra**ph-based **C**lassification syst**E**m (GraCE).

The novelty of this system stands in the generalisation obtained with the graph-based corpus representation. Each word in a dependency-parsed corpus is represented by a node in a graph. Arcs connect two nodes when their corresponding words are connected in a dependency relation in a sentence in corpus; arcs are labeled with the name of the dependency relation. Its main advantage is that all the dependency relations of a target word, extracted from different sentences, are incident arcs to its corresponding node in the graph. Thus, words that never co-occur in the same context in corpus, are linked in the graph through bridging words: words that appear in a dependency relation with each member of the pair but in different sentences. With this representation we address the data sparsity issue, aiming to overcome the reported major bottleneck of previous approaches: limited recall because information can only be gathered from co-occurrences in the same sentence of two related words.

The creation of the word pair representations contains two stages: (1) the representation of the input corpus as a graph, and (2) the extraction of meaningful patterns of co-occurrence for each semantic relation from the same corpus starting from the initial set of examples. The graph-based representation and the set of acquired patterns are used to create vectorial representations of target word pairs.

Next, it is presented an example of how the graph representation of the corpus addresses the lack of information in corpus and it formally introduces each step of the GraCE algorithm.

Example To illustrate the benefit of acquiring information about a word pair from the graph instead of using co-occurrence information, let us consider that, given the sentences (S6.1) and (S6.2) below, we want to classify the pair *(chisel, tool)* as an instance of the relation of hypernymy.

(S6.1) The students learned how to handle screwdrivers, hammers and other tools.

(S6.2) The carpenter handles the new chisel.

S6.1	S6.2
det(students-2, The-1)	det(carpenter-2, The-1)
nsubj(learned-3, students-2)	nsubj(handles-3, carpenter-2)
root(ROOT-0, learned-3)	root(ROOT-0, handles-3)
advmod(handle-6, how-4)	det(chisel-6, the-4)
mark(handle-6, to-5)	amod(chisel-6, new-5)
ccomp(learned-3, handle-6)	dobj(handles-3, chisel-6)
dobj(handle-6, screwdrivers-7)	
dobj(handle-6, hammers-9)	
conj_and(screwdrivers-7, hammers-9)	
cc(screwdrivers-7, and-10)	
amod(tools-12, other-11)	
dobj(handle-6, tools-12)	
conj_and(screwdrivers-7, tools-12)	

Table 6.1: The dependency relations returned by the Stanford parser in the collapsed format.

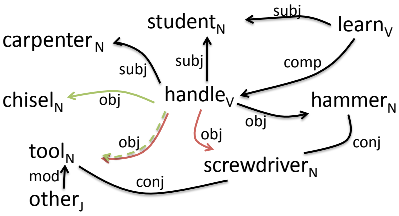


Figure 6.1: Dependency multigraph built from a two sentence corpus using GraCE. See text for details.

The word pair *(chisel, tool)* has a relation of hypernymy but its members do not co-occur in the same sentence. However, both words occur as objects of the verb *to handle* in different sentences, just like other hypernym word pairs such as *(hammer, tool)* and *(screwdriver, tool)* which do co-occur in the same sentence and are connected by the dependency patterns: $X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N$ and $X_N \xleftrightarrow{\text{conj}} Y_N$[1] in the graph. This shows that *handle* is one of the contexts shared between these semantically related words that provide information regarding a possible semantic relatedness between them. Leveraging only the information provided by each sentence, as existing pattern-based approaches do, no evidence is acquired regarding the semantic relation holding between *chisel* and *tool*. GraCE combines the dependency relations seen in each sentence, listed in Table 6.1, in the graph shown in Figure 6.1. In this graph, *chisel* and *tool* are connected by a path passing through the bridging word *handle*:

$$chisel_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} tool_N$$

This path shows that both *chisel* and *tool* could co-occur in a sentence as objects of the verb *to handle*, although they do not in the example two-

[1] The conjunction is marked with a bidirectional arrow because it is a bidirectional relation; in the graph, for each relation of conjunction $conj(w_1, w_2)$, we add two arcs $(w_1, conj, w_2)$ and $(w_2, conj, w_1)$

sentence corpus.

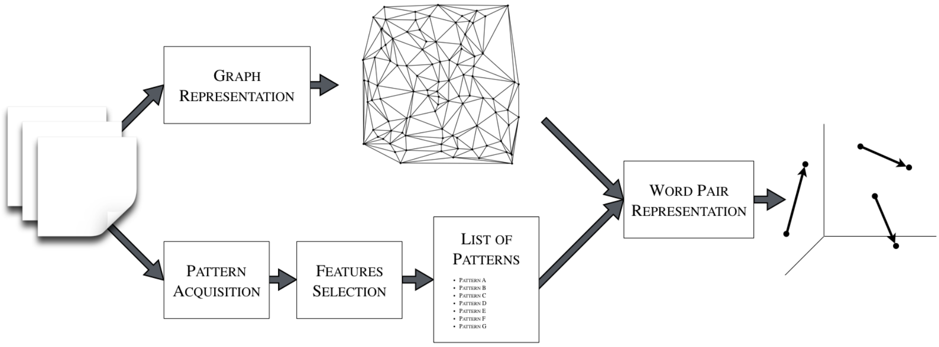


Figure 6.2: GraCE architecture.

In the following paragraphs we describe the algorithm for representing the relation holding between a word pair, using the graph based representation. As it can be observed in Figure 6.2, the corpus is used as input of two stages: to create a graph representing the connectivity of words based on their dependency relations, and to extract dependency-based patterns of contexts describing target lexical-semantic relations as meaningful features. The extraction of the patterns is split in two parts: the acquisition of the patterns and the selection of the most meaningful patterns. Finally, to create the word pair representations, information from the graph is combined with the final set of patterns.

Corpus representation The goal of the first step is to generate a directed graph connecting semantically associated words using observed dependency relations. Because two words may hold more than one dependency relation, two nodes may be linked by more than one arc. Formally, the corpus is represented as a graph $\mathbb{G} = (V, E)$, where V is a set of PoS-tagged lemmas in a corpus and E is the set of dependency relations connecting two lemmas from V in the corpus. From each parsed sentence of the corpus, a set of dependency relations linking the words in it is produced: $D = \{d_1 \ldots, d_{|D|}\}$, where $d = (w_i, dep, w_j)$ and w_i, w_j and dep denote PoS-tagged lemmas and a dependency relation, respectively. The graph $\mathbb{G}$ is created using all the dependency relations from D as arcs and their relation names al labels.

The output of this step is a labeled directed multigraph, where two words are connected by the set of arcs containing all the dependency relations holding between them in the corpus. See Figure 6.1 for an example of the graph.

Pattern Acquisition The goal of the second step is to collect features associated with each relation r from the parsed input corpus. The same approach is followed as in the acquisitions of patterns for PaCE, with the difference that these patterns are created only from the dependency paths that connect x and y in the dependency graph of each sentence. A dependency path is transformed in a pattern by delexicalising *x* and *y*. The PaCE system follows a pattern generalisation step based on PoS tags. Contrarily, the GraCE system uses only lexicalized patterns because the generalisation of the information is embedded in the structure of the corpus graph-based representation.

Given the example instances for the relation of hypernymy as in the previous example, i.e. the word pairs *(hammer, tool)* or *(screwdriver, tool)* from the sentence (S6.1), the system acquires the dependency patterns:

$$X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N$$

$$X_N \xleftrightarrow{\text{conj}} Y_N$$

as indicative dependency patterns for the relation of hypernymy.

Feature Selection Table 6.2 shows the total number of patterns acquired from each corpus. As in the PaCE system, for the word pair vectorial

	BNC		Wiki	
	$l_E = 2$	$l_E = 3$	$l_E = 2$	$l_E = 3$
Amph_rept	349	805	7824	51463
Appliance	1605	4327	39089	299399
Bird	661	1669	8693	61381
Building	5419	14837	77402	561877
Clothing	2278	6457	26551	186772
Container	1683	3858	18083	104018
Fruit	635	1820	8116	65802
Furniture	2053	5384	12729	79356
Gr_mam	3137	8532	42980	307647
Insect	455	1185	6011	39355
Mus_instr	707	1944	29526	227418
Tool	847	2137	12339	73682
Tree	421	1047	6617	50358
Veg	831	2763	9288	77872
Vehicle	4120	10413	116444	777800
Water_anim	606	1714	7646	51248
Weapon	1128	2891	51821	339683

Table 6.2: The number of dependency patterns acquired for GraCE$_2$ and GraCE$_3$ for each corpora.

representation, only the top t_G most meaningful patterns are considered in the final set of patterns $\mathbb{P}_\mathrm{G}$ to form a vectorial space. The features are selected based on the number of occurrences with unique word pairs, *ufr score*, or the their *tfidf scores* associated, as they were presented in Section 5.1.

Word pair representations Using the graph model $\mathbb{G}$ and the set of contextual patterns automatically acquired $\mathbb{P}_\mathrm{G}$, each word pair *(x,y)* is represented as a binary distribution over each pattern from $\mathbb{P}_\mathrm{G}$. Rather than using the raw input corpus to identify contexts of occurrence for the word pair *(x,y)* as PaCE does, GraCE uses the paths connecting *x* and *y* in $\mathbb{G}$.

These paths are then matched against the feature patterns from $\mathbb{P}_{\mathrm{G}}$, and the word pair *(x,y)* is represented as a binary vector encoding non-zero values for all the features connecting its members in $\mathbb{G}$, and zero otherwise. Binary weights are used because the feature values are derived from observing paths in the graph, which is a generalization of the corpus. The graph contains combinations of multiple dependency relations, extracted from various sentences, therefore, paths not observed in the corpus can be found in the graph.

	$X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N$	$X_N \xleftrightarrow{\text{conj}} Y_N$
(screwdriver, tool)	1	1
(hammer, tool)	1	1
(chisel, tool)	1	0

Table 6.3: Vectorial representations created with GraCE for (screwdriver, tool), (hammer, tool) and (chisel, tool) using the example corpus and the observed dependency patterns in the corpus.

6.1.1 Parameter Selection

In the previous lines we described the way GraCE creates the word pair representations. Next, we present a set of experiments used to set up the parameters for GraCE.

GraCE, as PaCE, contains a set of parameters that influence the patterns acquired in the $\mathbb{P}_G$, to create the final word pair representations. Pattern selection parameters refer to the maximal length of the dependency patterns, l_G, the maximal number of the patterns used t_G, and the selection scheme *ufr* or *tfidf*. $\mathrm{l_G}$ was tested for the values $\mathrm{l_G} = 2$ and $\mathrm{l_G} = 3$, larger path connecting to many words in the graph, and $\mathrm{t_G}$ was tested for the values $\mathrm{t_G} = 5000$, $\mathrm{t_G} = 8000$ and $\mathrm{t_G} = 10000$.

To test the effect of these parameters, we trained and tested GraCE in the in-domain setup using each combination of their values and the results

		BNC	Wikipedia					
			ufr			$tfidf$		
		5000	5000	8000	10000	5000	8000	10000
l_E=2	P	74 1	81 3	63 7	58 4	74 1	65 3	61
	R	80	81 1	73	69 7	80	75 3	72 7
	F	76 9	81 2	68	63 5	76 9	70	66 4
l_E=3	P	87.5	84 5	82 3	79	75 1	63	57 7
	R	85	82 3	81 2	80 4	80 6	73 3	69 7
	F	86.3	83 4	81 7	79 7	77 7	67 7	63 1

Table 6.4: Aggregated results obtained with the in-domain setup for the GraCE system using different parameters.

are shown in Table 6.4.

On BNC, using only dependency patterns created with dependency relations up to two arcs, the system achieves only 74.1 points in precision, 80 in recall and 76.9 in F-measure. It is when the paths created with three dependency relations are used, that the results improve with 13.4 points in precision, 5 in recall and 9.5 in F-measure. Therefore, although many word pairs are connected in graph at distance two, this dependency path length does not provide sufficient distinctive information. When larger paths are used, novel information is provided to the classifier, information that increases the precision of the system.

Contrarily, on Wikipedia, a 15 times larger corpus, using only patterns of length two, the classifier is able to make quite accurate classifications and the system is already able to achieve 81.3 points precision, 81.1 points recall and 81.2 in F-measure. However, using patterns up to three arcs, the results increase with 3.2 points in precision, 1.2 points in recall and 2.2 in F-score.

On Wikipedia, we also tested also a feature selection scheme for finding the most reliable features for the word pair representation. Note that the features are dependency patterns extracted from the dependency paths connecting instances of target relations in corpus. In GraCE the information provided by the dependency pattern is amplified by the generalisation process from the graph representation. When the information of the pattern is not precise enough for the target relation, it can generate a lot of erroneous information in the word pair representation. For this reason it is very important to use informative patterns for the semantic relation holding between the two words in the pair. However, an equilibrium has to be found between the number of patterns, too few patterns connect less word pairs decreasing the recall of the system, but too many patterns could decrease the precision of the system.

Comparing the results obtained with the ufr selection scheme and the results obtained with $tfidf$, the best results are obtained when the relevance of a word pair is given by the number of unique co-occurring word pairs. Therefore the patterns that are the most representative for a semantic relation tend to co-occur with more instances of the relation.

Taking a look at the results obtained when the number of features used to create the word pair representations $\mathrm{t_G}$ vary, the best results are obtained with $\mathrm{t_G}$=5000. This shows that using more features, less precise information is provided to the classifier and it is harder to make a decision due to the graph based corpus generalisation.

Therefore, the most effective parameters for the system are obtained using the $\mathrm{l_G}$=3, $\mathrm{t_G}$=5000 and the number of unique word pairs occurring with a pattern used as selection scheme, on both corpora used. This setup is used for the rest of experiments.

6.1.2 Graph-based Corpus Generalisation

After the parameters are set up for GraCE, we present an experiment that shows the effect of the graph-based generalisation of the corpus over the classification results.

To shed light over the generalisation process, we compare the results of GraCE with a Baseline system:

BASELINE The purported benefit of the GraCE model is that the graph enables identifying syntactic features between pair members that are never observed in the corpus, which increases the number of instances that can be classified without sacrificing accuracy. Therefore, to quantify the effect of the graph, we include a baseline system, denoted BL, that uses an identical setup to GraCE but where the feature vector for a word pair is created only from the dependency path features that were observed in the corpus (as opposed to the graph). Unlike the GraCE model which has binary weighting (due to the graph properties), the baseline model's feature values correspond to the frequencies with which patterns occur; following common practice, the values are log-normalized. Following the previous example, BL system lacks of enough information to represent the *(chisel, tool)* word pair, because *chisel* and *tool* are not seen in the same sentence in corpus, and consequently this word pair cannot be classified.

Evaluation

The improvement due to the generalisation of the corpus in a graph representation is reflected in Table 6.5 and Table 6.6, for the in-domain setup.

The main issue of systems that only use the information directly observable in corpus, like BL, is that a word pair has to co-occur in the same sentence a sufficient amount of times to provide enough distributional information for classification. In BNC, 27.2 % of the BLESS instances never co-occur, while in Wikipedia, a corpus 15 times larger than BNC, this number amounts to 12.2 %. Mainstream pattern-based systems are unable to associate any features to word pairs that do not co-occur and, consequently, to classify them. Therefore, the real upper-bound limit of this type of systems is the amount of word pairs co-occurring in the corpus, presented in the *UL-Recall BL* line in each results table. The recall achieved by GraCE overcomes this limitation of pattern-based systems: 23.6% and 4.8% of the correctly classified word pairs are never seen in the same sentence in BNC and in Wikipedia, respectively. The actual upper-limits of the GraCE

		ATTR	COORD	ACT	HYPO	MERO	ALL
BL	P	83.7	82.1	79	86.9	71.2	79.1
	R	51.2	48.5	66.2	18.3	47.7	50.8
	F	63.5	61	72	30.3	57.2	61.8
GRACE	P	86.7	93	84	93.3	86.6	88
	R	85.3	89.7	88.6	72.8	81.6	85.4
	F	86	91.3	86.3	81.8	84	86.6
UL-RECALL - BL	R	74	64.7	83	60	73.9	72.8
UL-RECALL - GRACE	R	99.9	98.2	99.9	99.6	99.3	99.3

Table 6.5: The results obtained by the BL and GraCE, across all relation types using BNC. The last lines presents the upper-limit for the BL system recall (UL-recall BL) and the upper-limit for the GraCE system recall (UL-recall GraCE).

		ATTR	COORD	ACT	HYPO	MERO	ALL
BASELINE	P	79	86.9	77.5	92.5	80.8	81.6
	R	68.4	75.5	73.7	43.4	65.1	68.6
	F	73.3	80.8	75.6	59.1	72.1	74.5
GRACE	P	83.8	97.4	84.4	70	80.4	84.5
	R	72.8	82.7	85.9	84.4	84.8	82.3
	F	77.9	89.5	85.1	76.5	82.5	83.4
UL-RECALL BL	R	84.4	85.9	92.9	79.1	90.8	87.8
UL-RECALL GRACE	R	100	98.8	100	100	99.3	99.7

Table 6.6: The results obtained by BL and GraCE across all relation types using Wikipedia corpus. The last lines presents the upper-limit for the BL system recall (UL-recall BL) and the upper-limit for the GraCE system recall (UL-recall GraCE).

system represent the number of word pairs that have both members occurring in corpus, presented in the *UL-Recall GraCE*. This capability makes GraCE improve the BL performance by 8.9 points in precision and 34.6 points in recall on BNC and 2.9 points in precision and 13.7 in recall on Wikipedia. Given that the BL system is built identically to GraCE but without using a graph, these results demonstrate the performance benefit of joining the distributional information of a corpus into a graph-based corpus representation.

6.2 Word Embeddings Representation Model

The present section introduces word pair representations based on Predictive VSM, also known as word embeddings, which encode information about the similarity between words as well as information about the semantic relations holding between word pairs (see Section 2.3.2). In this vectorial space, the word pair representations are created starting from the vectorial representation of the word pair members. Each word in corpus is represented as a vector, named word embeddings. We use the original Skip-gram model, improved with techniques of negative sampling and subsampling of frequent words, which achieved the best results for detecting semantically similar words [Mikolov et al., 2013a, Mikolov et al., 2013b]. This model uses co-occurrences of words in contexts created with their position in a window context. However, the dependency relations between words could create more suitable representations for our task. Therefore, to test the usefulness of the dependency relations between words for the representation of the lexical-semantic relation instances in Predictive VSM, a second set of vectorial representations is created using a dependency-based Skip-gram model [Levy and Goldberg, 2014a]. Both systems filtered out all the words occurring only once in corpus, used a window of 5, and negative sampling with $k = 5$, to create the vectors and 200-dimensional vectors are learned for each word.

Once the word embeddings are created, word pairs are represented applying vectorial operations over the word representations. Two representations are tested: the *vector offset* and the *vector concatenation*, both described more formally below.

Previous works collected vectorial similarities between word pair representations to find instances of the same semantic relation. This way, the information encoded in all the vectorial dimensions is equally important for the classification. The novelty of our proposal stands in the usage of a supervised SVM classifier, as in PaCE and GraCE, to discover patterns of features in the word pair representation for labelling a word pair as an instance of a target relation. We refer to these systems **W**ord **E**mbedding **C**lassification syst**E**m (WECE). Below we introduce each variation of the WECE system.

WECE$_{\textbf{offset}}$ [Mikolov et al., 2013c] shows that the vectorial representation of words provided by word embeddings captures syntactic and semantic regularities and that each relationship is characterized by a relation specific vector offset. Word pairs with similar offsets can be interpreted as word pairs with the same semantic relation. Therefore, given a target word pair *(x,y)*, the vectorial representation is calculated from the difference between its vectors:

$$v((x,y)) = v(x) - v(y)$$

Note that this operation depends on the order of the arguments, and is therefore potentially able to capture asymmetric relationships.

By subtracting the vectorial representations of the members, this representation relies only on a vectorial representation of the relation and eliminates the individual information of each member of the pair.

WECE$_{\textbf{concat}}$ This representation concatenates the embeddings of each member of a pair to create the vectorial representation of the pair. Formally, given a word pair (x, y), whose member vectorial representations are $v(x) = (x_1, x_2, \ldots, x_n)$, and $v(y) = (y_1, y_2, \ldots, y_n)$ respectively, the vectorial representation of (x, y) is defined as the concatenation of $v(x)$ and $v(y)$:

$$v((x,y)) = (x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n)$$

Consequently, the length of $v((x, y))$ is $2n$, where n is the dimension of the embedding space.

The concatenation of the vectorial representations of words tests if the information encoded directly in the embeddings reflects the semantic relation of the word pair. This way, the classification system does not leverage only the information about the relation holding between the two words, but it may handle also possible similarities in their distributional representations.

The $\text{WECE}_{\text{offset}}$ systems that use the word embeddings created with the original system is referred to as $\text{WECE}^{\text{bow}}_{\text{offset}}$, and when the dependency based embeddings are used ,$\text{WECE}^{\text{dep}}_{\text{offset}}$. Similarly, we defined $\text{WECE}^{\text{bow}}_{\text{concat}}$ and $\text{WECE}^{\text{dep}}_{\text{concat}}$.

6.2.1 Automatic Feature Similarities Detection

Word embeddings have been used in the **D**irectional **S**imilarity Model (DS)[Zhila et al., 2013] to measure relational similarity in the Sem-Eval 2012 - Task 2 [Jurgens et al., 2012]. This task can easily be extended to a classification setting: given a target word pair (x, y), the similarity is computed between (x, y) and prototypical instances $(x, y)_i$ of a target relation r to decide if (x, y) is also an instance of r.

DS relies on the assumption that word pairs having the same semantic relations have similar offset vectors. Therefore, to approximate how much a word pair is representative as an instance of a semantic relation, they calculate the cosine similarity between the target word pair and the vectorial representations of the prototypical word pairs. Thus, DS system represents a minimally-supervised system whose features are produced in an unsupervised way, through the embedding process, and are therefore not necessarily tuned for the acquisition of lexical-semantic instances. The WECE systems are intended to identify potential benefits when adding feature selection by means of the SVM[2]. Therefore, to show the contribution of the machine learning system, WECE systems are compared with our implementation of the DS systems.

DS systems As for the WECE systems, the DS systems use two types of embeddings: the word embeddings produced using the original Skip-gram model which was originally used in [Zhila et al., 2013] and the embeddings using the method of [Levy and Goldberg, 2014a], which include dependency parsing information. We refer to these systems as $\mathrm{DS}_{\mathrm{Zhila}}$ and $\mathrm{DS}_{\mathrm{Levy}}$, respectively. To create the vectorial representations of a word pair, DS systems use the offset vectors.

To calculate the similarity of (x, y) with the set of instances $(x, y)_i$ of the relation r_i, each system computes the average of the similarity of (x, y) with each $(x, y)_i$:

$$avg(cos((x, y), (x, y)_i))$$

The word pair is classified as an instance of the relation with the highest associated score.

DS$^{\mathrm{c}}$ systems DS$^{\mathrm{c}}$ differs from the DS system only in the way the similarity is computed: $\mathrm{DS}^{\mathrm{c}}_{\mathrm{Zhila}}$ and $\mathrm{DS}^{\mathrm{c}}_{\mathrm{Levy}}$ calculate first the centroid of the example instances as vectorial representation of the relation instances; the final similarity score is the similarity of (x, y) with the centroid of the class:

$$cos((x, y), centroid((x, y)_i))$$

[2] A similar system was created in the same time in another work [Levy et al., 2015b], but without any connection to our work. However, [Levy et al., 2015b] tests a supervised system only for the detection of the hypernyms.

	BNC			Wikipedia		
	P	R	F	P	R	F
DS_{ZHILA}	62.1	47.4	53.7	61.9	49.1	54.8
$\text{DS}^{\text{C}}_{\text{ZHILA}}$	66.4	63.6	65	66.6	64.1	65.3
DS_{LEVY}	53	49.2	51	61.3	54.5	57.7
$\text{DS}^{\text{C}}_{\text{LEVY}}$	55.3	58	56.6	68.5	64.9	66.6
$\text{WECE}^{\text{BOW}}_{\text{OFFSET}}$	**90**	**90.9**	**90.4**	**91.9**	**93.8**	**92.8**
$\text{WECE}^{\text{BOW}}_{\text{CONCAT}}$	**89.9**	**91**	**90.4**	90.1	91.8	90.9
$\text{WECE}^{\text{DEP}}_{\text{OFFSET}}$	85.3	86.5	85.9	88	91.9	89.9
$\text{WECE}^{\text{DEP}}_{\text{CONCAT}}$	85.9	87	86.5	88.6	92.2	90.3
UL-RECALL	99.3			99.7		

Table 6.7: Aggregated results obtained for the in-domain setup with the BLESS dataset.

Evaluation

Table 6.7 shows the results of the systems based on the predictive vectorial representations. Taking a look at DS systems, we observe that the best results are achieved by comparing the target word pair representation with the centroid vector of the class, while the corpus information used to train the vectors is not distinctive. However, the general results are quite low, achieving less than 70 points precision and recall. Instead, WECE systems achieve good results, having a recall of 88.85 points in average showing an improvement of 34.3 points in average over the DS systems which used the same embeddings. This difference highlights the importance of the SVM classifier for learning which dimensions of the embeddings reflect the lexical relation.

Nevertheless, these results go against the statements of [Levy et al., 2015b] that good results obtained by the supervised systems applied over the word embeddings are due to the *lexical memorization*. We will discuss this issue in Section 6.3.2, dedicated to analyse the results across all the relations.

Comparing all the WECE systems between them, we observe that the best results over BNC are achieved by the WECE$^{\text{bow}}$ systems that are trained using the bag-of-words model. WECE$^{\text{dep}}$ which uses only co-occurrences created from dependency relations achieves lower results in both precision and recall. On Wikipedia, the best results are obtained with WECE$^{\text{bow}}_{\text{offset}}$, while WECE$^{\text{dep}}$ achieves lower results in precision.

Comparing the results obtained by WECE$_{\text{concat}}$, which concatenates the vectorial representations, and therefore leverages information from the vectorial representation of the words, and the results obtained by WECE$_{\text{offset}}$, which uses the vector offset as representation of the relation holding between two words, we observe that they achieve very similar results. The embeddings are created to assign similar representation to similar words, therefore, the results show that information about the lexical similarity between pairwise words is not useful for the acquisition of lexical semantic relations.

6.3 Experiments

The objective of this work is to find word pair representations that are able to model even the semantic relation holding between word pairs that do not co-occur in text. In the previous sections two new systems were introduced: GraCE and WECE, both using vectors encoding distributional information of words and information about the context where a word pair co-occurs. GraCE proposes a graph-based representation of the corpus to encode these information types, while WECE combines them in a vectorial representation created with a neural network. Both use a machine learning classifier to tune the importance of each dimension of the vectors for the target lexical-semantic relation. Experiments presented in the last section were carried out to show the effects of the different representations over the BLESS dataset. Here we report on a series of experiments that: (1) confirm the results on a novel dataset, the K&H dataset which was created from WordNet, a standard lexical resource; (2) show the capabilities of the systems to address the classification for very different types of relations, three paradigmatic relations and two syntagmatic relations; (3) show the dependence on the domain information.

6.3.1 Systems Comparison

While in the first experiments we showed the importance of the graph representation for GraCE and the importance of the vector representation for the machine learning system for WECE separately, here we compare all the systems in an in-domain setup using the K&H dataset. The results are presented in Table 6.8.

	BNC			Wikipedia		
	P	R	F	P	R	F
PairClass	76.9	4.6	8.7	77.0	11.7	20.4
BL	82.6	7.7	14.2	89.4	16.2	27.5
GraCE	90.7	43.8	59.0	94.0	75.5	83.7
$\text{DS}^{\text{C}}_{\text{ZHILA}}$	37,1	19,8	25,8	36,1	29,2	32,3
$\text{DS}^{\text{C}}_{\text{LEVY}}$	34,8	22,6	27,4	31,6	26,3	28,7
$\text{WECE}^{\text{BOW}}_{\text{OFFSET}}$	96.0	59.1	73.1	96.8	87.7	92.0
$\text{WECE}^{\text{BoW}}_{\text{CONCAT}}$	**97.4**	60.0	74.2	**97.6**	**89.3**	**93.2**
$\text{WECE}^{\text{DEP}}_{\text{OFFSET}}$	87.9	63.1	73.5	95.4	86.1	90.5
$\text{WECE}^{\text{DEP}}_{\text{CONCAT}}$	93.1	**64.7**	**76.4**	96.7	88.4	92.4
UL-RECALL		81.4			99.6	

Table 6.8: Aggregated results obtained for the in-domain setup with the K&H dataset.

The results confirm that using a larger corpus improves recall. All the methods created in line of the Traditional VSM, PairClass and BL achieved a significantly larger recall on Wikipedia than on BNC. However, the recall is very limited for both corpora. If in BNC only 19.4% of the K&H instances co-occur, and although in Wikipedia the number of co-occurrences raises to 30.7%, BL only correctly classifies 7.7 % of the pairs on BNC and 16.2 % on Wikipedia.

GraCE, DC^c and WECE systems have the real-upper limit in the number of pairs containing words that occur in corpus, shown in the last line of the Table 6.8.

GraCE, using the graph generalisation, overcomes the actual limitation of pattern-based systems: 40% and 78.7% of the K&H instances that never co-occur in BNC and in Wikipedia, respectively, are correctly classified by GraCE. This ability causes GraCE to improve the BL performance by 8.1 points in precision and 36.1 points in recall on BNC and 4.6 points in precision and 59.3 in recall on Wikipedia.

However, analysing the false negatives of the GraCE classifier, we observed that even relying on a graph-based corpus representation to extract the distributional information of a word pair, many errors are still caused by the sparsity of their vectorial representation. For the word pairs that do not co-occur in the same sentence, the GraCE vector representations of correctly-classified pairs have a median of eight non-zero features, indicating that the graph was beneficial for still providing evidence of a relationship; in contrast, incorrectly-classified pairs had a median of only three non-zero features, suggesting that even using all the dependency properties of words from the corpus for finding "bridging words", data sparsity is still the major origin of classification errors.

By combining all the distributional information into a denser vector, WECE systems are able to improve GraCE's results by an average of 2.9 points in precision and 17.9 points in recall.

The importance of the SVM classifier, for learning which features of the embeddings reflect the lexical relation, is once more highlighted by WECE results with an increase by 62 points in precision and 46 in recall over DS^c.

Between the WECE systems, the WECE$^{\text{dep}}_{\text{concat}}$ achieves the best results over BNC, while WECE$^{\text{bow}}_{\text{concat}}$ over Wikipedia. Over this dataset, information about the lexical similarity of pairwise words included in the WECE$_{\text{concat}}$

was meaningful for the classifier. However, over the BLESS, this information did not show to be useful, therefore further investigations are necessary to determine when this type of information is useful.

6.3.2 Results per Relation Type

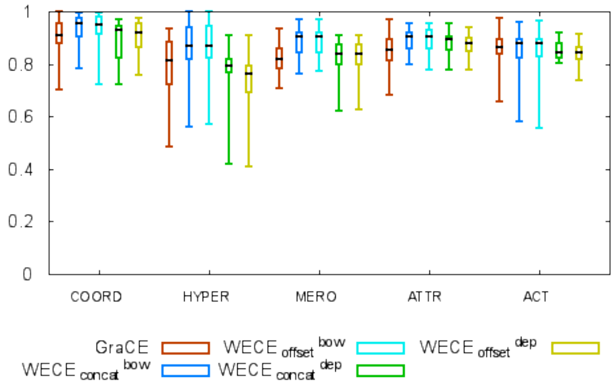


Figure 6.3: F1 distribution scores across domains obtained with BLESS dataset for each proposed system and relation type over BNC corpus.

In this work we are interested in creating a general approach for the classification of any lexical-semantic relation instance. In this experiment, we analyse the results obtained per relation in the previous experiments using BLESS dataset and K&H dataset.

Because BLESS contains 17 datasets, and we aim to show the differences in results across relations and their variation across domains, we show the box and whisker plot of the results obtained in the in-domain setup over the BNC Figure 6.3, and over the Wikipedia corpus in Figure 6.4.

The results confirm that the proposed systems achieve satisfactory results across all the relations, the median of the results being around 90

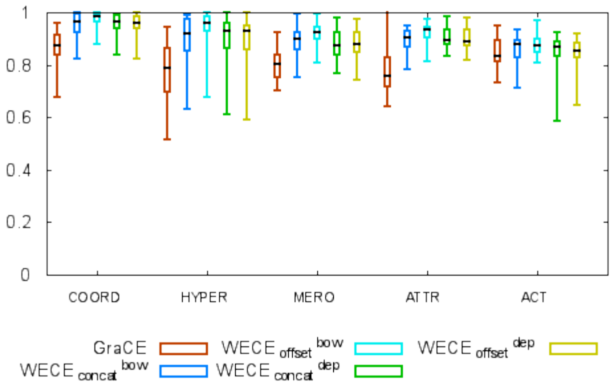


Figure 6.4: F1 distribution scores across domains obtained with BLESS dataset for each proposed system and relation type over Wikipedia corpus.

points in F1. The most accurate system is WECE^{bow}, which supports the assertion by [Levy and Goldberg, 2014a] that bag-of-word embeddings should offer superior performance over dependency-based embeddings on tasks involving semantic relations. The difference is more significant for BNC showing that this representations are more suitable for a smaller corpus. Carrying out an error analysis, the lowest results of the WECE systems are obtained in the domains with the fewest training instances, making apparent that word embedding systems are dependent on the number of training instances. For these domains, GraCE achieves better results.

Between all the results, the hypernymy relation achieves the results with the largest variation across domains , showing the complexity of the hypernymy relation.

The plots shown in Figure 6.3 and Figure 6.4 allow us to analyse also the *lexical memorisation* issue [Levy et al., 2015b]: when a supervised classifier is trained with various instances that contain the same word in a slot, any words occurring with that word will be labeled as an instance of the same relation r. We trained a multiclassifier over six classes of the

BLESS dataset and our classifier achieves good results in precision and recall over all the relations as one can observe from the results plots. Our classifier manages to associate the correct different relations for instances that share the same word on the same slot. For instance, the word *beet* occurs with: 11 attributes, 18 co-hyponyms, 14 actions, 5 hypernyms, 5 meronyms and 38 unrelated words being adjectives, nouns and verbs. The classifier is able to assign the correct label to each word pair but one attribute, one hypernym, two meronyms and one unrelated word that was classified as action.

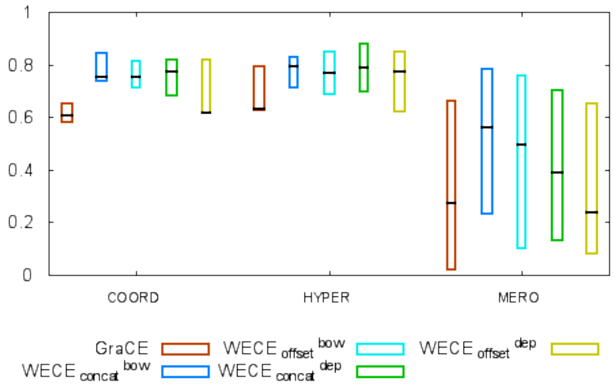


Figure 6.5: F1 distribution scores across domains obtained with K&H dataset for each proposed system and relation type over BNC corpus.

The results obtained with the K&H dataset are shown in Figure 6.5 for BNC and Figure 6.6 for Wikipedia. When we compare the two figures, we observe that using Wikipedia, the results improve considerably over all the relations. However, even if a corpus as large as Wikipedia is not available, using only the information over BNC, the WECE systems achieve satisfactory results, around 80 F1-points. Comparing the results across relations, we observe that the results for the relations of hypernymy and co-hyponymy are higher then the results for meronymy. On one hand,

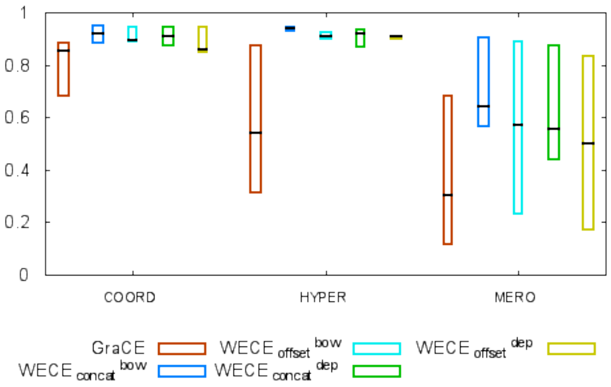


Figure 6.6: F1 distribution scores across domains obtained with K&H dataset for each proposed system and relation type over Wikipedia corpus.

one could think that a reason for this results could be the number of the available training instances. As shown in Table 4.2, the amount of training instances for meronyms acquired from WordNet is lower when compared with the instances provided for hypernymy and co-hyponymy. Indeed, the results for the relation of the meronymy over the *animals* and *plants* topical domain are lower, they containing fewer meronyms. However, in the *vehicles* topical domain, the K&N dataset provides only 189 hypernym pairs. Instead, the results of the systems for this domain and this relation are not affected by the number of training instances.

On the other hand, the amount of meronyms from K&H is larger than the meronyms provided by BLESS for each domain, which makes us think that a problem is also the polisemy of the words. For instance, WordNet provides as meronym for dog_n^1 the synset $flag_n^7$, which is not the most common sense for the word *flag*. Differently, because BLESS was manually developed, the meronyms provided by BLESS are words acquired from Wikipedia entries, providing words having their main sense the meronym sense.

From all the previous experiments, we conclude that GraCE achieves good results for BLESS datasets, over both corpora. However, this approach is considerable hindered by the words polisemy in the K&H dataset, where it needs an input corpus as Wikipedia to achieve satisfactory results. Instead, the reliability of the representations provided by the embeddings is shown, WECE system achieving the best results across each relation. Between the lexical relations tested over the BLESS dataset hypernymy achieves lower results, while in the K&H dataset the meronymy acquisition is considerably hindered by the words polysemy.

6.3.3 Domain-aware Training Instances

In all the previous experiments, the proposed systems were compared to test the importance of the novel representations that use a generalisation technique over the corpus information for creating representation word pairs that do not co-occur. However, the experiments were run only over one domain, therefore, the system takes advantage of the topical domain information. This topical domain information is observed in the patterns of context used, they keep specific information of the domain and benefit word pairs from that domain. For instance, the domain of vegetables, contains the patterns $x \xleftarrow{\text{obj\_inv}} eat_V \xrightarrow{\text{obj}} y$, which describes things that can be eaten. However, lexical-semantic relations are general relations that may be observed in all the topical domains. Therefore, one more experiment was carried out where the systems had to classify word pairs from a different domain than the domains of the training set. The objective is to assess the importance of the domain-aware training instances for the classification.

To test the importance of the domain-aware training instances we used the BLESS dataset, as it covers a larger number of domains, 17 when compared with the only 3 domains of the K&H dataset. The average results of the systems obtained for the *in-domain* setup across the BLESS dataset, presented in Section 6.1 and Section 6.2, are compared with the average

	BNC			Wikipedia		
	P	R	F	P	R	F
PairClass	78.9	43.2	55.8	81.2	40	53.6
BL	79.1	43.8	56.4	76.6	68.6	72.4
GraCE	71.7	40	51.4	77.8	38.8	51.8
$\text{DS}^{\text{C}}_{\text{Zhila}}$	47.6	55.3	51.2	50.1	54.3	52.1
$\text{DS}^{\text{C}}_{\text{Levy}}$	47.6	55.3	51.2	53.9	57.4	55.6
$\text{WECE}^{\text{BoW}}_{\text{offset}}$	68.0	**66.9**	67.5	71.3	69.7	70.5
$\text{WECE}^{\text{BoW}}_{\text{concat}}$	**83.8**	57.0	67.8	**84.6**	56.5	67.7
$\text{WECE}^{\text{Dep}}_{\text{offset}}$	68.7	62.3	65.4	74.6	**71.2**	**72.9**
$\text{WECE}^{\text{Dep}}_{\text{concat}}$	78.2	63.8	**70.3**	83.2	62.5	71.4

Table 6.9: Aggregated results obtained when the systems are tested out-of-domain with the BLESS dataset.

results obtained when the systems are trained *out-of-domain*. For the *out-of-domain* setup, one domain is left out from the training set and used for testing. The experiment was repeated for each domain and the average results are presented in Table 6.9.

When no examples from a domain are provided, a general significant decrease in performance is observed as presented in Figure 6.7 and Figure 6.8. As expected, GraCE is highly dependent of the lexical information showed by the 35.2 points decease in F1 over BNC and 31.6 over Wikipedia. However, even WECE systems are considerably hindered by this information: they decrease with 20.55 points F1 in average over BNC and 22.35 over Wikipedia.

The obtained results show that when the instances to be classified are less homogeneous, i.e. when the instances belong to different domains, none of the systems can achieve the level of performance reported for the in-domain setup. These were the expected results for the GraCE system due to the lexical features that it uses and which are domain dependent. However, the WECE systems are also affected by this lack of domain-aware training instances. In particular, $\text{WECE}_{\text{concat}}$ results decrease be-

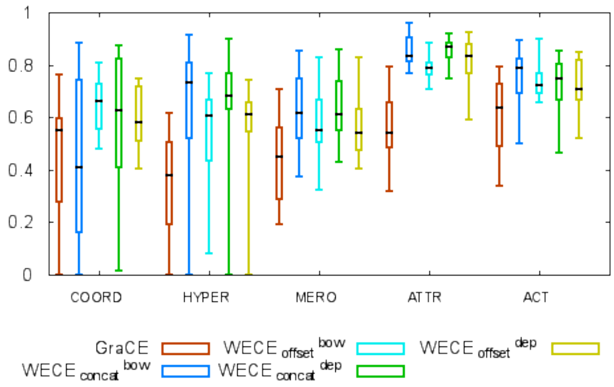


Figure 6.7: F1 distribution scores across domains for each proposed system and relation type obtained with the BLESS dataset in the out-of-domain setup over BNC corpus.

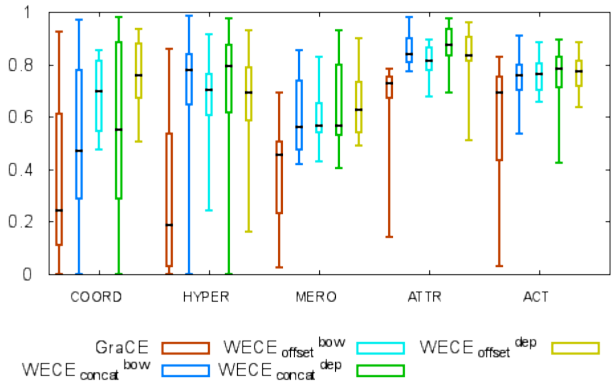


Figure 6.8: F1 distribution scores across domains for each proposed system and relation type obtained with the BLESS dataset in the out-of-domain setup over Wikipedia corpus.

cause similar embeddings are associated with similar words. When two words belong to two different topical domains, their embeddings are less similar and, therefore, the SVM system has difficulties to learn distinctive features for each lexical-semantic relation. WECE$_{\text{offset}}$, using the vector offset, eliminates the lexical information from the representation. As a consequence, this system achieves the best results to acquire information from different domains than the ones with which it was trained, especially for the acquisition of co-hyponyms. The analysis per relations shows that the most affected relation is the relation hypernymy and the less hindered relations are the syntagmatic relations when acquired with the WECE systems.

Chapter 7

CONCLUSIONS

To achieve human-like results many NLP systems must integrate human knowledge. This knowledge is encoded in lexical resources like taxonomies and ontologies, and consequently, there is a huge interest in the NLP domain to create lexical resources, for various languages and domains. The manual development of such resources involves a large amount of work, therefore, the automatisation of their acquisition becomes a real necessity.

This thesis addressed the automatic acquisition of lexical-semantic relations, starting from the latent-relational hypothesis: the contexts where two related words co-occur provide information about their semantic relation, as explained in Section 2.1. Systems that rely only in patterns of context are very precise, but they are upper-bound limited for recall: only semantically related word pairs that do co-occur in a particular input corpus can be acquired. But these are only a subset of all the semantically related word pairs, as our experiments with PaCE confirmed in Section 5.1. To be able to acquire more instances, the mainstream solution was provide more input corpus. However, large corpora are not always available for all the languages or all the domains. Therefore, the main goal of the research presented in this thesis was to search for word pair representations for the acquisition of lexical-semantic relation instances that were able to correctly represent even word pairs that do not co-occur.

Lexical-semantic relations are relations that emerge from the meaning of words, which is represented in terms of word distributional properties. To overcome the lack of information behind the recall problem of the pattern-based systems, we proposed novel representations that embed also information extracted from the distributional properties of words. Moreover, we focused on keeping similar precision scores as the previous systems that used only information about the co-occurrences of words. Our intuition was based on the fact that word pairs representation has to be created based on a generalisation of the corpus information, to provide also useful information about word pairs that do not co-occur. This generalization have to combine in an indistinctive form information about the contexts where the words co-occur with distributional properties of words, this way the two sources of information are capable to complete each other.

We implemented these intuitions in two systems GraCE and WECE. GraCE's novelty is the gathering of word distributional information from the proposed graph representation of input corpus created out of dependency relations. GraCE uses patterns of context as pattern-based system do, but it combines them with the information extracted from the graph, representing the distributional properties of words. The hypothesis behind GraCE was that there is a straight-forward correlation between word distributional properties, encoding contexts where each member of the word pair occurs, and the lexical-semantic relation holding between them.

The advantage of using a network of nodes connected by their dependency relations is that all the words are connected directly or indirectly by "bridging words". Therefore, words are not treated as independent features, like in the most common vectorial representations in the Traditional VSM and, the graph created with word distributional properties produces a generalisation over the information of actual co-occurrences in a corpus. As described in Section 6.1, the information extracted from the graph showed to be reliable and to keep the precision equals to the precision of other pattern-based systems as our PaCE, but considerably increasing the recall, because many word pairs that do not co-occur in corpus are correctly classified by GraCE. Thus, our hypothesis was validated.

As for WECE, it uses a novel word representation: word embeddings [Mikolov et al., 2013a]. They can be considered a more developed and more complicated representation of the intuition behind the network of words used in GraCE. Created by a neural network, word embeddings are a distributed vectorial representation in the sense that there is not a direct correspondence with distributional properties observed in a corpus. The information brought in the network by the co-occurrences of two words is spread through (almost) the entire network during the training time. The distributional information and the contexts where two words co-occur are encoded by the neural network in the vectorial word representations, which represent an (almost) perfect generalisation of the corpus information. This hypothesis is confirmed by the results shown in Section 6.2: WECE achieves considerably better results than GraCE, due to the density of the representation.

Although word embeddings have become popular only very recently (in fact at the final stage of this thesis) they have achieved impressive results in several tasks [Baroni et al., 2014]. The initial works that used word embeddings for the detection of the relations holding between words, in form of syntactic and semantic relations, have assumed that each relation is represented through a vector offset. Our assumption was that one lexical-semantic relation could not be properly represented by the same offset because these relations are more complex than the syntactic and semantic relations initially tested [Zhila et al., 2013, Levy and Goldberg, 2014b]. Moreover, the same initial approaches compared the vector offset of two word pairs to check if they have the same semantic relations. Therefore, the semantic similarity between two word pairs is based only on the similarity in the vectorial changes that have to be made to transform one member of the pair into the other, and ignoring any lexical similarity between word pair members.

Therefore, we tested two hypothesis using WECE systems. The first hypothesis was that lexical-semantic relations are not encoded in the offset but in the information that is encoded in each component of the embedding, and the linear combination of these components reflects the target relation. We used a machine learning system to discover a linear combination of the dimensions that reflect each target lexical-semantic relation. Indeed, the hypothesis was confirmed, as detailed in Section 6.2, WECE achieved considerable better results than DS, which uses only a simple vectorial similarity score between vector offsets. The second hypothesis was that in order to discover lexical-semantic relation instances, the representations should also include information about the lexical similarity between words. Word pair representations which use a vector offset to represent the target pair, miss the information about each member of the pair. Therefore, to test if the lexical similarity between pairwise words is important for the automatic acquisition of lexical semantic relation instances, we create WECE$_{\text{concat}}$, which represents a word pair by concatenating member's embeddings, thus, allowing the supervised system to learn features directly form the word embeddings. WECE$_{\text{concat}}$ and WECE$_{\text{offset}}$ achieved very similar results, showing that the lexical similarity between words is not very important for the acquisition of these relations. Slightly better results are achieved by WECE$_{\text{offset}}$ when the systems have to discover instances from different domains than the training used in the training step.

This thesis was focused on creating a general system for the automatic acquisition of different lexical-semantic relations, each of them with different distributional properties. In Section 6.3.2 we confirmed that similar results were achieved across all the addressed relations. While GraCE has a larger deviation in results and for some domains it achieves lower results, WECE achieves pretty good results across all the relations. The most complicated relation is the relation of hypernymy, for which all the systems achieved lower results. With these results we also showed that systems combining word embeddings and supervised systems do not suffer of "lexical memorization" as [Levy et al., 2015b] stated.

In addition to the empirical validation of our initial intuitions and hypothesis, which aimed to find the most suitable representation for the lexical-semantic relations holding between word pairs, our intuition was that the results achieved were influenced by the availability of the information about topical domain. Lexical-semantic relations are general relations, emerging from the meaning of words, which is related by the context where the words co-occur, context that may contain domain information. Indeed, the results from Section 6.3.3 showed that when domain information is not available an important decrease of the results is observed. Therefore, when a system is trained, it has to take into account domain information through the training word pairs. However, $\text{WECE}_{\text{offset}}$ systems are the less hindered systems in this setup.

7.1 Main contributions

Our research contributes in different ways to the state-of-the-art in the acquisition of lexical-semantic relation instances. In particular, our research focuses on the construction of new methods to improve the recall of the acquisition from a given corpus, and to avoid the necessity of traditional pattern-based systems to enlarge the input corpus for acquiring more information, as presented in the Section 5.2

The most meaningful contributions are outlined below:

- We showed that patterns of context information have to be combined with distributional information of words to improve the recall of pattern-based systems;
- We proposed two word pair representations which significantly improve state-of-the-art system recall by correctly classifying word pairs which do not co-occur in the same sentence;
- Although the representation of words as a graph is not new, we are the only ones, to our knowledge, representing the entire corpus in a

graph of words and making it the core of a general method for the acquisition of lexical-semantic relations; thus, we demonstrated that the information supplied, based on dependency relations, successfully allows the classification of semantically related words;

- We showed that Predictive VSM, i.e. word embeddings, are a more reliable representation for modelling the lexical-semantic relations of words; however, lexical relations are not encoded by the same offset vectors but, they are encoded by the information contained in the components of the vectors, and a supervised system has to be trained to weight the importance of each component;
- After the evidence that systems based only on patterns of context are very precise but they achieve a very low recall, we tried to improve the most general approaches that use dependency based lexicalized patterns; the mainstream approach uses only the shortest paths; we proposed a generalisation using PoS information and the usage of all the dependency patterns that do not co-occur a sufficient amount of times, but the results did not improve;
- We showed that systems achieve comparable results across five different relations, therefore each system created is a generic system for the detection of lexical-semantic relations;
- We tested each system in two setups: in-domain and out-of-domain, hence, showing the importance of the domain information for the acquisition of these types of relations;

All the systems used in this thesis, except `word2vec` system, were created by us during the thesis. We will provide to the community in a GitHub repository the following systems and resources:

- A system that takes as input a parsed text and provides a graph-based representation of the corpus;

- A system that acquires patterns of contexts from a given corpus and a dataset;
- The GraCE implementation;
- The PaCE implementation;
- All the WECE systems used in this thesis;
- Our implementation of PairClass;
- The graph-based representation of the BNC corpus and of the Wikipedia corpus;
- The vectorial representation of the BNC corpus and of the Wikipedia corpus as they were outputted by `word2vec`;
- The K&H dataset enhanced with co-hyponyms and meronyms.

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

- Necşulescu, Silvia; Mendes, Sara; Jurgens, David; Bel, Núria; Navigli, Roberto (2015). *"Reading Between the Lines: Overcoming Data Sparsity for Accurate Classification of Lexical Relationships"*. In Palmer, Martha (ed.) Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (SEM 2015). Denver, Colorado: Association for Computational Linguistics. p. 182–192. ISBN 978-1-941643-39-6
- Necşulescu, Silvia; Mendes, Sara; Bel, Núria (2014). *"Combining Dependency Information and Generalization in a Pattern-based Approach to the Classification of Lexical-Semantic Relation Instances"*. In Calzolari, Nicoletta (Conference Chair), Choukri, Khalid; Declerck, Thierry (et al.) (eds.) Proceedings of the Ninth International

Conference on Language Resources and Evaluation (LREC 2014): May 26-31, 2014 Reykjavik, Iceland. [s.l.]: ELRA. p. 4308-4315. ISBN 978-2-9517408-8-4

- Mendes, Sara; Necşulescu, Silvia; Bel, Núria (2012). *"Synonym extraction using a language graph model"*. In Barbu Mititelu, Verginica; Popescu, Octavian; Pekar, Viktor (ed.) Semantic Relations-II. Enhancing Resources and Applications (LREC 2012). Istanbul: 2012. p. 1-9

7.2 Future Directions

In the future, we plan to improve GraCE by analysing the importance of the edges for the acquisition process. For instance, calculating a statistical significance score, like log-likelihood ratio, and eliminating edges whose significance does not overcome a threshold. The intuition behind this process is that word pairs may be connected through paths containing dependency relations from different sentences, and some dependency relations are not reliable enough to enter in the distributional properties of a target word. At this moment we rely only on the SVM system to detect unreliable patterns. Moreover, we plan to apply a dimension reduction technique over the GraCE representations for finding similarities between patterns of contexts that are used as features.

The WECE system could be improved by providing the word embeddings to a Convolutional Neural-Network (CNN), which uses activation functions to detect meaningful features, instead of using an SVM system. CNN achieved very good results in image processing, and recently started to be used in NLP and achieved state-of-the-art results for many tasks.

This work may be also developed towards a complete system that is able to create a list of lexical-semantic relations from a given corpus. In the present work we detect the lexical-semantic relation holding between target word pairs. To create a complete system of acquisition of lexical-semantic word pairs, we want to develop also a technique which detects possible semantically related word pairs. This system could be created using the same graph-based representation of the corpus, and using the the Personalised Page Rank algorithm [Haveliwala, 2002], starting from a set of words from a target domain to detect words that are related with that domain, or starting from only one word to detect only words that enter in the semantic space of that word. Moreover, we could analyse the applicability of the GraCE and WECE systems to other relations than lexical-semantics. Although, our intuition is that GraCE is not very well fitted for this task, as it relies on the distributional properties of words, instead we expect a system based on word embeddings to achieve good results for the detection of almost any semantic relation, when enough training is available.

Bibliography

[bnc,]

[SAT,]

[Agichtein and Gravano, 2000] Agichtein, E. and Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

[Agirre et al., 2012] Agirre, A. G., Laparra, E., Rigau, G., and Donostia, B. C. (2012). Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *GWC 2012 6th International Global Wordnet Conference*, page 118.

[Almuhareb and Poesio, 2004] Almuhareb, A. and Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, volume 4, pages 158–165.

[Banerjee and Pedersen, 2002] Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer.

[Baroni et al., 2014] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52st Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 2014*, 1.

[Baroni and Lenci, 2010] Baroni, M. and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721.

[Baroni and Lenci, 2011] Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Baroni et al., 2010] Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

[Baroni and Zamparelli, 2010] Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1183–1193. ACL.

[Bel, 2010] Bel, N. (2010). Handling of missing values in lexical acquisition. In *In Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010).*

[Bengio et al., 2003a] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003a). A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

[Bengio et al., 2003b] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003b). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

[Bergsma et al., 2008] Bergsma, S., Lin, D., and Goebel, R. (2008). Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 59–68. ACL.

[Berland and Charniak, 1999] Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999)*, pages 57–64. ACL.

[Biemann, 2006] Biemann, C. (2006). Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, pages 73–80. ACL.

[Biemann and Riedl, 2013a] Biemann, C. and Riedl, M. (2013a). From global to local similarities: A graph-based contextualization method using distributional thesauri. *Graph-Based Methods for Natural Language Processing*.

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

[Boella and Di Caro, 2013] Boella, G. and Di Caro, L. (2013). Extracting definitions and hypernym relations relying on syntactic dependencies and support vector machines. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 2013)*, pages 532–537.

[Bollacker et al., 2008] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. ACM.

[Bollegala et al., 2009] Bollegala, D. T., Matsuo, Y., and Ishizuka, M. (2009). Measuring the similarity between implicit semantic relations from the web. In *Proceedings of the 18th international conference on World wide web*, pages 651–660. ACM.

[Budanitsky and Hirst, 2006] Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

[Bullinaria and Levy, 2007] Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

[Bunescu and Mooney, 2007] Bunescu, R. and Mooney, R. (2007). Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 2007)*, volume 45, page 576.

[Caraballo, 1999] Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999)*, pages 120–126. ACL.

[Cederberg and Widdows, 2003] Cederberg, S. and Widdows, D. (2003). Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the*

7th Conference on Natural Language Learning (CoNLL 2003), pages 111–118. ACL.

[Choi and Bae, 2004] Choi, K.-S. and Bae, H.-S. (2004). Procedures and problems in korean-chinese-japanese wordnet with shared semantic hierarchy. In *Global WordNet Conference*.

[Clark and Pulman, 2007] Clark, S. and Pulman, S. (2007). Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55.

[Clark and Weir, 2001] Clark, S. and Weir, D. (2001). Class-based probability estimation using a semantic hierarchy. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL-HLT 2001)*, pages 1–8. ACL.

[Clauset et al., 2008] Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.

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

[Cruse, 1986] Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press.

[Curran, 2003] Curran, J. R. (2003). *From Distributional to Semantic Similarity*. PhD thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

[Curran and Moens, 2002] Curran, J. R. and Moens, M. (2002). Improvements in Automatic Thesaurus Extraction. In *In Proceedings of the Workshop on Unsupervised Lexical Acquisition*, pages 59–66.

[Curran et al., 2007] Curran, J. R., Murphy, T., and Scholz, B. (2007). Minimising semantic drift with mutual exclusion bootstrapping. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, volume 6.

[Davidov and Rappoport, 2006] Davidov, D. and Rappoport, A. (2006). Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of COLING-ACL*, pages 297–304.

[De Marneffe et al., 2006] De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

[Dorow et al., 2004] Dorow, B., Widdows, D., Ling, K., Eckmann, J.-P., Sergi, D., and Moses, E. (2004). Using curvature and markov clustering in graphs for lexical acquisition and word sense discrimination. *arXiv preprint cond-mat/0403693*.

[Erk, 2007] Erk, K. (2007). A simple, similarity-based model for selectional preferences. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, 45(1):216.

[Erk and Padó, 2008] Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 897–906. ACL.

[Erk et al., 2010] Erk, K., Padó, S., and Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

[Evert, 2005] Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. PhD thesis, University of Stuttgart.

[Firth, 1957] Firth, J. (1957). A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32.

[Fountain and Lapata, 2012] Fountain, T. and Lapata, M. (2012). Taxonomy induction using hierarchical random graphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT 2012)*, pages 466–476. ACL.

[Fu et al., 2014] Fu, R., Guo, J., Qin, B., Che, W., Wang, H., and Liu, T. (2014). Learning semantic hierarchies via word embeddings. In *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL 2014)*.

[Gildea and Jurafsky, 2002] Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

[Girju et al., 2003] Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2003)*, pages 1–8. ACL.

[Girju et al., 2006] Girju, R., Badulescu, A., and Moldovan, D. I. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

[Golub and Kahan, 1965] Golub, G. and Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224.

[Gorman and Curran, 2006] Gorman, J. and Curran, J. R. (2006). Scaling distributional similarity to large corpora. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, pages 361–368. ACL.

[Grefenstette and Sadrzadeh, 2011] Grefenstette, E. and Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 1394–1404. ACL.

[Grefenstette, 1994] Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. The Springer International Series in Engineering and Computer Science. Springer US.

[Guevara, 2010] Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. ACL.

[Gunning et al., 2010] Gunning, D., Chaudhri, V. K., Clark, P. E., Barker, K., Chaw, S.-Y., Greaves, M., Grosof, B., Leung, A., McDonald, D. D., Mishra, S., et al. (2010). Project halo update - progress toward digital aristotle. *AI Magazine*, 31(3):33–58.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18.

[Harris, 1954] Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

[Haveliwala, 2002] Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM.

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling 1992)*, pages 539–545.

[Herdağdelen and Baroni, 2009] Herdağdelen, A. and Baroni, M. (2009). Bagpack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 33–40. ACL.

[Hindle, 1990] Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics (ACL 1990)*, pages 268–275. ACL.

[Hoffmann et al., 2011] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 541–550. ACL.

[Hovy et al., 2009] Hovy, E., Kozareva, Z., and Riloff, E. (2009). Toward completeness in concept extraction and classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 948–957.

[Huang et al., 2012] Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 873–882. ACL.

[Hughes and Ramage, 2007] Hughes, T. and Ramage, D. (2007). Lexical semantic relatedness with random graph walks. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 581–589.

[Igo and Riloff, 2009] Igo, S. P. and Riloff, E. (2009). Corpus-based semantic lexicon induction with web-based corroboration. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, pages 18–26. ACL.

[Ittoo and Bouma, 2010] Ittoo, A. and Bouma, G. (2010). On learning subtypes of the part-whole relation: do not mix your seeds. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1328–1336. ACL.

[Jurgens et al., 2012] Jurgens, D., Turney, P. D., Mohammad, S. M., and Holyoak, K. J. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364. ACL.

[Kilgarriff and Yallop, 2000] Kilgarriff, A. and Yallop, C. (2000). What's in a thesaurus? In *In Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC 2000)*.

[Kozareva and Hovy, 2010] Kozareva, Z. and Hovy, E. H. (2010). A semi-supervised method to learn and construct taxonomies using the web. In

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2010, pages 1110–1118.

[Kozareva et al., 2008] Kozareva, Z., Riloff, E., and Hovy, E. H. (2008). Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, volume 8, pages 1048–1056. Citeseer.

[Leacock and Chodorow, 1998] Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

[Lenat et al., 1985] Lenat, D. B., Prakash, M., and Shepherd, M. (1985). Cyc: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI magazine*, 6(4):65.

[Lenci et al., 2000] Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., et al. (2000). Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

[Lenci and Benotto, 2012] Lenci, A. and Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 75–79. ACL.

[Levy and Goldberg, 2014a] Levy, O. and Goldberg, Y. (2014a). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*. ACL.

[Levy and Goldberg, 2014b] Levy, O. and Goldberg, Y. (2014b). Linguistic regularities in sparse and explicit word representations. In *Proceed-*

ings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014). ACL.

[Levy and Goldberg, 2014c] Levy, O. and Goldberg, Y. (2014c). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems (NIPS 2014)*, pages 2177–2185.

[Levy et al., 2015a] Levy, O., Goldberg, Y., and Dagan, I. (2015a). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics (TACL 2015)*, 3:211–225.

[Levy et al., 2015b] Levy, O., Remus, S., Biemann, C., Dagan, I., and Ramat-Gan, I. (2015b). Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT 2015), Denver, CO*.

[Li and Abe, 1998] Li, H. and Abe, N. (1998). Generalizing case frames using a thesaurus and the mdl principle. *Computational linguistics*, 24(2):217–244.

[Lin, 1998a] Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING 1998)*, pages 768–774. ACL.

[Lin, 1998b] Lin, D. (1998b). An information theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann.

[Lin and Pantel, 2002] Lin, D. and Pantel, P. (2002). Concept discovery from text. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7. ACL.

[Ling et al., 2013] Ling, X., Clark, P., and Weld, D. S. (2013). Extracting meronyms for a biology knowledge base using distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 7–12. ACM.

[Liu and Singh, 2004] Liu, H. and Singh, P. (2004). Conceptnet – a practical common sense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

[Lund and Burgess, 1996] Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

[McCarthy and Carroll, 2003] McCarthy, D. and Carroll, J. (2003). Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

[Mihalcea and Tarau, 2004] Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*. ACL.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS 2013)*, pages 3111–3119.

[Mikolov et al., 2013c] Mikolov, T., Yih, W., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the*

Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2008), pages 746–751.

[Miller, 1995] Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

[Minkov and Cohen, 2008] Minkov, E. and Cohen, W. W. (2008). Learning graph walk based similarity measures for parsed text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 907–916. ACL.

[Minkov and Cohen, 2012] Minkov, E. and Cohen, W. W. (2012). Graph based similarity measures for synonym extraction from parsed text. In *Proceedings of the Workshop on Graph-based Methods for Natural Language Processing*, pages 20–24. ACL.

[Mintz et al., 2009] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009)*, pages 1003–1011. ACL.

[Mitchell and Lapata, 2008] Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of the 2008 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2008)*, pages 236–244.

[Mitchell and Lapata, 2010] Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.

[Nakov and Hearst, 2008] Nakov, P. and Hearst, M. A. (2008). Solving relational similarity problems using the web as a corpus. In *The 46th*

Annual Meeting of the Association for Computational Linguistics (ACL 2008), pages 452–460.

[Nastase and Szpakowicz, 2006] Nastase, V. and Szpakowicz, S. (2006). Matching syntactic-semantic graphs for semantic relation assignment. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 81–88. ACL.

[Navigli and Lapata, 2010] Navigli, R. and Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(4):678–692.

[Navigli and Ponzetto, 2010] Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 216–225. ACL.

[Navigli and Velardi, 2005] Navigli, R. and Velardi, P. (2005). Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(7):1075–1086.

[Navigli and Velardi, 2010] Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1318–1327, Uppsala, Sweden. ACL.

[Navigli et al., 2011] Navigli, R., Velardi, P., and Faralli, S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 1872–1877.

[Necsulescu et al., 2015] Necsulescu, Silvia Bel, N., Mendes, S., Jurgens, D., and Navigli, R. (2015). Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (*SEM 2015).*

[Ó Séaghdha and Copestake, 2009] Ó Séaghdha, D. and Copestake, A. (2009). Using lexical and relational similarity to classify semantic relations. In *Proceedings of EACL*, pages 621–629. ACL.

[Ó Séaghdha and Korhonen, 2012] Ó Séaghdha, D. O. and Korhonen, A. (2012). Modelling selectional preferences in a lexical hierarchy. 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*, pages 170–179. ACL.

[Padó and Lapata, 2007] Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

[Padó et al., 2007] Padó, S., Padó, U., and Erk, K. (2007). Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-CoNLL 2007)*, pages 400–409.

[Pantel and Lin, 2002] Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM.

[Pantel and Pennacchiotti, 2006] Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference* *on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 113–120. ACL.

[Pantel and Ravichandran, 2004] Pantel, P. and Ravichandran, D. (2004). Automatically labeling semantic classes. In *Proceedings of Human Language Technology/North American chapter of the Association for Computational Linguistics (HLT/NAACL 2004)*, volume 4, pages 321–328.

[Pantel et al., 2004] Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards terascale knowledge acquisition. In *Proceedings of the 20th international conference on Computational Linguistics*, page 771. ACL.

[Pease et al., 2002] Pease, A., Niles, I., and Li, J. (2002). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

[Phillips and Riloff, 2002] Phillips, W. and Riloff, E. (2002). Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 125–132. ACL.

[Platt, 1999] Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in kernel methods*, pages 185–208. MIT Press, Cambridge, MA, USA.

[Poesio et al., 2002] Poesio, M., Ishikawa, T., im Walde, S. S., and Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. In *In Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC 2002*, pages 1220–1224.

[Rapp, 2003] Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, pages 315–322.

[Resnik, 1996] Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1):127–159.

[Riloff and Jones, 1999] Riloff, E. and Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI 1999)*, pages 474–479.

[Riloff and Shepherd, 1997] Riloff, E. and Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*.

[Ritter et al., 2009] Ritter, A., Soderland, S., and Etzioni, O. (2009). What is this, anyway: Automatic hypernym discovery. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*.

[Roark and Charniak, 1998] Roark, B. and Charniak, E. (1998). Noun-phrase co-occurrence statistics for semiautomatic semantic lexicon construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL - COLING 1998)*, pages 1110–1116. ACL.

[Rooth et al., 1999] Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1999). Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 104–111. ACL.

[Rothenhäusler and Schütze, 2009] Rothenhäusler, K. and Schütze, H. (2009). Unsupervised classification with dependency based word spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 17–24. ACL.

[Sahlgren, 2006] Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University.

[Salton et al., 1997] Salton, G., Singhal, A., Mitra, M., and Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207.

[Schütze and Walsh, 2008] Schütze, H. and Walsh, M. (2008). A graph-theoretic model of lexical syntactic acquisition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 917–926. ACL.

[Schwartz et al., 2014] Schwartz, R., Reichart, R., and Rappoport, A. (2014). Minimally supervised classification to semantic categories using automatically acquired symmetric patterns. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1612–1623.

[Scott and Matwin, 1998] Scott, S. and Matwin, S. (1998). Text classification using wordnet hypernyms. In *Use of WordNet in natural language processing systems: Proceedings of the conference*, pages 38–44.

[Snow et al., 2004] Snow, R., Jurafsky, D., and Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of Neural Information Processing Systems (NIPS 2004)*.

[Stamou et al., 2002] Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2002). Balkanet: A multilingual semantic network for the balkan languages. In *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25.

[Steyvers and Tenenbaum, 2005] Steyvers, M. and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78.

[Suchanek et al., 2008] Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

[Surdeanu et al., 2012] Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 455–465. ACL.

[Szpektor et al., 2004] Szpektor, I., Tanev, H., Dagan, D., Coppola, B., et al. (2004). Scaling web-based acquisition of entailment relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

[Tanev and Magnini, 2008] Tanev, H. and Magnini, B. (2008). Weakly supervised approaches for ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 129–143.

[Thater et al., 2010] Thater, S., Fürstenau, H., and Pinkal, M. (2010). Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 948–957. ACL.

[Thelen and Riloff, 2002] Thelen, M. and Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 214–221. ACL.

[Toutanova et al., 2004] Toutanova, K., Manning, C. D., and Ng, A. Y. (2004). Learning random walk models for inducing word dependency distributions. In *Proceedings of the twenty-first international conference on Machine learning*, page 103. ACM.

[Tsang and Stevenson, 2010] Tsang, V. and Stevenson, S. (2010). A graph-theoretic framework for semantic distance. *Computational Linguistics*, 36(1):31–69.

[Turney et al., 2005] Turney, P. et al. (2005). Measuring semantic similarity by latent relational analysis. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005)*.

[Turney, 2006a] Turney, P. D. (2006a). Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling/ACL 2006)*, pages 313–320. ACL.

[Turney, 2006b] Turney, P. D. (2006b). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

[Turney, 2008a] Turney, P. D. (2008a). The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research (JAIR)*, 33:615–655.

[Turney, 2008b] Turney, P. D. (2008b). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912.

[Turney, 2012] Turney, P. D. (2012). Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research (JAIR)*, pages 533–585.

[Turney, 2013] Turney, P. D. (2013). Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *Transactions of the Association for Computational Linguistics (TACL)*, pages 353–366.

[Turney et al., 2010] Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37(1):141–188.

[Van de Cruys, 2010] Van de Cruys, T. (2010). A non-negative tensor factorization model for selectional preference induction. *Natural Language Engineering*, 16(04):417–437.

[Van de Cruys, 2014] Van de Cruys, T. (2014). A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 26–35.

[Van Hage et al., 2006] Van Hage, W. R., Kolb, H., and Schreiber, G. (2006). A method for learning part-whole relations. In *The Semantic Web-ISWC 2006*, pages 723–735. Springer.

[Voorhees, 1998] Voorhees, E. M. (1998). Using wordnet for text retrieval. *WordNet: An Electronic Lexical Database*, pages 285–303.

[Vossen et al., 1997] Vossen, P. et al. (1997). Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.

[Weeds, 2003] Weeds, J. (2003). *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, University of Sussex.

[Weeds et al., 2014] Weeds, J., Clarke, D., Reffin, J., Weir, D., and Keller, B. (2014). Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 2249–2259.

[Weeds and Weir, 2003] Weeds, J. and Weir, D. (2003). A general framework for distributional similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 81–88.

[Weeds et al., 2004] Weeds, J., Weir, D., and McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*. ACL.

[Widdows, 2003] Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 197–204. ACL.

[Widdows, 2008] Widdows, D. (2008). Semantic vector products: Some initial investigations. In *Second AAAI Symposium on Quantum Interaction*, volume 26, page 28th.

[Widdows and Dorow, 2002] Widdows, D. and Dorow, B. (2002). A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics (COLING 2002).*

[Winston et al., 1987] Winston, M. E., Chaffin, R., and Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive science*, 11(4):417–444.

[Wu and Weld, 2010] Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 118–127.

[Yang and Callan, 2009] Yang, H. and Callan, J. (2009). A metric-based framework for automatic taxonomy induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-COLING 2009)*, pages 271–279. ACL.

[Yokoi, 1995] Yokoi, T. (1995). The edr electronic dictionary. *Communications of the ACM*, 38(11):42–44.

[Zhila et al., 2013] Zhila, A., Yih, W.-t., Meek, C., Zweig, G., and Mikolov, T. (2013). Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2013)*, pages 1000–1009.

[Zhitomirsky-Geffet and Dagan, 2005] Zhitomirsky-Geffet, M. and Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of Association for Computational Linguistics.*

[Zhitomirsky-Geffet and Dagan, 2009] Zhitomirsky-Geffet, M. and Dagan, I. (2009). Bootstrapping distributional feature vector quality. *Computational linguistics*, 35(3):435–461.