Automatic induction of verb classes using clustering

Lin Sun

University of Cambridge
Computer Laboratory
Girton College

December 2012

This dissertation is submitted for
the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the regulation length of 60 000 words, including tables and footnotes.

Automatic induction of verb classes using clustering

Lin Sun

Summary

Verb classifications have attracted a great deal of interest in both linguistics and natural language processing (NLP). They have proved useful for important tasks and applications, including e.g. computational lexicography, parsing, word sense disambiguation, semantic role labelling, information extraction, question-answering, and machine translation (Swier and Stevenson, 2004; Dang, 2004; Shi and Mihalcea, 2005; Kipper *et al.*, 2008; Zapirain *et al.*, 2008; Rios *et al.*, 2011). Particularly useful are classes which capture generalizations about a range of linguistic properties (e.g. lexical, (morpho-)syntactic, semantic), such as those proposed by Beth Levin (1993). However, full exploitation of such classes in real-world tasks has been limited because no comprehensive or domain-specific lexical classification is available.

This thesis investigates how Levin-style lexical semantic classes could be learned automatically from corpus data. Automatic acquisition is cost-effective when it involves either no or minimal supervision and it can be applied to any domain of interest where adequate corpus data is available. We improve on earlier work on automatic verb clustering. We introduce new features and new clustering methods to improve the accuracy and coverage. We evaluate our methods and features on well-established cross-domain datasets in English, on a specific domain of English (the biomedical) and on another language (French), reporting promising results. Finally, our task-based evaluation demonstrates that the automatically acquired lexical classes enable new approaches to some NLP tasks (e.g. metaphor identification) and help to improve the accuracy of existing ones (e.g. argumentative zoning).

Acknowledgements

I wish to thank my supervisor, Anna Korhonen. She patiently provided vision, encouragement and advice necessary for me to proceed with my doctoral program and complete my dissertation. I want to thank Anna for her unflagging encouragement and serving as a role model to me as a member of academia. She has always given me a great freedom to pursue independent work.

I am grateful to Yuval Krymolowski and Thierry Poibeau for their practical support and advice. Yuval has made various datasets available and answered my endless queries when I started my research. Thierry Poibeau provided valuable advice on French verb clustering and constructed the gold-standard for me to start experimenting. In addition, I would like to thank Cédric Messiant for making the French subcategorization lexicon available.

I would like to thank Ekaterina Shutova and Yufan Guo for collaboration on task-based evaluations. Special thanks to Ilona Silins, Ulla Stenius, Johan Hogberg, Diarmuid Ó Séaghdha and Ian Lewin for a great collaboration on cancer risk assessment.

I would also like to thank my examiners, Simone Teufel and Diana McCarthy who provided constructive feedbacks and detailed comments for this thesis.

I am grateful to Girton College and my college supervisor Frances Gandy. The financial support from the British Telecom and EPSRC are gratefully acknowledged.

I wish to thank my wife Yu Zhang and my parents for their continuous encouragement. I would like to thank my friends Yichi Zhang, Shuguang Li, Yixi Sui, Likan Fu and Hui Li for their support, understanding and the extra burden they have to shoulder.

Contents

List of Abbreviations

ACC	Accuracy
AGG	Agglomerative Clustering
ANLT	Alvey Natural Language Tools
ASSCI	A Subcategorization Frame Acquisition System for French Verbs
AZ	Argumentative Zoning
BL	Baseline
BNC	The British National Corpus
CO	Co-occurrence
COMLEX	COMLEX Syntax Dictionary
DA	Diathesis Alternation
EM	Expectation-Maximisation Algorithm
F	F-Measure
GENIA	The GENIA project
GR	Grammatical Relation
HGFC	Hierarchical Graph Factorization Clustering
IGNG	Incremental Growing Neural Gas
JSD	Jensen–Shannon divergence
K-Means	K-Means Clustering Algorithm
LexSchem	LexSchem Subcategorization Lexicon for French Verbs
LP	Lexical Preference

MDL	Minimum Description Length
MNCut	Multiway Normalizaed Cut Spectral Clustering
mPUR	Modified Purity
MRD	Machine Readable Dictionary
NLP	Natural Language Processing
NMI	Normalized Mutual Information
NP	Noun Phrase
OBJ	Object
P	Precision
PC	Pairwise Clustering
POS	Part-of-Speech
PP	Prepositional Phrase
R	Recall
$\mathrm{R}_{\mathrm{adj}}$	Adjusted Rand Index
RASP	Robust Accurate Statistical Parsing System
SCF	Subcategorization Frame
SP	Selectional Preference
SPEC	Spectral Clustering
SUBJ	Subject
VALEX	VALEX Subcategorization Lexicon for English Verbs
VN	VerbNet
VP	Verb Phrase

Chapter 1

Introduction

Verb classifications have attracted a great deal of interest in both linguistics and natural language processing (NLP). Verb classes which capture generalizations about a wide range of linguistic properties (e.g. lexical, (morpho-)syntactic, semantic), such as those proposed by Beth Levin (1993), have been of particular interest for NLP. They have proved useful for various important tasks and applications, including e.g. computational lexicography, parsing, word sense disambiguation, semantic role labelling, information extraction, question-answering, and machine translation (Swier and Stevenson, 2004; Dang, 2004; Shi and Mihalcea, 2005; Kipper *et al.*, 2008; Zapirain *et al.*, 2008; Rios *et al.*, 2011). However, full exploitation of such classes in real-world tasks has been limited because no comprehensive or domain-specific lexical classification is available.

This thesis investigates how Levin-style classes could be learned automatically from corpus data. Unsupervised or minimally supervised acquisition is cost-effective and can be applied to any domain of interest provided adequate corpus data is available. We improve on earlier work on automatic verb clustering. We introduce new features and new clustering methods to improve the accuracy and coverage. We evaluate our methods and features on general English, a specific domain (the biomedical) and another language (French), as well as in the context of NLP tasks, reporting promising results. Our task-based evaluation demonstrates that the automatically acquired lexical classes enable new approaches to some NLP tasks (e.g. metaphor identification) and help to improve the accuracy of existing ones (e.g. argumentative zoning).

This introductory chapter discusses the need for automatic lexical acquisition and the approaches proposed so far (section 1.1). It then discusses the central role of a verb in the syntactic structure and the semantics of a sentence (section 1.2). Section 1.3 gives an overview of the linguistic theory behind Levin-style classification, describes its importance for NLP, and discusses the automatic acquisition of verb classifications from corpus data. Section 1.4 summarizes the contributions of our work and section 1.5 provides an overview of the organization of this thesis.

1.1 Automatic lexical acquisition

The lexicon plays a central role in NLP. Rich lexical knowledge-bases are an important component of NLP applications (e.g. Word-Sense Disambiguation and Information Extraction).

Building large, explicit enough lexicons for NLP use has proved difficult. Manual construction of a large-scale lexicon involves many years of lexicographic work. The use of computers has enabled automating the task. Much of the early work on computational lexicography exploited information in machine readable dictionaries (MRD). Since MRDs were constructed primarily for human use, the conversion to a satisfactory computational lexicon proved difficult (Tuells, 1997). Also, it turned out that manually built lexicons have gaps and inconsistencies which are difficult to detect automatically (e.g. Boguraev and Briscoe (1987)). It is also costly to extend such resources to cover new information (e.g. statistical information).

Researchers therefore started to acquire lexical information automatically from corpus data. This automatic approach can solve the above mentioned problems: it can obtain a good coverage, and it can gather statistical information and can be easily applied to sub-languages and domains. Automatic methods have been developed for many areas of the lexicon: e.g. subcategorization frames (SCF) (Briscoe and Carroll, 1997; Korhonen, 2002; Messiant *et al.*, 2008; Lippincott *et al.*, 2012), selectional preferences (SPs) (Brockmann and Lapata, 2003; Erk, 2007; Bergsma *et al.*, 2008; Ó Séaghdha, 2010; Ó Séaghdha and Korhonen, 2012), diathesis alternations (Lapata, 1999; McCarthy and Korhonen, 1998) and word sense induction (Pantel and Lin, 2002; Navigli and Crisafulli, 2010; Ponzetto and Navigli, 2010; Lau *et al.*, 2012). Many approaches have aimed to minimise data annotation and employed semi-supervised or unsupervised methods. Many current methods need further refinement, but resources produced by some methods have already proved useful for NLP tasks and applications e.g., question answering (Lin and Pantel, 2001), unsupervised POS tagging (Clark and Tim, 2003) and text entailment (Zhang *et al.*, 2010).

1.2 The role of a verb in a sentence

The verb is central to the syntax and semantic of a sentence. The main verb of a sentence determines the number and the role of participants in the event the sentence is describing. For example, the verb *give* may select as its subject an entity that is able to give (AGENT), as its object an entity which is being given (THEME), and as another object an entity which is being given to (RECIPIENT). From the syntactic point of view, these requirements can be realized as a dative construction, e.g. *John gives an apple to Lucy* or a double object construction, e.g. *John gives Lucy an apple.*

When constructing a sentence, the speaker will find the participants that are compatible with the selectional restrictions or preferences of the verb in question, and will compose the sentence according to the syntactic mapping rules. Therefore, verbs play a key role in the meaning and the structure of sentences.

1.3 Verb classification

Verbs can be classified according to their syntax (Boguraev *et al.*, 1987; Grishman *et al.*, 1994), semantics (Fellbaum, 1998) or other properties. A number of linguists (Pinker, 1989; Jackendoff, 1990; Levin, 1993) have shown that verbs which share commonality in meaning often share also commonality in their (morpho-)syntactic behaviour and can be grouped into lexical classes according to a wider range of linguistic properties, e.g. Pinker (1989), Jackendoff (1990) and Levin (1993). Such classes (often called lexical-semantic classes) generalize over a range of linguistic properties of verbs without defining the idiosyncratic details for each verb. For example, the Levin class of COOK verbs includes the verbs *cook, bake, boil, roast and heat* which share similar meaning and the following syntactic frames and alternations between frames (Levin, 1993):

- **Causative/Inchoative Alternation**
  1. Jennifer baked the potatoes.
  2. The potatoes baked.
- **Middle Alternation**
  1. Jennifer baked Idaho potatoes.
  2. Idaho potatoes bake beautifully.
- **Instrument subject Alternation**
  1. Jennifer baked the potatoes in the oven.
  2. This oven bakes potatoes well.

The largest and the most widely deployed English verb classification in NLP is the classification of Levin (1993). This classification provides a summary of the variety of theoretical research done on lexical-semantic verb classification over the past decades. Verbs which display the same or a similar set of diathesis alternations are assumed to share certain meaning components and are organized into a semantically coherent class. Although alternations are chosen as the primary means for identifying verb classes, additional properties related to subcategorization, morphology and extended meanings of verbs are taken into account as well. VerbNet (Kipper-Schuler, 2005) [1] – an extensive on-line lexicon for English verbs – provides detailed syntactic-semantic descriptions of Levin's classes as well as additional classes organized into a refined taxonomy. The resulting taxonomy classifies over 6272 verbs into 270 first level classes.

Levin-style verb classes are interesting for NLP because they can help to reduce the redundancy in the lexicon (since verbs in a class share similar properties). They can also alleviate the problem of data sparseness which affects many NLP tasks by predicting the properties of member verbs, when not enough empirical evidence is available. VerbNet classes have been used to help tasks such as parsing, word sense disambiguation, semantic role labelling, information extraction, question-answering, and machine translation (Swier and Stevenson, 2004; Dang, 2004; Shi and Mihalcea, 2005; Zapirain *et al.*, 2008; Rios *et al.*, 2011). Such verb classifications could be particularly useful for domains. One such domain is the biomedical domain. According to Ananiadou and McNaught (2005), current manually built lexical resources (e.g. the UMLS specialist lexicon (Browne *et al.*, 2003)) do not provide enough coverage for domain specific language properties, e.g. word usage and word relation.

Because verb classes are sensitive to meaning variations between different text types and domains, and manual classification of large numbers of verbs is not practical, automatic verb classification has received a considerable amount of interest (Schulte im Walde, 2006; Joanis *et al.*, 2007; Sun *et al.*, 2008b; Li and Brew, 2008; Korhonen *et al.*, 2008; Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009b). Automatic classification is not only cost-effective but it also gathers the important statistical information from data and can easily be applied to new domains and usage patterns. Like manual classification, automatic classification is based on the assumption that the shared semantic behaviour of verbs can be largely inferred by their shared (morpho-)syntactic behaviour.

The methods proposed so far use machine learning techniques to classify syntactic and semantic features extracted from corpus data using part-of-speech (POS) tagging or statistical parsing techniques. Both supervised and unsupervised approaches have been proposed. Primarily unsupervised approaches are particularly interesting because they are easy to port between different tasks and domains. However, the accuracy of such approaches shows room for improvement. Also, because most approaches have been evaluated on a general language, only on one language, and against pre-determined gold standards, we do not know how useful they are to practical tasks.

1.4 Our contributions

The aim of this thesis is to improve the accuracy of automatic, primarily unsupervised verb classification and to evaluate the approach not only using well-established gold

[1] See http://verbs.colorado.edu/verb-index/index.php for details.

standard for English, but also on a different domain and a language, and in the context of NLP tasks.

Our contributions include:

- Introducing two clustering methods (MNCut spectral clustering (SPEC) and hierarchical graph factorization (HGFC) to the field of NLP. The methods outperform the best performing methods in previous works.
- Exploring a wide range of features including the selectional preference (SP) feature. The verb clustering performance is improved by using the SP features.
- Investigating whether the method and features developed for general English can be applied to a new language (French) and new domain (biomedicine). The results show that both the method and the features transfer well to French and biomedical domain.
- Performing two task based-evaluations of the verb clusters. To the author's knowledge, this is the first task-based evaluations on automatically acquired verb clusters. The results demonstrate that the automatically acquired verb clusters can be very useful for NLP applications.

The following subsections describe these contributions in detail.

1.4.1 Clustering methods

We introduce two novel clustering methods – a new variation of spectral clustering and HGFC which are new not only for verb classification but also for the field of NLP. SPEC has previously been used for a verb classification task (Brew and Schulte im Walde, 2002), but we use a new version (Maila and Shi, 2001) which consistently outperforms other previous methods in our studies. We also introduce a method to automatically detect the number of clusters for SPEC (Zelnik-Manor and Perona, 2004). In addition, we introduce HGFC (Yu *et al.*, 2006) as a hierarchical verb clustering method. Hierarchical verb clustering is important as all the existing manually built classifications are hierarchical in nature (e.g. VerbNet). All previous hierarchical verb clustering experiments (Schulte im Walde and Brew, 2001; Stevenson and Joanis, 2003; Ferrer, 2004) have used linkage based clustering methods. We address two problems with the linkage method: 1) error propagation (i.e. when a verb is misclassified at a level, the error propagates to all the upper/lower levels) and 2) local pairwise merging (only two clusters can be combined at any level). HGFC solves these two problems and shows improved performance on small and large-scale verb classification tasks. The method can also be modified to integrate prior knowledge about the task in the form of soft constraints. In other words, the HGFC can be adjusted to integrate semi-supervision in situations where some gold standard data is available.

1.4.2 Features

We explore a wide range of features in our studies, focussing in particular on SP features. Previous works have used SPs acquired from WordNet or GermaNet and have reported that the feature offers no significant improvement over syntactic features (Schulte im Walde, 2006; Joanis, 2002). This is contradictory to manual verb classification where selectional preferences are often discussed. Using a new clustering and SP acquisition methods, we demonstrate that such semantic information can be very useful. The performance is greatly improved over the use of purely syntactic features.

1.4.3 Language and domain transferability

We investigate whether a verb clustering approach initially developed using general English can be transferred to a new language and to a new domain. For many languages there is no Levin-style verb classification. We conduct experiments on French using SPEC and features similar to our work in English. The result confirms that the clustering techniques and features can be applied to a new language without substantial changes. However, to obtain better results, a large corpus and improved language-specific NLP tools (e.g. tagger, parser) would be ideal. We also apply our methods (SPEC and HGFC) and features to the biomedical domain. We perform experiments on Korhonen *et al.* (2008)'s dataset. The results show that our clustering outperforms previous results on the same data using similar features. The SP feature is the best feature in all the experiments, demonstrating that our methods can achieve good performance in a new domain. Domain specific features could be tried to further improve performance.

1.4.4 Task-based evaluation

VerbNet has proved useful for many practical NLP tasks, e.g. (Swier and Stevenson, 2004; Dang, 2004; Shi and Mihalcea, 2005; Zapirain *et al.*, 2008). Automatically acquired classification has not been applied to any NLP task. In collaboration with other researchers, we apply our clustering results to two tasks: metaphor identification and argumentative zoning. In the metaphor identification task, automatically obtained noun and verb classes are used to detect the source and target concept domains. The system is the first of its kind and it is capable of identifying metaphorical expressions with a high precision (0.79). The coverage is also better compared against the WordNet baseline. In the argumentative zoning task, verb classes are used as an additional feature for detecting the categories based on scientific discourse (e.g. Introduction, Background and Conclusion) in biomedical abstracts. The results show that this feature improves performance over the raw verb feature. In fact, it makes the raw verb feature redundant: adding the raw verb feature over the verb class feature actually decreases the performance.

1.5 Overview of subsequent chapters

The chapters of this thesis are organized as follows:

Chapter 2 *(Background to verb classification)* introduces the background and motivates our work. We discuss current verb resources and the benefits of Levin-style verb classification, including its use in NLP applications. We survey approaches to automatic verb classification, and the data, features and methods used in each approach are discussed. We finally identify the problems which need to be addressed in current approaches and define the scope of our work.

Chapter 3 *(Verb clustering using selectional preferences)* examines the role of semantic features (SPs) for verb clustering. We use a clustering method – MNCUT SPEC – which is new to NLP and particularly useful for handling high dimensional features. We also introduce a method for SP acquisition which is based on the same clustering algorithm. Using this approach, we show on two well-established test sets that automatically acquired SPs can be highly useful for verb clustering. This result contrasts with most previous works but is in line with theoretical work on verb classification which relies not only on syntactic but also on semantic features (Levin, 1993).

Chapter 4 *(Hierarchical verb clustering using graph factorization)* investigates hierarchical clustering of verbs. Most previous research on verb clustering has focused on acquiring flat classifications, although many manually built classifications are taxonomic in nature. We introduce a new graph-based method – HGFC – to hierarchical verb clustering which avoids some of the problems reported with the frequently used agglomerative method (AGG)[2]. We modify HGFC so that it can automatically determine the tree structure for clustering, and propose two extensions to it. The first involves automatically determining the number of clusters to be produced. The second involves adding soft constraints to guide the clustering performance, useful for domains where some prior classification is available. The results are promising. On a flat test set, the unconstrained version of HGFC outperforms AGG and performs similarly with the best current flat clustering method (SPEC). On the hierarchical test sets, the unconstrained and constrained versions of HGFC outperform AGG clearly at all levels. The constrained version of HGFC detects the missing hierarchy from the existing gold standards with high accuracy. When the number of clusters and levels is learned automatically, the unconstrained method produces a fairly accurate multi-level hierarchy. Finally, our qualitative evaluation shows that both constrained and unconstrained versions of HGFC are capable of learning valuable novel information not included in gold standards.

[2] We also use the name *linkage method* in this thesis.

Chapter 5 *(Verb classification in the biomedical domain)* explores the domain-specific application of verb clustering methods developed using general, cross-domain data. We apply our two clustering methods (SPEC and HGFC) to a 3 level hierarchical gold standard which consists of verbs in biomedical texts. We use the same features as in Korhonen *et al.* (2008), and extract the features from biomedical journals. We demonstrate that SPEC outperforms the previous best method (PC) on 3 levels. We then use HGFC to produce a 3 level hierarchy. We show that the resulting hierarchy is more accurate than the 3 level flat clustering produced by PC. We conclude that both clustering methods that are developed for the general domain can be applied to the biomedical domain without significant change.

Chapter 6 *(Cross-linguistic potential of verb classification)* investigates the cross-linguistic potential of Levin style verb clustering. We apply the SPEC clustering method developed for general English to French. We develop our initial gold standard based on the translation of a widely used English gold standard (Sun *et al.*, 2008b). The SCF features are acquired from a French subcategorization lexicon - LexSchem (Messiant *et al.*, 2008). The results show that not only the general methodology but also the best performing features are transferable between the languages, making it possible to learn useful VerbNet style classes for French automatically without language specific tuning.

Chapter 7 *(Task-based evaluation of verb classification)* evaluates the use of verb clustering in two NLP tasks - metaphor identification and argumentative zoning. [3] For metaphor identification, the verb and noun clusters enable a novel approach for unrestricted text. Starting from a limited set of metaphorical seeds, we use noun and verb clustering to harvest the target concepts (noun clusters) associated with the same source domain (verbs clusters). The selectional strength filter is used to filter out the verbs that are not prone to metaphoricity. Finally, we show that the system is capable of capturing the regularities behind metaphor production and annotating a wider range of previously unseen metaphors. In the argumentative zoning task, verb classes are used as additional feature to identify the argumentative discourse categories in scientific abstracts. The verb class feature greatly improves the classification accuracy.

Chapter 8 *(Conclusion)* summarises the contributions of our work and proposes directions for future research.

[3] These are the collaboration works with Ekaterina Shutova and Yufan Guo. Our contribution is to provide the verb and noun clusters.

Chapter 2

Background to verb classification

2.1 Verb resources in NLP

Verbs are central to the syntactic structure and the meaning of sentences. Many computational resources and classifications have been developed for verbs. They can be classified into three types:

Syntactic resources : Comlex (Grishman *et al.*, 1994) and ANLT (Boguraev *et al.*, 1987) dictionaries are examples of syntactic verb resources. These dictionaries are mainly manually developed. A verbal entry comprises verb forms and subcategorization information.

Semantic resources : FrameNet (Baker *et al.*, 1998), PropBank (Kingsbury and Palmer, 2002) and WordNet (Miller, 1995) are mainly semantic resources of English verbs. FrameNet groups verbs according to the conceptual structures (frames) and their combinatorial patterns. PropBank provides information about predicate-argument structures of verbal predicates. The core arguments of each verb are numbered. For a given verb, the argument with the same number always has the same semantic meaning across different syntactic frames. WordNet groups words into synsets (synonym sets), and records the semantic relation between synsets. These resources contain little syntactic information, or the syntactic information is bound to the semantics of individual verbs (e.g. PropBank and WordNet).

Syntactic-semantic resources : In Levin (1993)'s verb classification, verbs are grouped in terms of shared meaning components, similar (morpho-)syntactic behavior of words and a number of other properties. VerbNet (Kipper-Schuler, 2005) extends this classification with a large number of additional verbs and classes.

For many NLP applications, a Levin style verb classification is useful for its ability to capture generalizations about a range of linguistic properties. For example, in the semantic role labelling task, Zapirain *et al.* (2008) discovered that a system trained on

PropBank was too focussed on verb specific knowledge and VerbNet classes provided a better starting point. Also WordNet has been criticized for lack of generalization, for example: Palmer (2000) argues that WordNet's level of sense distinction is too fine-grained for a computational lexicon. For example, the word *lose* has the sense *fail to keep or to maintain*: "She lost her purse when she left it unattended on her seat" and *miss from one's possessions*: "I have lost my glasses.". In both cases, the word *lose* takes an animate subject and a solid thing as object. The sense distinction comes from the future event specifying whether the object can be found which is a very fine-grained distinction for a computational lexicon. In contrast, Levin's classification is more consistent with the ideal sense distinction criteria proposed by Palmer:

- different predicate argument structures
- different semantic class constraints on verb arguments
- different lexical co-occurrences, such as prepositions

We will discuss the benefits of Levin's classes further in section 2.3 after first introducing Levin's classification.

2.2 Levin's verb classes

Different subcategories of verbs impose different constraints on the number and the type of arguments they take. The syntactic variation can be captured by frames called subcategorization frames (SCFs). For example, the verb *eat* can take a simple NP frame which consists of one subject and one direct object (Example 1).

Example 1 *I ate an apple.*

On the other hand, the verb *put* cannot take this frame (Example 2). It requires at least three syntactic arguments: a subject, a direct object and an indirect object, as shown in the SCF NP-PP in Example 3.

Example 2 *I put the apple** [1]

Example 3 *I put the apple on the table.*

The semantics of the verb is said to determine its syntactic behaviour. To prove the interdependency, Hale and Keyser (1987) presented the example below showing that the verb

[1] An asterisk (*) indicates that the form or construction is not found in natural language.

semantics is a key to the verb syntactic behaviour. Suppose we don't know the meaning of the archaic English word 'gally', as in *The sailor gallied the whales*. Some people might suggest that the word means 'see', as in *The sailor saw the whales*, while others might suggest that it means 'frighten', as in *The sailor frightened the whales*. At this point, we don't know which meaning is correct. To solve this problem, Hale and Keyser looked at the middle transitivity alternation, where the subject of the intransitive usage of the verb can be placed as the object the transitive usage. For example, for the verb 'read', we can say 'this book reads well'; or we can move the subject 'book' to the object position, as in 'I read this book'. The people who believe 'gally' means 'see' would not allow for the sentence 'Whales gally easily', because middle transitivity alternation is not allowed for 'see' (**Whales see easily*). The people who believe the meaning is 'frighten' would allow for the middle transitivity alternation, as in *Whales frighten easily*. In fact, the verbs with middle transitivity alternation would normally cause a change of state, e.g. *frighten, open, split and crush*. From this example, we can see that a verb's behaviour has a very strong relationship with its semantics.

On the basis of the inter-dependency between this type of diathesis alternation (DA) behaviour of verbs and their meanings, Levin (1993) manually created a classification for English verbs. The verbs within the same class share some meaning component, similar syntactic frames and possibly other linguistic features (e.g. zero nominals). Levin's central thesis is that "the behaviour of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning". According to Levin, a verb's syntactic behaviour and semantics are linked. The syntactic frames are understood as a direct reflection of the underlying meaning components which have selectional restrictions on arguments. For example:

- Heat *radiates* from the sun
- The sun *radiates* heat

According to Levin, verbs taking this alternation (substance/source alternation) express substance emission (e.g. *bleed, leak, sweat*). They take two arguments: (i) a source (e.g. *sun*) and (ii) the substance emitted from the source (e.g. *heat*). The semantic role of the subject of the intransitive use of the verb is the same as the semantic role of the object of the transitive use, and the semantic role of the object of intransitive use is the same as that of the subject of the transitive use.

Based on the previous research on DAs (e.g. Pinker (1989); Jackendoff (1990)), Levin defined 78 possible DA types. These alternations concern changes in verbs' transitivity or within the arguments of VP, or involve the introduction of oblique complements, reflexives, passives, there-insertion, different forms of inversions or specific words. They are mainly restricted to verbs taking NP and PP complements.

Then, Levin analysed 3104 verbs according to the alternations, and associated each verb with a set of relevant alternations. She argued that verbs which behave similarly with respect to alternations share certain meaning components, and can thus be grouped together to form a semantically coherent class.

Finally, the verbs with the same or similar alternation behaviour were assigned to the same class. Levin classified the verbs into 49 broad semantically motivated verb classes, some of which divide further into subclasses, making the total number of classes 192.

For example, the class of 'COOK' verbs includes verbs such as *cook, bake, boil, roast, heat* ... which share the following syntactic behaviour and alternations (Levin, 1993):

1. Causative/Inchoative Alternation
   (a) Jennifer baked the potatoes.
   (b) The potatoes baked.
2. Middle Alternation:
   (a) Jennifer baked Idaho potatoes.
   (b) Idaho potatoes bake beautifully.
3. Instrument subject Alternation
   (a) Jennifer baked the potatoes in the oven.
   (b) This oven bakes potatoes well.
4. Resulative Phrase
   (a) Jennifer boiled the pot dry.
   (b) Jennifer baked the potatoes to a crisp.

Besides listing the relevant alternations, Levin also included some alternations that the verbs cannot take. Levin demonstrated that the resulting verb classes can capture the inter-dependency between the syntax and the verb meaning.

Despite the popularity of the classification in NLP applications, Levin's original resource has a few limitations:

1. The classification is not exhaustive in terms of breadth or depth of coverage. More work is needed to cover a larger set of DAs and further to extend and refine verb classification.

2. The classification lacks a comprehensive hierarchical organization of the types found in other computational lexical resources, such as WordNet and FrameNet. The taxonomy is only three levels with both missing upper and lower parts, and only a few classes have subclasses. A more complete taxonomy could benefit various NLP applications.

3. The classification does not provide explicit description of syntactic and semantic properties of member verbs (e.g. only some syntactic frames are listed that participate in alternations).

2.2.1 VerbNet

VerbNet was designed to address these limitations in Levin's original classification. It is currently the most extensive on-line verb lexicon available for English. It provides detailed syntactic-semantic descriptions of Levin classes organized into a refined taxonomy (additional classes and subclasses). Each verb class is described by thematic roles, selectional restrictions, semantic predicates and argument and frame types containing syntactic descriptions. The resource has recently been extended with additional classes for verbs not covered by Levin (Kipper *et al.*, 2006b). It now provides 270 fine-grained classes, which cover a wider number of verb senses and member verbs (6272).

VerbNet has been used to aid a number of NLP applications such as automatic verb acquisition (Swift, 2005), semantic role labelling (Swier and Stevenson, 2004), robust semantic parsing (Shi and Mihalcea, 2005), word sense disambiguation (Dang, 2004), building conceptual graphs (Hensman and Dunnion, 2004), and creating a unified lexical resource for knowledge extraction (Croch and King, 2005), among others. However, the resource is still not comprehensive. It provides no statistical information about the likelihood of different classes for individual verbs, and is not helpful for processing texts in specific domains where verb senses (and thus classes) may only partially overlap with the ones in general language.

2.2.2 Levin's verb classification and NLP applications

Verb classes can help to reduce the redundancy in the lexicon, since verbs in a class share similar properties. They can alleviate the problem of data sparseness which affects many NLP tasks by predicting the properties of member verbs, when not enough empirical evidence is available. One such task is semantic role labelling. Most work on this task takes the supervised approach which relies on a large amount of manually tagged corpora such as PropBank and FrameNet as training data. Swier and Stevenson (2004) take an unsupervised approach instead which takes advantage of the classification in VerbNet.

Experiencer V Cause
Experiencer V Cause Prep(in) Oblique
Experiencer V Oblique Prep(for) Cause

Table 2.1: An example VerbNet entry for the class containing verb *admire*

The probability model is $p(r|v, s, n)$, which is the probability of a semantic role r given the verb v, the slot s and the noun n. Consider the following example sentence:

Kiva[EXPERIENCER] admires Mat[CAUSE].

The model estimates $p(EXPERIENCE|admire, SUBJECT, Kiva)$ for the subject slot. The bootstrapping algorithm initiates the probability model by making initial unambiguous role assignment according to the frames in the relevant VerbNet entry (an example is shown in table 2.1). It then iteratively updates the probability model and makes further assignment to the ambiguous slots. When the evidence for a slot is below a threshold, the model falls back on the statistics over the verb and noun classes. Thus, the backoff model is $p(r|VerbClasss, s, NounClass)$. This creates a simplistic probabilistic model with the emphasis on class generalization. This unsupervised method achieves an error reduction of 50-65% over an informed baseline[2], demonstrating the potential of the approach for a task that has previously relied on large amounts of manually tagged training data.

Verb classifications have also been used to help tasks such as parsing, word sense disambiguation, information extraction, question-answering, and machine translation (Swier and Stevenson, 2004; Dang, 2004; Shi and Mihalcea, 2005; Zapirain *et al.*, 2008; Rios *et al.*, 2011).

2.3 Automatic verb classification

Although Levin-style classes have proved helpful for a number of NLP tasks, large-scale exploitation in real-world or highly domain-sensitive tasks has been limited because no fully accurate or comprehensive lexical classification is available. There is no such resource because manual classification of large numbers of words has proved very time-consuming. In addition, class-based differences are typically manifested in differences in the statistics over usages of syntactic-semantic features. This statistical information is difficult to collect by hand as it is highly domain-sensitive, i.e. it varies with predominant word senses, which change across corpora and domains (e.g. biomedical domain).

In recent years, automatic induction of verb classes from corpus data has become increasingly popular (Merlo and Stevenson, 2001; Korhonen *et al.*, 2003; Schulte im Walde,

[2] The baseline assigns all slots the role with highest probability given the slot class (subject, object, indirect object and PP object).

2006; Joanis *et al.*, 2007; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009b); Some works have focussed on multilingual (Ferrer, 2004; Falk *et al.*, 2012) and domain specific classification (Korhonen *et al.*, 2006b, 2008).

Automatic verb classification is important, because it provides the opportunity to acquire and tune classifications for the application and domain in question. Automatic classification is not only cost-effective but it also gathers important statistical information from corpora and can be applied to new domains, provided relevant data is available.

Various approaches have been proposed for verb classification. Both supervised and unsupervised (including semi-supervised) approaches have been used to classify features extracted from corpus data (raw, tagged or parsed). Although the results have been generally encouraging, the accuracy of automatic classification shows room for improvement. In the following, we review the previous works on automatic verb classification. For each work, the data, features, clustering methods and results are described.

2.3.1 Classification methods and performance

Supervised approaches

A supervised approach assigns verbs into one of several pre-defined verb classes. Supervised methods yield optimal performance where adequate and accurate training data are available. However, such methods require a pre-defined verb classification as part of the training data. Therefore, they are unable to detect new classes, and the cost of applying the method to a new domain is high.

Merlo and Stevenson (2001) presented an automatic classification of English intransitive verbs into just three syntactic classes: *unergative*, *unaccusative* and *object-drop* verbs. The classification was based on the argument structure and the thematic relations of verbs. The features used were transitivity, causativity, animacy, passive voice and verb tense. Each verb was associated with a vector of 5 features. The vectors served as input to a decision tree classification algorithm. 60 verbs are classified into three classes with the accuracy[3] of 69.8%.

Joanis *et al.* (2007) classified 845 verbs into 14 (some coarse, some fine-grained) Levin classes. 11 evaluations were reported for 2-14 way classifications. The features included: syntactic slot, syntactic slot overlap, tense, voice, aspect and animacy. Support Vector Machines (Vapnik, 1995) were used for classification. The best accuracy on the 14 way test was 58.4%.

Sun *et al.* (2008b) classified 204 verbs into 17 Levin classes using rich SCF based features (SCFs parameterized for prepositions, tense and voice). The features were extracted

[3] The evaluation measures used in this section are described in section 2.3.3.

from the automatically acquired VALEX lexicon (Korhonen *et al.*, 2006a). Three classification methods were used: Support Vector Machines, Maximum Entropy and Gaussian method. The best classification accuracy was obtained with the Gaussian method (64%). The best performing feature is SCF parameterized with prepositions. Ó Séaghdha and Copestake (2008) classified Sun *et al.*'s best performing feature using the distributional kernel method. A better accuracy was obtained at 67%.

Li and Brew (2008) investigated a range of feature sets for classifying English verbs. Joanis *et al.* (2007)'s test set and a large-scale test set of 48 Levin classes involving 1300 verbs were used. Bayesian Multinomial Regression (Genkin *et al.*, 2008) was employed for classification. The feature sets included Joanis *et al.*'s feature sets, CO, dependency relation, SCF and their combinations. The SCFs combined with COs gave the best result at 66.3% in accuracy for Joanis *et al.*'s 14-way test-set.

Unsupervised and semi-supervised approaches

Unsupervised methods use clustering to infer verb classes based on the similarity between verbs[4]. Such methods require minimal prior knowledge on the resulting classification: a similarity measure and the number of classes (optional for some methods). This approach has the benefit that it can also be used to discover novel information in corpus data. This is particularly useful for supplementing or improving existing classifications or learning new classifications for languages and domains where no manually built classifications are available. The resulting classification reflects the natural groupings of the data (Jain *et al.*, 1999). This can help to understand the set of features that characterizes the verb classes and the underlying linguistic phenomena. For example, if a clustering experiment shows that DAs naturally group verbs into Levin style classes, it can be an empirical proof for Levin's verb classification.

Sometimes, partial information of the verb classification is available prior to clustering. This information can be incorporated as a guideline for the clustering algorithm. This forms a basis of a semi-supervised approach[5]. Since in the existing verb clustering exper-

[4] There are unsupervised methods other than clustering methods: for example dimensionality reduction techniques such as Non-negative matrix factorization (Seung and Lee, 2001), Principal component analysis (Wold *et al.*, 1987) and Singular value decomposition (Golub and Reinsch, 1970) and Neural Network model (Kohonen, 1990). Because verb classification is a task that groups verbs together, clustering is the most straightforward approach. However, note that some methods that are not traditional clustering methods can be used for clustering, e.g. Non-negative matrix factorization (Ding *et al.*, 2010; Yang and Oja, 2012).

[5] We use the term semi-supervised instead of weakly or lightly supervised learning. In some works, the two terms are used interchangeably (e.g. Jones (2004); Pham *et al.* (2005)). Yet in weakly-supervised learning, the training dataset (the seeds) is typically small. At least one labelled sample is required for each class. In semi-supervised learning, the definition tends to be more flexible: the training data does not need to be small, but the use of a large amount of unlabelled data is the focus (e.g. Søgaard (2011)).

iments, the supervision(e.g. constraints) is always added over an unsupervised method, we included semi-supervised approaches in the current section.

Brew and Schulte im Walde (2002) used spectral clustering algorithm (SPEC) (Ng *et al.*, 2002) to cluster 57 German verbs into 14 classes on the basis of cues related to SCF frequency information. A set of 38 SCFs was employed, some of which were parameterized for prepositions. The frequency information was extracted automatically from the data parsed using a statistical parser. The performance of SPEC was compared to that of the K-means algorithm. The results showed that the SPEC outperformed K-means in all the experiments yielding an F-Measure of 0.48 at best.

Schulte im Walde (2003) performed a larger experiment with 168 German verbs belonging to 43 classes. The features were similar to the ones used in Brew and Schulte im Walde (2002). Some frames were parameterized for SPs. 0.182 adjusted Rand Index was obtained using K-means for classification.

Stevenson and Joanis (2003) classified 260 verbs into 13 Levin's classes using agglomerative hierarchical clustering. The features were similar to the features used in Joanis *et al.* (2007). They experimented with unsupervised, manual and semi-supervised feature selection. The semi-supervised method samples a small set of seed verbs whose classification is known. The features were selected on the basis of the supervised classification result. The semi-supervised feature selection gave the best accuracy at 38%, outperforming the manually selected features.

Korhonen *et al.* (2003) conducted an experiment of classifying 110 highly polysemous English verbs into 34 classes. They achieved 60% modified purity. Like Schulte im Walde (2003), they clustered verbs according to SCF frequency information, but employed a technique capable of dealing with sense variation. SCFs were extracted from corpus data using the SCF acquisition system of Briscoe and Carroll (1997). They parameterized two of the SCFs for prepositions, and applied the Information Bottleneck and the Nearest Neighbours method to assign verbs into classes corresponding to (any of) senses in the corpus data.

Korhonen *et al.* (2006b) extended this system with additional clustering methods (Probabilistic Latent Semantic Analysis and a modified version of the Information Bottleneck) and applied it to the biomedical domain. 192 medium to high frequency verbs were selected from biomedical journals for the experiment. The resulting classification was highly accurate (77% F-measure) and domain specific. Using the same gold standard, Korhonen *et al.* (2008) investigated a wide range of feature sets (e.g. SCF, voice, SP). A novel clustering algorithm - pairwise clustering (Puzicha *et al.*, 2000) was used. The best performing feature was SCFs parameterized for prepositions and lexical preferences (LPs). The LPs were acquired automatically from corpus data using an unsupervised method.

In verb clustering, the supervision is often not available for all the classes. Therefore, we use the term semi-supervised learning.

Vlachos *et al.* (2009b) used the Dirichlet process mixture model to cluster verbs in Sun *et al.* (2008b) test set using SCFs parameterized for prepositions. The result is 57% in V-measure (Rosenberg and Hirschberg, 2007) for the 30 clusters setting against the 17 gold standard classes. Furthermore, a small number of pairwise constraints specifying if two verbs *must link* or *must not link* were added to the algorithm. The performance was improved by 8%.

Falk *et al.* (2012) used existing syntactic and semantic lexical resources to cluster 2183 French verbs. These verbs are from the gold standard classes we developed and introduce in chapter 6 of this thesis. The following features were used: SCF, VerbNet thematic grids, syntactic features (symmetric arguments, predicate, sentential argument, optional object and passive build with *it*) and semantic features (location role, concrete object, asset role and plural role). Incremental Growing Neural Gas (IGNG) with Feature Maximization was used as the clustering method (Lamirel *et al.*, 2011). The method was shown to outperform the K-means, because it is suitable for the relatively small and clean features extracted from the lexicon. 0.70 F-Measure was achieved using a combination of thematic grid, SCF and syntactic features. Detailed discussion on this work can be found in section 6.8.

In addition to the flat clustering, Ferrer (2004) applied hierarchical clustering to 514 Spanish verbs and evaluated against a hierarchical gold standard resembling that of Levin's classification in English (Vázquez *et al.*, 2000). R_{adj} of 0.07 was reported for a 15-way classification, which is comparable to the result of Stevenson and Joanis (2003).

2.3.2 Features and feature extraction

The main feature of manual verb classification is a DA which manifests at the level of syntax as alternating sets of SCFs. Since automatic detection of DAs is challenging (McCarthy, 2001), most work on automatic classification has focused on syntactic features, exploiting the fact that similar alternations tend to result in similar syntactic behaviour. The syntactic features have been shallow syntactic slots (e.g. NPs preceding or following the verb) extracted using a lemmatizer or a chunker, or verb SCFs extracted using a chunker or a parser. These feature types have been refined with information about prepositional preferences (PPs) of verbs. Joanis *et al.* (2007) have reported better results using syntactic slots, while several others have obtained good results using SCFs, e.g. (Schulte im Walde, 2006; Li and Brew, 2008). While SCFs correspond better (than syntactic slots) with the features used in manual work, optimal results have required including in SCFs additional information about adjuncts (not only arguments) of verbs (Sun *et al.*, 2008a) which are typically not used in manual classification.

Recent research has also experimented with replacing or supplementing SCFs with information about basic lexical context (co-occurrences (COs)) of verbs, or LPs in specific grammatical relations (GRs) associated with verbs in parsed data (for example, the type and frequency of prepositions in the indirect object relation) (Li and Brew, 2008). Some experiments have also explored the usefulness of verb tense (e.g. the part-of-speech tags of verbs), voice (the knowledge whether the verb was used in active or passive) and/or aspect for verb classification (Joanis *et al.*, 2007; Korhonen *et al.*, 2008).

While most work has focussed on syntactic or lexical features, a few attempts have also been made to refine syntactic features with semantic information about verb selectional preferences (SPs). Following Merlo and Stevenson (2001); Joanis *et al.* (2007) used a simple 'animacy' feature which was determined by classifying e.g. pronouns and proper names in data to this single SP class. Joanis (2002) employed as SP models the top level WordNet (Miller, 1995) classes (Schulte im Walde (2006) tried a similar approach for German). Finally, combinations of lexical, syntactic, semantic and other features have been explored. We describe below the features and the feature extraction methods used in previous works.

Transitivity

The transitive and intransitive usage of verbs can be used to distinguish certain groups of verbs (Merlo and Stevenson, 2001). For example, among the three groups of verbs: unergative (e.g. *race*), unaccusative (e.g. *melt*) and objective-drop (e.g. *play*), we are expecting the frequency of the transitive usage as follows: object-drop >unaccusative >unergative.

An approximation of this feature was used by Merlo and Stevenson (2001): A verb occurrence preceded by forms of the verb *be*, or immediately followed by a potential object (e.g. noun, pronoun and determiner) was counted as transitive; otherwise, the occurrence was counted as intransitive. As the underlying corpus was tagged but not parsed, the identified transitive/intransitive uses are expected to include some noise.

Causativity

Causativity is related to the causative alternation. Consider the example *The sun melted the ice./The ice melted.* In this alternation, the thematic role of the subject of the intransitive is same as the thematic role of the object of the transitive (Merlo and Stevenson, 2001). For verbs that don't take this alternation, this pattern is infrequent. For example, in the unexpressed object alternation (*Mike ate the cake./Mike ate.*), the thematic role of the subject in the intransitive use is different from the thematic role of the object in the transitive use.

In Merlo and Stevenson (2001), causativity feature was approximated as follows: The subjects and objects of a verb were collected into two multi-sets, respectively. The overlap of these two multi-sets were calculated in order to find out the number of times the same noun was used as both subject and object. For example, given the subject and object multisets {a,a,a,b} and {a}, the overlap is {a,a,a}. The feature value is the ratio between the cardinality of the overlap, and the sum of the cardinality of the subject and object multisets. For the last example, the ratio would be 3/5.

Animacy

Certain groups of verbs can more frequently take agentive subjects. For example, the unergative and object-drop verbs take agentive subjects in both transitive and intransitive. However, the unaccusative verbs assign an agentive role only in transitive. In the intransitive use, the unaccusative verbs assign a theme role to the subject. Thus, the agentive subject is less frequently found with unaccusatives than with unergative or object-drop verbs. Since agents are more likely to be animate nouns than the themes, the unaccusative verbs are expected to be less likely taking animate nouns as subject when compared to unergative and object-drop verbs (Merlo and Stevenson, 2001).

Merlo and Stevenson approximated the animacy of the subjects by all pronouns except *it*. They followed the hypothesis, presented in Silverstein (1976); Dixon (1994), that those pronouns most often refer to animate entities. The pronouns that occur in the subject position were counted. The feature value is the ratio of the count of pronoun subjects to the count of all subjects.

Stevenson and Joanis (2003) and Joanis *et al.* (2007) extended this approach by including proper noun phrases that are labelled as a person by a chunker (Abney and Abney, 1991). The chunker includes a crude named entity recognition system which recognizes person names by matching a list of English first names and titles.

Voice

The passive/active voice of the verb is related to the transitivity alternations. The passive use of the verb implies a transitive use of a verb.

In Merlo and Stevenson (2001), verbs tagged with VBD (the past tense tag) were considered as active voice. A token tagged as VBN (the past participle tag) and the closest preceding auxiliary *be* were considered as passive use. This feature was also used by Stevenson and Joanis (2003) and Joanis *et al.* (2007).

Verb Tense

In analogy to the voice feature, the VBN/VBD tag is related to the transitivity alternation. The past participle form of the verb implies a transitive use of a verb. In addition, the middle voice (found in middle alternation) is often in present tense (Joanis *et al.*, 2007).

In Merlo and Stevenson (2001), the occurrences of VBD/VBN tags are simply counted within the tagged corpus. In Stevenson and Joanis (2003) and Joanis *et al.* (2007), tags VB (base form), VBP (present tense, not third person singular), VBZ (present tense, third person singular) and VBG (present particle) are also included.

Lexical Preferences

This feature encodes the frequency of the syntactic slots corresponding to verbal arguments (e.g. subject and direct and indirect object slots). These argument slots are components of a subcategorization frame, but are here considered independently of their co-occurrence with other slots.

In Stevenson and Joanis (2003) and Joanis *et al.* (2007), these features were extracted from the tagged and NP-chunked BNC corpus. In Joanis *et al.* (2007), the subject slots that were found in transitive and intransitive usages were considered separately. In addition, the frequency of two word classes on the slots were considered which reflect the number of DAs, including the reflexive pronouns (e.g. *Jill dressed hurriedly./Jill dressed herself hurriedly.*) and the words *it/there* (e.g. *A problem developed./There developed a problem.*). Furthermore, the overlap of the word lemmas in the slots was also a feature. This overlap feature was inspired by Causativity feature (Merlo and Stevenson, 2001) described in this section. Since the same semantic argument can occur in different slots in alternating frames, the degree to which these two slots containing the same entities is an indicator of the verb's participation in an alternation. For example, given the alternation *The sky cleared/The clouds cleared from the sky*, an overlap feature of the subject and the indirect object slot is added. The overlap value was calculated in the same way as in Merlo and Stevenson (2001) (described above in the causativity section).

In the recent works, the heads of arguments were detected from grammatical relations found by statistical parsers. In Li and Brew (2008), the lexical heads of arguments were identified using grammatical relations produced by the C&C CCG parser (Clark and Curran, 2007). Korhonen *et al.* (2008) used the RASP parser (Briscoe *et al.*, 2006) to extract grammatical relations.

Subcategorization Frames

DAs show at alternating sets of SCFs (e.g. in the causative/inchoative alternation, an NP frame alternates with an intransitive frame: *Tony broke the window* ↔ *The window broke*). Most work (e.g. Korhonen *et al.* (2003, 2006b, 2008); Li and Brew (2008)) on automatic classification has exploited the fact that similar alternations tend to result in similar SCFs.

Many recent verb classification works (Korhonen *et al.*, 2003, 2006b; Joanis *et al.*, 2007; Korhonen *et al.*, 2008) have used some version of the comprehensive SCF acquisition system originally developed by Briscoe and Carroll (1997). The system makes use of the RASP toolkit (Briscoe *et al.*, 2006). The corpus data are first tokenized, tagged, lemmatised and parsed using RASP. SCF patterns are then extracted from the parser output and classified using a classifier which distinguishes between over 160 SCF types, a superset of those in the ANLT (Boguraev *et al.*, 1987) and COMLEX (Grishman *et al.*, 1994) syntax dictionaries. The SCFs abstract over lexically-governed particles, prepositions and specific predicate selectional preferences. A statistical filtering component may optionally be applied which removes noisy SCFs from the lexicon. A large SCF lexicon, VALEX (Korhonen *et al.*, 2006a), has been constructed using this system. The VALEX lexicon (typically an unfiltered noisy version of the lexicon, as this produced the best results for the task) was used for verb classification by Sun *et al.* (2008b) and Sun *et al.* (2008a).

Li and Brew (2008) extracted SCFs by matching the label of a grammatical relation to a list of syntactic constituents that are found in subcategorization frame definitions. For example, NP1 is the subject of the verb, NP2 is the object of the verb and PP is the prepositional phrase. In the first step, a lexical frame is constructed. For a sentence *he broke the door with a hammer*, the list of grammatical relations include (dobj broke_1 door_3), (dobj with_4 hammer_6), (iobj broken_1 with_4) and (ncsubj broken_1 He_0); the identified lexical frame is NP1(he)-V-NP2(door)-PP(with:hammer). Then, the SCF is constructed as NP1-NP2-PP(with).

SCFs can be further refined by adding lexical, syntactic and semantic information:

SCF+preposition

Some of the SCFs can be parameterized with the prepositional phrase involved. For example, consider the SCF (NP-PP): *she puts the flowers on the table./ she removed the flowers from the table.* As the examples show, *put* and *remove* have same SCF NP-PP. However, *put* prefers a PP headed by *on*, while *remove* prefers a PP headed by *from*. Therefore, to tell the difference between *put* and *remove* (which belong to different verb classes, those of PUTTING and REMOVING verbs) it helps to know the prepositions occurring in their SCFs. In this example, the parameterized SCF would be NP-PP(on) and NP-PP(from).

In Korhonen *et al.* (2003), only two high frequency SCFs (PP and NP-PP) were parameterized with preposition information. In recent works, the parameterization was extended to all the SCFs that involve PPs (e.g. Sun *et al.* (2008b); Korhonen *et al.* (2008)). For example, the SCF NP-P-NP-ING has different variations, depending on the preposition in question, e.g. *he* ***attributed*** *his failure* ***to*** *buying his books; he* ***told*** *her* ***about*** *climbing the mountain*).

(a)

Feature name	Feature value	Type of feature
NP-PP	10	SCF features
PP	20	
VBD	15	Tense features
VBN	15	

(b)

Feature name	Feature value	Type of feature
NP-VBD-PP	6	NP-PP parameterized by POS tag
NP-VBN-PP	4	
VBD-PP	9	PP parameterized by POS tag
VBP-PP	11	

Table 2.2: Example of two types of SCF+tense features

SCF+tense

This feature supplements a SCF with the POS tag of the main verb. The tense feature is related to the passive voice and middle voice of a verb, which are in turn related to certain DAs. Further description can be found in the tense feature section.

In previous works, two variants of this feature have been used. In the first one, the frequencies of POS tags are calculated over all the SCFs of a verb. This feature is actually a simple concatenation of SCF features and tense features. In the second variant, the frequencies of POS tags are calculated specific to each SCF of the verb. These two types of features are exemplified in table 2.2.

SCF+voice

The active and passive voices are related to the transitivity of the verb use (described in the voice feature section). In previous works, two sub-types of the feature were used: 1) the frequency of the active and passive occurrences of the verb calculated over all the SCFs of the verb 2) the frequency of the active and passive occurrences of the verb calculated specific to each SCF of the verb. Table 2.3 illustrates these two feature types.

SCF+lexical preference

This feature supplements a SCF with information about LPs of the verbs in the following slots: subject, direct object, second object, and the NP within the PP complement.

For an example sentence *The sun melted the ice*, the SCF NP is parameterized by the lexical head of the arguments in the subject and direct object slots. The SCF is transformed to nsubj:sun-NP-dobj:ice after parameterization.

(a)

Feature name	Feature value	Type of feature
NP-PP	10	SCF features
PP	20	
ACTIVE	25	Voice features
PASSIVE	5	

(b)

Feature name	Feature value	Type of feature
NP-ACTIVE-PP	9	NP-PP parameterized by voice
NP-PASSIVE-PP	1	
ACTIVE-PP	16	PP parameterized by voice
PASSIVE-PP	4	

Table 2.3: Example of two types of SCF+voice features

The LP features are extracted from the grammatical relation produced by a statistical parser (Li and Brew, 2008; Korhonen *et al.*, 2008). To relieve the sparse data problem, frequency threshold can be added to remove the noise (Korhonen *et al.*, 2008).

SCF+selectional preference

To overcome the data sparseness problem with the LP feature, selectional preferences (SPs) can be used instead of actual nouns. SPs can be strong indicators of DAs (McCarthy and Korhonen, 1998) and fairly precise semantic descriptions, including information about verb selectional restrictions, can be assigned to the majority of Levin classes, as demonstrated by VerbNet (Kipper-Schuler, 2005). SP acquisition from undisambiguated corpus data is arguably difficult (Brockmann and Lapata, 2003; Erk, 2007; Bergsma *et al.*, 2008). The traditional way of SP acquisition is to use a lexical resource. Joanis (2002); Schulte im Walde (2006) employed top level WordNet (Miller, 1995) and GermaNet (Kunze and Lemnitzer, 2002) classes as SP models in verb classification. Joanis (2002) obtained no improvement over syntactic features, whereas Schulte im Walde (2006) obtained insignificant improvement. Korhonen *et al.* (2008) combined SPs with SCFs when clustering biomedical verbs. The SPs were acquired automatically from syntactic slots of SCF using PC clustering. The SP clusters offered no improvement over the SCF+LP features.

As an example, if word w_1 is in SP cluster c_1 and word w_2 is in cluster c_2, the original SCF+LP feature ncsubj:w_1_NP_dobj:w_2 is transformed to ncsubj:c_1_NP_dobj:c_2. If w_3 is also in the cluster c_2, then the SCF+LP feature ncsubj:w_1_NP_dobj:w_3 would have the same representation in SCF+SP.

Co-occurrence

Co-occurrence (CO) shows the words that occur in the context of a verb. The verbs that occur in similar contexts tend to have similar meaning, according to the distributional hypothesis (Harris, 1954). However, COs are generally not considered particularly sensitive to argument structure (Rohde *et al.*, 2004).

This feature was introduced by Li and Brew (2008). The word lemmas that occur in the fixed length window around verb are collected as features. A stopword list was used to filter out function words. Li and Brew also used an extended feature which integrated some syntactic information: 1) All the prepositions are kept, as they are known to carry information about the lexical meaning of the verb. 2) All verbs that occur in the context of the target verb are replaced with their POS tags. Li and Brew assumed that most verbs tend to have a strong selectional preference for their nominal arguments, but not for their verbal arguments.

Table 2.4 summarizes all the existing features and the extraction method that are reused in this thesis. Table 2.5 shows an example for each feature type.

2.3.3 Evaluation measures

Many evaluation measures have been used in verb classification experiments. In this section, we focus on clustering evaluation measures that have been used in unsupervised verb classification experiments. Typical objective functions in clustering have a goal of attaining high intra-cluster similarity and low inter-cluster similarity. This is an internal criterion for the quality of the clustering (Manning *et al.*, 2008). However, a good result by an internal criterion might not match the gold standard which is an indicator of the effectiveness in application. Therefore, we describe the external criterion that evaluates how well the clustering matches the gold standard. We will omit evaluation measures which do not consider the gold standard, for example: the Mean Silhouette (Stevenson and Joanis, 2003) and Cumulative Micro-Precision (Falk *et al.*, 2012). These evaluation measures are forms of internal criterion which do not reflect the effectiveness of the clustering in application.

Modified Purity, Weighted class accuracy and F-Measure have been used in many previous verb clustering experiments, e.g. by Korhonen *et al.* (2008); Ó Séaghdha and Copestake (2008).

Modified purity (mPUR) is a global measure which evaluates the mean precision of clusters. Each cluster is associated with its prevalent class. The number of verbs in a cluster K that take this class is denoted by $n_{prevalent}$(K). Verbs that do not take it are considered as errors. Clusters where $n_{prevalent}$(K) = 1 are disregarded, so as not to introduce a bias

Feature Name	Description and extraction method
F-CO	Co-occurrence (Li and Brew, 2008)
F-LP	Lexical preference, extracted as in Korhonen *et al.* (2008) using RASP parser
F-PP	Prepositional preference, a subset of F-LP which only include the type and frequency of prepositions in the indirect object relation
F-SCF	Basic SCF, extracted using Preiss *et al.* (2007)'s system
F-SCF(B)	Basic SCF feature, as in Sun *et al.* (2008b), extracted from the VALEX lexicon.
F-SCF+CO	The concatenation of the F-SCF and F-CO
F-SCF+TENSE(A)	F-SCF with the tense of the verb. The frequency of verbal POS tags is calculated over all SCFs (Korhonen *et al.*, 2008).
F-SCF+TENSE(B)	Same as above, but the frequency of verbal POS tags is calculated specific to each SCF.
F-SCF+VOICE(A)	F-SCF with the active/passive voice of the verb. The frequency of the voice is calculated over all SCFs (Korhonen *et al.*, 2008)
F-SCF+VOICE(B)	Same as above, but the frequency of voice is calculated specific to each SCF.
F-SCF+PP(A)	F-SCF with two high frequency PP frames parameterized for prepositions: the PP and NP-PP frames (Korhonen *et al.*, 2008).
F-SCF+PP(B)	F-SCF with all PP frames parameterized for prepositions (Korhonen *et al.*, 2008).
F-SCF+LP(A)	F-SCF is parameterized by the F-LP in all argument slots (Korhonen *et al.*, 2008).
F-SCF+LP(B)	Filter F-SCF+LP(A) by only keeping those raw argument head types which occur with four or more verbs with frequency of ≥ 3 (Korhonen *et al.*, 2008).
F-SCF+SP(A)	F-SCF is parameterized by the F-SP in all argument slots. As in Korhonen *et al.* (2008), the SPs are acquired automatically by clustering the argument head. The number of clusters was set to 10.
F-SCF+SP(B)	The number of clusters was set to 20.
F-SCF+SP(C)	The number of clusters was set to 50.

Table 2.4: Summary of previously proposed features that are reused in this thesis

towards singletons:

$$\text{mPUR} = \frac{\sum_{n_{prevalent(k_i)}>2} n_{prevalent(k_i)}}{\text{number of verbs}}$$

The second measure is weighted class accuracy (ACC): the proportion of members of dominant clusters DOM-CLUST$_i$ within all classes c_i.

$$\text{ACC} = \frac{\sum_{i=1}^{C} \text{verbs in DOM-CLUST}_i}{\text{number of verbs}}$$

Feature Name	Example features
F-CO	she_-1, flower_+2, on_+3 ...
F-LP	nsubj:she, dobj:flower, iobj:on, idobj:table
F-PP	iobj:on
F-SCF	NP-PP
F-SCF+CO	NP-PP, she_-1, flower_+2, on_+3 ...
F-SCF+TENSE(A)	NP-PP, VBZ
F-SCF+TENSE(B)	NP-PP_VBZ
F-SCF+VOICE(A)	NP-PP, ACTIVE
F-SCF+VOICE(B)	NP-PP_ACTIVE
F-SCF+PP	NP-PP:on
F-SCF+LP	she-NP:flower-PP:on_table
F-SCF+SP	SP1-NP:SP5-PP:on_SP7

Table 2.5: Example of features extracted from the sentence *She puts the flower on the table.* SP noun cluster 1 contains pronouns like he, she and they; Cluster 5 contains nouns like flower, vine and weed; Cluster 7 contains nouns like table, chair and bench.

mPUR and ACC can be seen as a measure of precision(P) and recall(R) respectively. F-measure is calculated as the harmonic mean of P and R:

$$\mathrm{F} = \frac{2 \cdot \mathrm{mPUR} \cdot \mathrm{ACC}}{\mathrm{mPUR} + \mathrm{ACC}}$$

If the number of clusters is not pre-defined, a high F-Measure would be easily achieved, as the mPUR and ACC tend to increase when the number of clusters is large. Information theory based evaluation measures (e.g. NMI, V-MEASURE (Rosenberg and Hirschberg, 2007)) have been used to compare the verb clustering results with different numbers of clusters (Vlachos *et al.*, 2009b). NMI measures the amount of statistical information shared by two random variables representing the clustering result and the gold standard labels. Given random variables A and B:

$$\mathrm{NMI}(A,B) = \frac{I(A;B)}{[H(A)+H(B)]/2}$$

$$\mathrm{I}(A,B) = \sum_{k}\sum_{j} \frac{|(v_k \cap c_j|}{N} \log \frac{N|v_k \cap c_j|}{|v_k||c_j|}$$

where $|v_k \cap c_j|$ is the number of shared members between cluster v_k and gold standard class c_j. Geiß (2011) proved that V-MEASURE is equivalent to NMI (they are identical under a certain condition).

Adjusted rand index (R_{adj}) was employed in Schulte im Walde (2006). The R_{adj} views clustering as a series of decisions, one for each pair of verbs. It measures the disagreement and agreement of these pairs. Given a gold standard with G classes, and a clustering with C clusters, a $C \times G$ contingency table N defines the agreement between gold standard

Data	Work	Method	Result(in F)
Joanis *et al.* (2007)	Li et al. 2008	supervised	66.3
	Joanis et al. 2008	supervised	58.4
	Stevenson et al. 2003	semi-supervised unsupervised	29 31
Sun *et al.* (2008b)	Sun et al. 2008	supervised unsupervised	62.50 51.6
	Ó Séaghdha et al. 2008	supervised	67.3

Table 2.6: Previous verb classification results on two small gold standards

and the clustering. n_{ij} is the size of the intersection between class i and cluster j. The formula of R$_{adj}$ is (Hubert and Arabie, 1985):

$$\mathrm{R}_{adj} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}/\binom{n}{2}}{\frac{1}{2}[\sum_i\binom{n_{i\cdot}}{2} + \sum_j\binom{n_{\cdot j}}{2}] - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}/\binom{n}{2}}$$

2.4 The current challenges

In table 2.6, we list a few previous classification results on two small gold standards[6]. Although improved accuracy can be observed in more recent works, there is still much room for further improvement. Even the supervised approaches do not achieve performance better than an accuracy of 70. Large-scale, cross-lingual and domain specific experiments are also needed. We discuss a few important challenges currently faced by verb classification. The first three challenges are addressed in this thesis. The last challenge – polysemy is discussed in the future work section (section 8.2).

2.4.1 Semantic information

The current features used in verb classification are mainly syntactic in nature, e.g. SCF-based and lexical features. However, the member verbs of Levin's classes are not only similar in terms of their syntactic behaviour, but also share meaning components. Levin discusses selectional preferences with many verb classes and fairly precise semantic descriptions. Information about verb selectional restrictions has been assigned to the majority of VerbNet classes. It seems intuitive that semantic features should be useful for verb classification.

Previous works have mainly experimented with SPs acquired from a lexicon, because the automatic acquisition of SPs from corpus is challenging (Brockmann and Lapata, 2003;

[6] Note that although the results are obtained using the same gold standards, the results are not mutually fully comparable because of differences in corpus data, features, and the number of test verbs actually used in experiments. However, they serve to show the upper bound of previous works in English.

Erk, 2007; Bergsma *et al.*, 2008). In these previous works, SPs do not offer improvement over the syntactic features (Joanis, 2002; Schulte im Walde, 2006; Korhonen *et al.*, 2008). In verb classification, SPs are needed for argument slots of SCFs. The data may be sparse, and the resulting feature space is very high-dimensional. This large feature space is a challenge for the current clustering methods. New clustering methods that are good in handling high-dimensional feature space may help.

2.4.2 Hierarchical classification

The existing gold standards for verb classification (e.g. Levin's classification and VerbNet) are hierarchical in nature. Different levels of sub-classes form a tree structure. Yet current works mostly focus on flat classification of verbs. Some perform multiple levels of flat clustering and evaluate against a hierarchical gold standard (Korhonen *et al.*, 2008), and others perform hierarchical clustering but evaluate against a flat gold standard (e.g. Stevenson and Joanis (2003)). We found that only Ferrer (2004) performed hierarchical clustering on Spanish verbs, and evaluated against a small hierarchical gold standard. Moreover, all previous works use linkage hierarchical clustering. This method has a few problems, e.g. the cut-off value is difficult to determine (Stevenson and Joanis, 2003). A recent graph-based method (Yu *et al.*, 2006) avoids some problems of the linkage method, and performs better on many tasks. Yu *et al.*'s method can be a good starting point for improving the performance of hierarchical verb clustering.

2.4.3 Task-based evaluation

The manually created verb classification in VerbNet has proved useful for many practical NLP tasks. However, to our knowledge, automatic verb classification has not been evaluated in the context of a NLP task yet. Automatic classification is typically evaluated against a manual gold standard. Although gold standard evaluation gives some idea of an accuracy of the method, few gold standards are perfect and the required accuracy or ideal granularity of a verb classification may change from one task to another. Moreover, some tasks involve specific domains for which no manually built classifications are available that could serve as a gold standard. It is therefore important to evaluate verb clustering in the context of tasks.

2.4.4 Polysemy

Polysemy is frequent in language. It is estimated that 25% of the verbs in VerbNet are polysemous (Abend *et al.*, 2008). In particular, many high frequency verbs have several senses and can be members of several classes. Most work on automatic classification has bypassed this issue by assuming a single class for each verb – usually the one corresponding to its predominating (the most frequent sense) in language according to e.g. WordNet. This is not only an oversimplified model for the real-world application of verb classes but also the predominating sense is not static but varies across domains and sub-languages.

Few attempts have been made to address this problem. Korhonen *et al.* (2003) performed a clustering experiment with highly polysemous verbs. They constructed a polysemous gold standard for around 200 English verbs and examined whether a soft clustering method (Information Bottleneck) could be used to assign these verbs to several classes. The clustering turned out hard, with the majority of verbs being assigned to one class only. Yet the investigation showed that polysemy has a considerable impact on verb classification: optimal results were obtained when clustering was evaluated against the polysemous gold standard, not the monosemous version of it which assumed the predominant sense according to WordNet.

Clearly polysemy is an issue that needs to be dealt with, and this amounts to both extending gold standards to capture non-predominant senses as well as finding a suitable ML method. A multi-label classification method was used for supervised adjective classification (Boleda *et al.*, 2007) which might yield useful results also with verbs. For unsupervised learning, we argue that the soft clustering methods (e.g. Gaussian mixture model and EM) are not suitable for modelling polysemy. The probability of a verb belonging to more than one cluster can not be modelled by these soft clustering approaches. Probabilistic clustering models which explicitly model the overlap between lexical categories might be of use (Heller *et al.*, 2008). We discuss this in detail in the section 8.2.2.

Chapter 3

Verb clustering using selectional preferences

3.1 Introduction

[1] As discussed in previous chapters, both supervised and unsupervised machine learning (ML) methods have been proposed for verb classification and used to classify a variety of features extracted from raw, tagged and/or parsed corpus data. The best performing features on cross-domain verb classification have been syntactic in nature (e.g. syntactic slots, SCFs). Disappointingly, semantic features have not yielded significant additional improvement, although they play a key role in manual and theoretical work on verb classification and could thus be expected to offer a considerable contribution to classification performance.

We further investigate the potential of semantic features – verb SPs – for the task. We introduce a novel approach to verb clustering which involves the use of (i) a SCF acquisition system by (Preiss *et al.*, 2007) which produces rich lexical, SCF and syntactic data, (ii) novel syntactic-semantic feature sets extracted from this data which incorporate a variety of linguistic information, including SPs, and (iii) a new variation of spectral clustering based on the MNCut algorithm (Maila and Shi, 2001) which is well-suited for dealing with the resulting, high dimensional feature space.

Using this approach, we show on two well-established test sets that automatically acquired SPs can be highly useful for verb clustering. They yield high performance when used in combination with syntactic features. We obtain our results using a fully unsupervised approach to SP acquisition which differs from previous approaches employed in verb classification in that it does not exploit WordNet (Miller, 1995) or other lexical resources. It is based on clustering argument head data in the grammatical relations associated with verbs.

[1] The research reported in this chapter was published in Sun and Korhonen (2009).

We describe our features in this section and the clustering methods in section 3.2. Experimental evaluation and results are reported in sections 3.3 and 3.4, respectively. Section 3.5 provides discussion and describes related work, and section 3.6 concludes.

SP acquisition from undisambiguated corpus data is arguably challenging (Brockmann and Lapata, 2003; Erk, 2007; Bergsma *et al.*, 2008). It is especially so in the context of verb classification where SP models are needed for specific syntactic slots for which the data may be sparse, and the resulting feature vectors integrating both syntactic and semantic features may be high dimensional. However, we wanted to investigate whether better results could be obtained if the features were optimised for richness, the feature extraction for accuracy, and a clustering method capable of dealing with the resulting high dimensional feature space was employed.

3.1.1 Feature extraction

We adopted SCF acquisition system which has proved more accurate than previous comparable systems for English [2] but which has not been employed for verb clustering before: the system of Preiss *et al.* (2007). This system tags, lemmatizes and parses corpus data using the current version of the RASP toolkit (Briscoe *et al.*, 2006), and on the basis of resulting grammatical relations (GRs) assigns each occurrence of a verb to one of 168 verbal SCFs classes[3].

The system provides a filter which can be used to remove adjuncts from the resulting lexicon. We do not employ this filter since adjuncts have proved informative for verb classification (Sun *et al.*, 2008b; Joanis *et al.*, 2007). However, we do frequency-based thresholding to minimise the noise (e.g. erroneous SCFs) and sparse data in verb classification and to ensure that only features supported by several verbs are used in classification: we only consider SCFs and GRs which have frequency larger than 40 with 5 or more verbs[4].

The system produces a rich lexicon which includes raw and processed input sentences and provides a variety of material for verb clustering, including e.g. (statistical) information related to the POS tags, GRs, SCFs, argument heads, and adjuncts of verbs. Using this material, we constructed a wide range of feature sets for experimentation, both shallow and deep syntactic and semantic features. As described below, some of the feature types have been employed in previous works and some are novel.

F1:	F-CO
F2:	F-PP
F3:	F-LP
F4:	F-SCF
F5:	F-SCF+CO
F6:	F-SCF+TENSE(B)
F7:	F-SCF+PP(B)
F8:	F-SCF(B)

Table 3.1: The mapping to the features in table 2.4. F8 is extracted from the VALEX lexicon (Korhonen *et al.*, 2006a) for the comparison to Preiss *et al.* (2007)'s SCF acquisition system.

3.1.2 Feature sets

Table 3.1 provides the mapping of our features to the features showed earlier in table 2.4 which were used in previous experiments. F8 was included to enable comparing the contribution of the SCF system to that of an older, comparable system which was used for constructing the VALEX lexicon (Korhonen *et al.*, 2006a).

The following 9 feature sets are novel, so they are not in table 2.4. They build on F7, refining it further. F9-F11 refine F7 with information about LPs:

F9: F7 with F3 (subject only)

F10: F7 with F3 (object only)

F11: F7 with F3 (subject, object, indirect object)

F12-17 refine F7 with SPs. We adopt a fully unsupervised approach to SP acquisition. We acquire the SPs by

1. taking the GR relations (subject, object, indirect object) associated with verbs,
2. extracting all the argument heads in these relations which occur with frequency > 20 with more than 3 verbs, and
3. clustering the resulting N most frequent argument heads into M classes using the SPEC method described in the following section.

[2] See Preiss *et al.* (2007) for the details of evaluation.

[3] We used an implementation of the SCF classifier provided by Paula Buttery.

[4] These and other threshold values mentioned in this chapter were determined empirically on corpus data.

We tried the N settings $\{200, 500\}$ and the M settings $\{10, 20, 30, 80\}$. The best settings $N = 200, M = 20$ and $N = 500, M = 30$ are reported. We enforce the features to be shared by all the potential members of a verb class. The expected class size is approximately N/K, and we allow for 10% outliers (the features occurring less than $(N/K) \times 0.9$ verbs are thus removed).

The resulting SPs are combined with SCFs in a similar fashion as LPs are combined with SCFs in F9-F11:

F12-F14: as F9-F11 but SPs (20 clusters from 200 argument heads) are used instead of LPs

F15-F17: as F9-F11 but SPs (30 clusters from 500 argument heads) are used instead of LPs

All the features (including features used in other chapters) are summarized in appendix A.

3.2 Clustering methods

We use two clustering methods: (i) pairwise clustering (PC), which obtained the best performance in comparison with several other methods in work on biomedical verb clustering (Korhonen *et al.*, 2008), and (ii) a method which is new to the task (and to the best of our knowledge, to NLP): a variation of spectral clustering which exploits the MNCut algorithm (Maila and Shi, 2001) (SPEC). SPEC has been shown to be effective for high dimensional and non-convex data in NLP (Chen *et al.*, 2006) and it has been applied to German verb clustering by Brew and Schulte im Walde (2002). However, previous work has used Ng *et al.* (2002)'s algorithm, while we adopt the MNCut algorithm. The latter has shown a wider applicability (von Luxburg, 2007; Verma and Meila, 2003) and it can be justified from the random walk view, which has a clear probabilistic interpretation.

Clustering groups a given set of items (verbs in our experiment) $V = \{v_n\}_{n=1}^{N}$ into a disjoint partition of K classes $I = \{I_k\}_{k=1}^{K}$. Both our algorithms take a similarity matrix as input. We construct this from the skew divergence (Lee, 2001). We choose this similarity measure by following Brew and Schulte im Walde (2002)'s work on German verb clustering. An alternative distributional similarity measure is Jensen-Shannon divergence (Lin, 1991). These two measures are compared in appendix B.

The skew divergence between two feature vectors v and v' is:

$$d_{skew}(v, v') = D(v' || a \cdot v + (1 - a) \cdot v') \tag{3.1}$$

where D is the KL-divergence. v is smoothed with v'. The level of smoothing is controlled by a whose value is set to a value close to 1 (e.g. 0.9999). We symmetrize the skew divergence as follows:

$$d(v,v')_{sskew} = \frac{1}{2}(d_{skew}(v,v') + d_{skew}(v',v))$$

SPEC is typically used with the Radial Basis Function (RBF) kernel. We adopt a new kernel similar to the symmetrized KL divergence kernel (Moreno *et al.*, 2004) which avoids the need for scale parameter estimation.

$$w(v,v') = \exp(-d_{sskew}(v,v'))$$

The similarity matrix W is constructed where $W_{ij} = w(v_i, v_j)$.

3.2.1 Pairwise clustering

PC (Puzicha *et al.*, 2000) is a method where a cost criterion guides the search for a suitable partition. This criterion is realized through a cost function $H(S,M)$ where

(i) $S = \{\mathrm{sim}(a,b)\},\ a,b \in A$: a collection of pairwise similarity values, each of which pertains to a pair of data elements $a,b \in A$.

(ii) $M = (A_1, \ldots, A_k)$: a candidate clustering configuration, specifying assignments of all elements into the disjoint clusters (that is $\cup A_j = A$ and $A_j \cap A_{j'} = \phi$ for every $1 \le j < j' \le k$).

The main idea underlying the clustering criteria is the preference of configurations in which similarity of elements within each cluster is generally high and similarity of elements that are not in the same cluster is correspondingly low.

The cost function is defined as follows:

$$\begin{aligned} &H = -\sum n_j \cdot \mathrm{Avgsim}_j\,, \\ &\mathrm{Avgsim}_j = \tfrac{1}{n_j \cdot (n_j - 1)} \sum_{\{a,b \in A_j\}} \mathrm{sim}(a,b) \end{aligned}$$

where n_j is the size of the j^{th} cluster and Avgsim_j is the average similarity between cluster members. We used the skew divergence as the similarity measure.

3.2.2 Spectral clustering

In SPEC, the similarities W_{ij} are viewed as the weight on the edges ij of a graph G over V. The similarity matrix W is thus the adjacency matrix for G. The degree of a vertex i is $d_i = \sum_{j=1}^{N} w_{ij}$. A cut between two partitions A and A' is defined to be $\mathrm{Cut}(A,A') = \sum_{m \in A, n \in A'} W_{mn}$.

In the MNCut algorithm, the similarity matrix W is transformed to a stochastic matrix P.

$$P = D^{-1}W \tag{3.2}$$

The degree matrix D is a diagonal matrix where $D_{ii} = d_i$.

It was shown by Maila and Shi (2001) that if P has the K leading eigenvectors that are piecewise constant[5] with respect to a partition I^* and their eigenvalues are not zero, then I^* minimizes the multiway normalized cut(MNCut):

$$\text{MNCut}(I) = K - \sum_{k=1}^{K} \frac{\text{Cut}(I_k, I_k)}{\text{Cut}(I_k, I)} \tag{3.3}$$

P_{mn} can be interpreted as the transition probability between vertices m, n. The criterion can thus be expressed as:

$$\text{MNCut}(I) = \sum_{k=1}^{K}(1 - P(I_k \rightarrow I_k|I_k)) \tag{3.4}$$

which is the sum of transition probabilities across different clusters. The criterion finds the partition where the random walks are most likely to happen within the same cluster.

In practice, the K leading eigenvectors of P are not piecewise constant. But we can extract the partition by finding the approximately equal elements in the eigenvectors using a clustering algorithm like K-Means.

The numerator of MNCut is similar to the cost function of PC. The main differences between the two algorithms are: 1) MNCut takes into account of cross cluster similarity, while PC does not. 2) PC optimizes the cost function using deterministic annealing, whereas SPEC uses eigensystem decomposition.

The SPEC algorithm is based on the MNCut algorithm (Maila and Shi, 2001).

[5] The eigenvector v is piecewise constant with respect to I if $v(i) = v(j) \forall i, j \in I_k$ and $k \in 1, 2...K$

```
Input: Dataset $S$, Number of clusters $K$
    1. Compute similarity matrix $W$ and Degree matrix $D$
    2. Construct stochastic matrix $P$ using equation 3.2
    3. Compute the eigenvalues and eigenvectors $\{\lambda_n, x_n\}_{n=1}^{N}$ of $P$, where $\lambda_n \geq \lambda_{n+1}$,
       form a matrix $X = [x_2, \ldots, x_k]$ by stacking the eigenvectors in columns.
    4. Form a matrix $Y$ from $X$ by normalizing the row sums to have norm 1:
       $Y_{ij} = X_{ij} / (\sum_j X_{ij}^2)^{\frac{1}{2}}$
    5. Consider the row of $Y$ to be the transformed feature vectors for each verb
       and cluster them into clusters $C_1 \ldots C_k$ using $K$-means clustering algorithm.
Output: Clusters $C_1 \ldots C_k$
```

3.3 Experimental evaluation

3.3.1 Test sets

We employed two test sets which have been used to evaluate previous work on English verb classification:

T1 The test set of Joanis *et al.* (2007) provides a classification of 835 verbs into 15 (some coarse, some fine-grained) Levin classes. 11 tests are provided for 2-14 way classifications. We employ the 14-way classification because this corresponds the closest to our target (Levin's fine-grained) classification[6]. We select 586 verbs according to Joanis *et al.*'s selection criteria, resulting in 10-120 verbs per class. We restrict the class imbalance to 1:1.5[7]. This yields 205 verbs (10-15 verbs per class), which is similar to the sub-set of T1 employed by Stevenson and Joanis (2003).

T2 The test set of Sun *et al.* (2008b) classifies 204 verbs to 17 fine-grained Levin classes, so that each class has 12 member verbs.

Table 3.2 shows the classes in T1 and T2. The class names in the first column and the number in the second column correspond to classes in Levin's classification. For example, the class 51.3.2 (Run) contains verbs *fly, gallop, glide, jog, march, run, slide, stroll, swim, travel, trot, walk*. Joanis *et al.* (2007) did not include the actual verbs used in the experiment. Thus, we show the list of verbs used in T1 in appendix D.

[6] However, the correspondence is not perfect, with half of the classes including two or more of Levin's fine-grained classes.

[7] Otherwise, in the case of a large class imbalance the evaluation measure would be dominated by the classes with large population.

T1		T2	
Object Drop	26.1, 26.3, 26.7	Remove	10.1
Recipient	13.1, 13.3	Send	11.1
Admire	31.2	Get	13.5.1
Amuse	31.1	Hit	18.1
Run	51.3.2	Amalgamate	22.2
Sound	43.2	Characterize	29.2
Light & Substance	43.1,43.4	Peer	30.3
Cheat	10.6	Amuse	31.1
Steal & Remove	10.5,10.1	Correspond	36.1
Wipe	10.4.1, 10.4.2	Manner of speaking	37.3
Spray / Load	9.7	Say	37.7
Fill	9.8	Nonverbal expression	40.2
Putting	9.1-6	Light	43.1
Change of State	45.1, 45.2, 45.3, 45.4	Other change of state	45.4
		Mode with motion	47.3
		Run	51.3.2
		Put	9.1

Table 3.2: Levin classes in T1 and T2

3.3.2 Data processing

For each verb in T1 and T2, we extracted all the occurrences (up to 10,000) from the raw corpus data gathered originally for constructing the VALEX lexicon (Korhonen *et al.*, 2006a). The data was gathered from five corpora, including the BNC (Leech, 1992), the Guardian corpus, the Reuters corpus (Rose *et al.*, 2002), the North American News Text Corpus (Graff, 1995) and the data used for two Text Retrieval Evaluation Conferences[8] (TREC-4 and TREC-5). The average frequency of verbs in T1 was 1448 and T2 2166, showing that T1 is a more sparse dataset.

The data was first processed using the feature extraction module. Table 3.3 shows (i) the total number of features in each feature set and (ii) the average per verb in the resulting lexicons for T1 and T2.

We normalized the feature vectors by the sum of the feature values before applying the clustering techniques. Since both clustering algorithms have an element of randomness, we run them multiple times. The step 5 of SPEC (K-means) was run for 50 times. The result that minimizes the distortion (the distances to cluster centroid) is reported. PC was run 20 times, and the results are averaged.

[8] http://trec.nist.gov/data/docs_eng.html

		T1		T2	
		total	avg	total	avg
CO	F1	1328	764	743	382
LP (p)	F2	61	37	55	25
LP (all)	F3	2521	526	1481	295
SCF	F4	88	46	86	38
SCF+CO	F5	1466	833	856	422
SCF+POS	F6	319	114	299	87
SCF+P	F7	282	96	273	76
SCF (V)	F8	-	-	92	45
SCF+LP (s)	F9	1747	324	1474	225
SCF+LP (o)	F10	2817	424	2319	279
SCF+LP (all)	F11	4250	649	3515	426
SCF+SP20 (s)	F12	821	235	690	145
SCF+SP20 (o)	F13	792	218	706	135
SCF+SP20 (all)	F14	1333	357	1200	231
SCF+SP30 (s)	F15	977	274	903	202
SCF+SP30 (o)	F16	1026	273	1012	205
SCF+SP30 (all)	F17	1720	451	1640	330

Table 3.3: (i) The total number of features and (ii) the average per verb for all the feature sets. In the feature name, *s* means subject slot; *o* means object and indirect object slot; *p* means the preposition and *all* indicates information on subject, object and indirect object slots. Appendix A contains more details on each feature.

3.3.3 Evaluation measures

To facilitate meaningful comparisons, we employed the *m*PUR, ACC and F as used e.g. by Korhonen *et al.* (2008) and Ó Séaghdha and Copestake (2008). These measures are described earlier in section 2.3.3.

The random baseline (BL) is calculated as follows:

$$\text{BL} = 1/\text{number of classes}$$

3.3.4 Statistical significance test

We performed one-tailed McNemar's test (McNemar, 1947) on the major findings. The test is widely used in previous NLP (Escudero *et al.*, 2000; Chambers *et al.*, 2007; Guo *et al.*, 2011a) and machine learning studies (Dietterich (1998) and Salojärvi *et al.* (2005)). The result of the test is equivalent to Cochran's Q test (Cochran, 1950) when the number of treatments (number of clusterings) is two (Tate and Brown, 1970). Since the test requires the response variable to be binary, we extend the test as in Dietterich (1998).

Number of verb pairs misclustered in both A and B	Number of verb pairs misclustered in A but not in B
Number of verb pairs misclustered in B but not in A	Number of verb pairs clustered correctly in both A and B

Table 3.4: The example contingency table for McNemar's test

For two clustering results A and B, we record how each verb pair was clustered and construct the contingency table 3.4.

The value in the table is calculated as following: we convert the clustering result to a set of binary indicators on paired verbs. For example, given verbs v_1, v_2 and v_3, we will have pairs: (v_1, v_2), (v_1, v_3) and (v_2, v_3). A pair is considered to be clustered correctly if two verbs that are in the same cluster are also in the same class as in the gold standard, or two verbs that are not in the same cluster are also not in the same class as in gold standard. We report the statistical test decision directly, e.g. "significant at $p < 0.05$" or simply "$p < 0.05$".

3.4 Results

3.4.1 Quantitative evaluation

Table 3.5 includes the F-measure results for all feature sets when the two methods (PC and SPEC) are used to cluster verbs in the test sets T1 and T2, respectively.

A number of tendencies can be observed in the results. Firstly, the results for T2 are clearly better than those for T1. Including a higher number of verbs lower in frequency from classes of variable granularity, T1 is probably a more challenging test set than T2. T2 is controlled for the number and frequency of verbs to facilitate cross-class comparisons. While this may contribute to better results, T2 is a more accurate test set for us in the sense that it offers a better correspondence with our target (fine-grained Levin) classes.

Secondly, the difference between the two clustering methods is clear: the new SPEC outperforms PC on both test sets and across all the feature sets except F2, F6 and F7. The average improvement in F is 10.49 ($p < 0.05$). The performance of the two methods is still fairly similar with the more basic, less sparse feature sets (F1-F2, F4, F6-7) but when the more sophisticated feature sets are used (F3, F5, F9-F17) SPEC performs considerably better. This demonstrates that it is a better suited method for high dimensional feature sets.

Comparing the feature sets, the simple co-occurrence based F1 performs significantly better than the random baseline ($p < 0.05$). F2 and F3 which exploit lexical data in the

		T1		T2	
		PC	SPEC	PC	SPEC
	BL	7.14	7.14	5.88	5.88
CO	F1	15.62	33.85	17.86	40.94
LP (p)	F2	40.40	38.97	50.98	49.02
LP (all)	F3	42.94	47.50	41.08	74.55
SCF	F4	34.22	36.16	52.33	57.78
SCF+CO	F5	26.43	28.70	19.52	29.10
SCF+POS	F6	36.14	34.75	44.44	46.70
SCF+P	F7	43.57	43.85	63.40	63.28
SCF (V)	F8	-	-	34.08	38.30
SCF+LP (s)	F9	47.72	56.09	65.94	71.65
SCF+LP (o)	F10	43.09	48.43	57.11	73.97
SCF+LP (all)	F11	45.87	54.63	56.30	72.97
SCF+SP20 (s)	F12	46.67	**57.75**	39.52	71.67
SCF+SP20 (o)	F13	44.95	51.70	40.76	70.78
SCF+SP20(all)	F14	48.19	55.12	39.68	73.09
SCF+SP30 (s)	F15	45.89	56.10	64.44	**80.35**
SCF+SP30 (o)	F16	42.01	48.74	52.75	70.52
SCF+SP30(all)	F17	46.66	52.68	51.07	68.67

Table 3.5: Results on testsets T1 and T2

argument head positions of GRs prove significantly better than F1 ($p < 0.05$). F3 yields surprisingly good results on T2: it is the second best feature set on this test set. Also on T1, F3 performs significantly better than the SCF-based feature sets F4-F7 ($p < 0.05$). This demonstrates the usefulness of lexical data when obtained from argument positions in relevant GRs.

Our basic SCF feature set F4 performs significantly better than the comparable feature set F8 obtained from the VALEX lexicon ($p < 0.05$). The difference is 19.50 in F-measure. As both lexicons were extracted from the same corpus data, the improvement can be attributed to improved parser and SCF acquisition performance (Preiss *et al.*, 2007).

F5-F7 refine the basic SCF feature set F4 further. F5 which combines a SCF with CO information proved the best feature set in the supervised verb classification experiment of Li and Brew (2008). In our experiment, F5 produces a significantly lower result than CO and SCF alone (i.e. F1 and F4) using SPEC ($p < 0.05$). However, our corpus is smaller (Li and Brew used the large Gigaword corpus), our SCFs are different, and our approach is unsupervised, making meaningful comparisons difficult.

F6 combines F4 with information about verb tense. This was not helpful: F6 produces worse results than F4. The difference is not significant on T1 ($p = 0.27$), but significant on T2 ($p < 0.05$). F7, on the other hand, yields better results than F4 on both test sets

($p < 0.05$). This demonstrates what the previous research has shown: SCFs perform better when parameterized for prepositions.

Looking at our novel feature sets F9-F17, F9-F11 combine the most accurate SCF feature set F4 with the LP-based features F2-F3. Although the feature space becomes more sparse, all the feature sets outperform F2-F3 on T1 ($p < 0.05$). On T2, F3 performs exceptionally well, and thus yields a better result than F9-F11, but F9-F11 nevertheless perform significantly better than the best SCF-based feature set F4 alone ($p < 0.05$). The differences among F9, F10 and F11 are small on T2, but on T1 F9 yields the best performance. It could be that F9 works the best for the more sparse T1 because it suffers the least from data sparsity (it uses LPs only for the subject relation).

F12-F17 replace the LPs in F9-F11 by semantic SPs. When only 20 clusters are used as SP models and acquired from the smaller sample of (200) argument heads (F12-F14), SPs do not perform better than LPs on T2. A small improvement can be observed on T1, especially with F12 which uses only the subject data (yielding the best F-measure on T1: 57.75%, $p < 0.05$ when compared to the result of the second best non-SP feature F9). However, when 30 more fine-grained clusters are acquired from a bigger sample of (500) argument heads (F15-F17), lower results can be seen on T1. On T2, on the other hand, F15 yields dramatic improvement and we get the best performance for this test set: 80.35% F-measure ($p < 0.05$ when compared to the result of the second best feature F3).

The fact that no improvement is observed when using F16 and F17 on T2 could be explained by the fact that SPs are stronger for the subject position, which also suffers less from the sparse data problem than e.g. object position. The fact that no improvement is observed on T1 is likely to be due to the fact that verbs have strong SPs only at the finer-grained level of Levin classification. Recall that in T1, as many as half of the classes are coarser-grained.

3.4.2 Qualitative evaluation

The best performing feature sets on both T1 and T2 were thus our new SP-based feature sets. We conducted qualitative analysis of the best 30 SP clusters in the T2 data created using SPEC to find out whether these clusters were really semantic in nature, i.e. captured semantically meaningful preferences. As no gold standard specific to our verb classification task was available, we did manual cluster analysis using VerbNet (VN) as an aid. In VN, Levin classes are assigned with semantic descriptions: the arguments of SCFs involved in DAs are labeled with thematic roles, some of which are labeled with selectional restrictions.

From the 30 thematic role types in VN, as many as 20 are associated with the 17 Levin classes in T2. The most frequent role in T2 is agent, followed by theme, location, patient, recipient, and source. From the 36 possible selectional restriction types, 7 appear in

Human	mother, wife, parent, girl, child
Role	patient, student, user, worker, teacher
Body-part	neck, shoulder, back, knee, corner
Authority	committee, police, court, council, board
Organization	society, firm, union, bank, institution
Money	cash, currency, pound, dollar, fund
Amount	proportion, value, size, speed, degree
Time	minute, moment, night, hour, year
Path	street, track, road, stair, route
Building	office, shop, hotel, hospital, house
Region	site, field, area, land, island
Technology	system, model, facility, engine, machine
Task	operation, test, study, analysis, duty
Arrangement	agreement, policy, term, rule, procedure
Matter	aspect, subject, issue, question, case
Problem	difficulty, challenge, loss, pressure, fear
Idea	argument, concept, idea, theory, belief
Power	control, lead, influence, confidence, ability
Form	colour, style, pattern, shape, design
Item	letter, book, goods, flower, card

Table 3.6: Cluster analysis: 20 clusters, their SP labels (assigned by the author of this thesis), and prototypical member nouns

T2; the most frequent ones being +animate and +organization, followed by +concrete, +location, and +communication.

As SP clusters capture selectional *preferences* rather than *restrictions*, we examined manually whether the 30 clusters (i) capture semantically meaningful classes, and whether they (ii) are plausible given the VN semantic descriptions/restrictions for the classes in T2.

The analysis revealed that all the 30 clusters had a predominant, semantically motivated SP supported by the majority of the member nouns. Although many clusters could be further divided into more specific SPs (and despite the fact that some nouns were clearly misclassified), we were able to assign each cluster a descriptive label characterizing the predominant SP. Table 3.6 shows 15 sample clusters, the SP labels assigned to them, and a number of example nouns in these clusters.

When comparing each SP cluster against the VN semantic descriptions/restrictions for T2, we found that each predominant SP was plausible. Also, the SPs frequent in our data were also frequent among the 17 classes according to VN. For example, the many SP clusters labeled as arrangements, issues, ideas and other abstract concepts were also frequent in T2, e.g. among COMMUNICATION (37), CHARACTERISE (29.2), AMALGAMATE (22.2) and other classes.

This analysis showed that the SP models which performed well in verb clustering were semantically meaningful for our task. An independent evaluation using one of the standard datasets available for SP acquisition research (Brockmann and Lapata, 2003) is of course needed to determine how well the acquisition method performs in comparison with other existing methods.

Finally, we evaluated the quality of the verb clusters created using the SP-based features. We found that some of the errors were similar to those seen on T2 when using syntactic features: errors due to polysemy and syntactic idiosyncracy. However, a new error type clearly due to the SP-based feature was detected. A small number of classes got confused because of strong similar SPs in the subject (agent) position. For example, some PEER (30.3) verbs (e.g. *look, peer*) were found in the same cluster with SAY (37.7) verbs (e.g. *shout, yell*) – an error which purely syntactic features do not produce. Such errors were not numerous and could be addressed by developing more balanced SP models across different GRs.

3.5 Discussion and related work

Although features incorporating semantic information about verb SPs make theoretical sense, they have not proved equally promising in previous experiments, which have compared them against syntactic features in verb classification. Joanis *et al.* (2007) incorporated an 'animacy' feature (a kind of a 'SP') which was determined by classifying e.g. pronouns and proper names in data to this single SP class. A small improvement was obtained when this feature was used in conjunction with syntactic features in supervised classification.

Joanis (2002) and Schulte im Walde (2006) experimented with more conventional SPs with syntactic features in English and German verb classification, respectively. They employed top level WordNet (Miller, 1995) and Germanet (Kunze and Lemnitzer, 2002) classes as SP models. Joanis (2002) obtained no improvement over syntactic features, whereas Schulte im Walde (2006) obtained insignificant improvement.

Korhonen *et al.* (2008) combined SPs with SCFs when clustering biomedical verbs. The SPs were acquired automatically from syntactic slots of SCFs (not from GRs as in our experiment) using PC clustering. A small improvement was obtained using LPs extracted from the same syntactic slots, but the SP clusters offered no improvement. Schulte im Walde *et al.* (2008) proposed an interesting SP acquisition method, which involves combining EM training and the MDL principle for a verb classification incorporating SPs. However, no comparison against purely syntactic features is provided.

In our experiment, we obtained a considerable improvement over syntactic features, despite using a fully unsupervised approach to both verb clustering and SP acquisition. In addition to the rich, syntactic-semantic feature sets, our good results can be attributed

		Method	Result
T1	Li et al. 2008	supervised	66.3
	Joanis et al. 2008	supervised	58.4
	Stevenson et al. 2003	semi-supervised unsupervised	29 31
	SPEC	unsupervised	57.55
T2	Sun et al. 2008	supervised unsupervised	62.50 51.6
	Ó Séaghdha et al. 2008	supervised	67.3
	SPEC	unsupervised	80.35

Table 3.7: Previous verb classification results

to the clustering technique capable of dealing with them. The potential of SPEC for the task was recognised earlier by Brew and Schulte im Walde (2002). Although a different version of the algorithm was employed and applied to German (rather than to English), and although no SP features were used, these earlier experiments did demonstrate the ability of the method to perform well in high dimensional feature space.

To get an idea of how our performance compares with that of related approaches, we examined works on verb classification (supervised and unsupervised), which were evaluated on the same test sets using comparable evaluation measures. These works are summarized in table 3.7. ACC and F-measure are shown for T1 and T2, respectively. It is important to note here that although the gold-standards (T1 and T2) employed by these works are the same, the feature extraction methods and the corpora used by each method are different and therefore the results are not directly comparable. However, the results shown serve to show the current best performances on the task.

On T1, the best performing supervised method reported so far is that of Li and Brew (2008). Li and Brew used Bayesian Multinomial Regression for classification. A range of feature sets integrating COs, SCFs and/or LPs were evaluated. The combination of COs and SCFs gave the best result, shown in the table. Joanis *et al.* (2007) report the second best supervised result on T1, using Support Vector Machines for classification and features derived from linguistic analysis: syntactic slots, slot overlaps, tense, voice, aspect, and animacy of NPs. Stevenson and Joanis (2003) report a semi- and unsupervised experiment on T1. A feature set similar to that of Joanis *et al.* (2007) was employed (features were selected in a semi-supervised fashion) and hierarchical clustering was used.

Our unsupervised method SPEC performs substantially better than the unsupervised method of Stevenson et al. and nearly as well as the supervised approach of Joanis *et al.* (2007) (note, however, that the different experiments involved different sub-sets of T1 so are not entirely comparable).

On T2, the best performing supervised method so far is that of Ó Séaghdha and Copestake (2008) which employs a distributional kernel method to classify SCF features parameterized for prepositions in the automatically acquired VALEX lexicon. Using exactly the same data and feature set, Sun *et al.* (2008b) obtain a slightly lower result when using a supervised method (Gaussian) and a notably lower result when using an unsupervised method (PC clustering). Our method performs considerably better and also outperforms the supervised method of Ó Séaghdha and Copestake (2008).

3.6 Summary

We introduced a new approach to verb clustering which involves the use of (i) rich lexical, SCF and GR data produced by a SCF system, (ii) novel syntactic-semantic feature sets which combine a variety of linguistic information, and (iii) a new variation of SPEC which is particularly suited for dealing with the resulting, high dimensional feature space. Using this approach, we showed on two well-established test sets that automatically acquired SPs can be highly useful for verb clustering. This result contrasts with most previous works but is in line with theoretical work on verb classification which relies not only on syntactic but also on semantic features (Levin, 1993).

In addition to the ideas mentioned earlier, future work could look into optimal ways of acquiring SPs for verb classification. Considerable research has been done on SP acquisition, most of which has involved collecting argument headwords from data and generalizing to WordNet classes. Brockmann and Lapata (2003) have showed that WordNet-based approaches do not always outperform simple frequency-based models, and a number of techniques have been proposed which may offer ideas for refining our current unsupervised approach (Erk, 2007; Bergsma *et al.*, 2008). The number and type (and combination) of GRs for which SPs can be reliably acquired, especially when the data is sparse, requires also further investigation.

In addition, it would be interesting to investigate other potentially useful features for verb classification (e.g. named entities and preposition classes) and explore semi-automatic ML technology and active learning for guiding the classification.

Chapter 4

Hierarchical verb clustering using graph factorization

4.1 Introduction

[1]Most works on verb classification have focussed on acquiring and evaluating flat classifications (Schulte im Walde, 2006; Joanis *et al.*, 2007; Sun *et al.*, 2008b; Li and Brew, 2008; Korhonen *et al.*, 2008; Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009b). Levin's classification is not flat, but taxonomic in nature, which is practical for NLP purposes since applications may differ in terms of the granularity they require from a classification.

In this chapter, we experiment with hierarchical Levin-style clustering. We adopt as our baseline method a well-known hierarchical method – agglomerative clustering (AGG) – which has been previously used to acquire flat Levin-style classifications (Stevenson and Joanis, 2003) as well as hierarchical verb classifications not based on Levin (Ferrer, 2004; Schulte im Walde, 2008). The method has also been used in the related task of noun clustering (Ushioda, 1996; Matsuo *et al.*, 2006; Bassiou and Kotropoulos, 2011).

We introduce then a new method called Hierarchical Graph Factorization Clustering (HGFC) (Yu *et al.*, 2006). This graph-based, probabilistic clustering algorithm has some clear advantages over AGG (e.g. it delays the decision on a verb's cluster membership at any level until a full graph is available, minimising the problem of error propagation) and it has been shown to perform better than several other hierarchical clustering methods in recent comparisons (Yu *et al.*, 2006). The method has been applied to the identification of social network communities (Lin *et al.*, 2008), but has not been used (to the best of our knowledge) in NLP before.

We modify HGFC with a new tree extraction algorithm which ensures a more consistent result, and we propose two novel extensions to it. The first is a method for automatically

[1] The research reported in this chapter was published in Sun and Korhonen (2011).

determining the tree structure (i.e. number of clusters to be produced for each level of the hierarchy). This avoids the need to pre-determine the number of clusters manually. The second is addition of soft constraints to guide the clustering performance (Vlachos *et al.*, 2009b). This is useful for situations where a partial (e.g. a flat) verb classification is available and the goal is to extend it.

Adopting a set of lexical and syntactic features which have performed well in previous works, we compare the performance of the two methods on test sets extracted from Levin and VerbNet. When evaluated on a flat clustering task, HGFC outperforms AGG and performs very similarly with the best flat clustering method reported on the same test set in section 3. When evaluated on a hierarchical task, HGFC performs considerably better than AGG at all levels of gold standard classification. The constrained version of HGFC performs the best, as expected, demonstrating the usefulness of soft constraints for extending partial classifications.

Our qualitative analysis shows that HGFC is capable of detecting novel information not included in our gold standards. The unconstrained version can be used to acquire novel classifications from scratch while the constrained version can be used to extend existing ones with additional class members, classes and levels of the hierarchy.

4.2 Target classification and test sets

The taxonomy of Levin (1993) classifies over 3000 verbs in 57 top level classes, some of which divide further into subclasses. The extended version of the taxonomy in VerbNet (Kipper-Schuler, 2005) classifies 5757 verbs. Its 5-level taxonomy includes 101 top level and 369 subclasses. We used three gold standards (and corresponding test sets) extracted from these resources in our experiments:

T3: The first gold standard is a flat gold standard which includes 13 classes appearing in Levin's original taxonomy (Stevenson and Joanis, 2003). We included this small gold standard in our experiments so that we could compare the flat version of our method against previously published methods. Stevenson and Joanis (2003) did not include the actual list of verbs used in the experiment. Therefore, we selected 20 verbs from each class which occur at least 100 times in our corpus. This is also the approach used by Stevenson and Joanis. This gave us 260 verbs in total. The actual verbs are listed in appendix D.

T4: The second gold standard is a large, hierarchical gold standard which we extracted from VerbNet as follows: 1) We removed all the verbs that have less than 1000 occurrences in our corpus. 2) In order to minimise the problem of polysemy, we assigned each verb to the class which, according to VerbNet, corresponds to its predominant sense in WordNet (Miller, 1995). 3) In order to minimise the sparse data problem with very fine-grained classes, we converted the resulting classification into a 3-level representation so that the classes at the 4th and 5th level were combined. For example, the sub-classes of *Declare* verbs (numbered as 29.4.1.1.{1,2,3}) were combined into 29.4.1. 4) The classes that have fewer than 5 members were discarded. The total number of verb senses in the resulting gold standard is 1750, which is 33.2% of the verbs in VerbNet. T4 has 51 top level, 117 second level, and 133 third level classes.

T5: The third gold standard is a subset of T4 where singular classes (top level classes which do not divide into subclasses) are removed. This gold standard was constructed to enable proper evaluation of the constrained version of HGFC (introduced in the following section) where we want to compare the impact of constraints across several levels of classification. T5 provides classification of 357 verbs into 11 top level, 14 second level, and 32 third level classes.

For each verb appearing in T3-T5, we extracted all the occurrences (up to 10,000) from the raw corpus data used for constructing VALEX (Korhonen *et al.*, 2006a), including the BNC (Leech, 1992), the Guardian corpus, the Reuters corpus (Rose *et al.*, 2002), the North American News Text Corpus (Graff, 1995) and the data used for two Text Retrieval Evaluation Conferences (TREC-4 and TREC-5).

4.3 Method

4.3.1 Features and feature extraction

We used in the experiments the features that proved the best in the earlier experiment reported in chapter 3: F-SCF, F-SCF+PP(B) and F-SCF+LP(A). The description of these features and the feature extraction methods are given in table 2.4 (all the features used in this thesis are summarized in appendix A). However, we only used the best syntactic features here. Although the semantic feature – verb SPs – was the best feature (when used in combination with syntactic features) in chapter 3, we left it for future work because we noticed that different levels of classification are likely to require semantic features at different granularities.

4.3.2 Clustering

We introduce the agglomerative clustering (AGG) and Hierarchical Graph Factorization Clustering (HGFC) methods in the following two subsections, respectively. The subsequent two subsections present our extensions to HGFC: (i) automatically determining the cluster structure and (ii) adding soft constraints to guide clustering performance.

Agglomerative clustering

AGG is a method which treats each verb as a singleton cluster and then successively merges the closest two clusters until all the clusters have been merged into one. We used the SciPy's implementation (Oliphant, 2007) of the algorithm. The cluster distance is measured using linkage criteria. We experimented with four commonly used linkage criteria: Single, Average, Complete and Ward's (Ward Jr., 1963). Ward's criterion performed the best and was used in all the experiments in this chapter. It measures the increase in variance after two clusters are merged. The output of AGG tends to have an excessive number of levels. Cut-based methods (Wu and Leahy, 1993; Shi and Malik, 2000) are frequently applied to extract a simplified view. We followed previous verb clustering works and cut the AGG hierarchy manually.

AGG suffers from two problems. The first is error propagation. When a verb is misclassified at a lower level, the error propagates to all the upper levels. The second is local pairwise merging, i.e. the fact that only two clusters can be combined at any level. For example, in order to group clusters representing Levin classes 9.1, 9.2 and 9.3 into a single cluster representing class 9, the method has to produce intermediate clusters, e.g. 9.{1,2} and 9.3. Such clusters do not always have a semantic interpretation. Although they can be removed using a cut-based method, this requires a pre-defined cut-off value which is difficult to set (Stevenson and Joanis, 2003). In addition, a significant amount of information is lost in pair-wise clustering. In the above example, only the clusters 9.{1,2} and 9.3 are considered, while alternative clusters 9.{1,3} and 9.2 are ignored. Ideally, information about all the possible intermediate clusters should be aggregated, but this is intractable in practice.

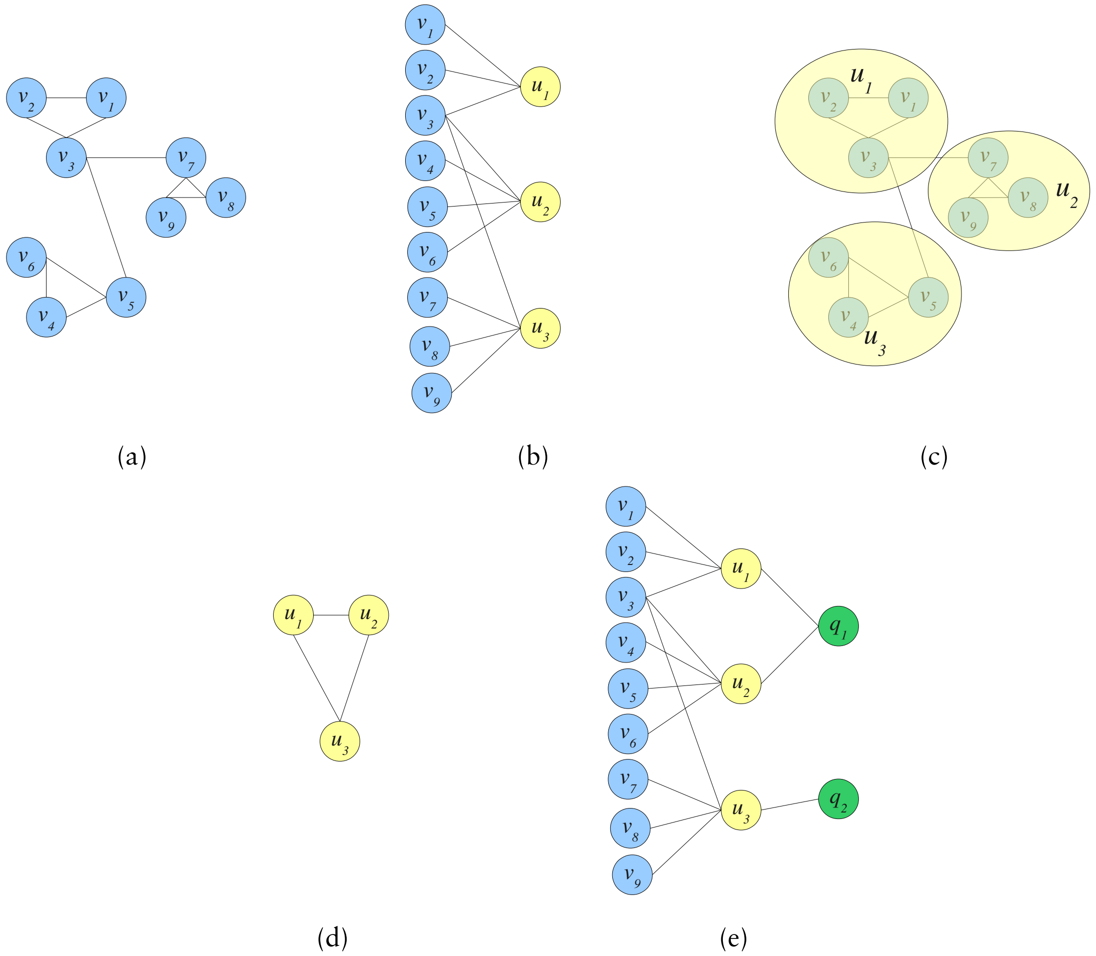
Hierarchical Graph Factorization Clustering

Our new method HGFC derives a probabilistic bipartite graph from the similarity matrix (Yu *et al.*, 2006). The local and global clustering structures are learned via the random walk properties of the graph.

The method does not suffer from the above problems with AGG. Firstly, there is no error propagation because the decision on a verb's membership at any level is delayed until the full bipartite graph is available and until a tree structure can be extracted from it by aggregating probabilistic information from all the levels. Secondly, the bipartite graph enables the construction of a hierarchical structure without any intermediate classes. For example, we can group classes 9.{1,2,3} directly into class 9.

We calculated the similarity matrix using JSD instead of skew divergence. When compared to JSD, skew divergence has an extra parameter (the smoothing factor). The value of this parameter is difficult to set automatically, as there is no labelled training data available for parameter estimation. The details of JSD and other similarity measures are summarized in appendix B. Given a set of verbs, $V = \{v_n\}_{n=1}^{N}$, we compute a similarity matrix W where $W_{ij} = \exp(-d_{jsd}(v_1, v_2))$. W can be encoded by a undirected graph G (Figure 4.1(a)), where the verbs are mapped to vertices and the W_{ij} is the edge weight between vertices i and j.

The graph G and the cluster structure can be represented by a bipartite graph $K(V, U)$. V are the vertices on G. $U = \{u_p\}_{p=1}^{m}$ represent the hidden m clusters. For example, looking at Figure 4.1(b), V on G can be grouped into three clusters u_1, u_2 and u_3. The matrix B denotes the $n \times m$ adjacency matrix, with b_{ip} being the connection weight between the vertex v_i and the cluster u_p. Thus, B represents the connections between clusters at an upper and lower level of clustering. A flat clustering algorithm can be induced by computing B.



(a)

(b)

(c)

(d)

(e)

Figure 4.1: (a) An undirected graph G representing the similarity matrix; (b) The bipartite graph showing three clusters on G; (c) The induced clusters U; (d) The new graph G_1 over clusters U; (e) The new bipartite graph over G_1

The bipartite graph K also induces a similarity (W') between v_i and v_j: $w'_{ij} = \sum_{p=1}^{m} \frac{b_{ip}b_{jp}}{\lambda_p} = (B\Lambda^{-1}B^T)_{ij}$ where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_m)$. Therefore, B can be found by approximating the similarity matrix W of G using W' derived from K. Given a distance function ζ

between two similarity matrices, B approximates W by minimizing the cost function $\zeta(W, B\Lambda^{-1}B^T)$. The coupling between B and Λ is removed by setting $H = B\Lambda^{-1}$:

$$\min_{H,\Lambda} \zeta(W, H\Lambda H^T), s.t. \sum_{i=1}^{n} h_{ip} = 1 \tag{4.1}$$

We use the divergence distance: $\zeta(X, Y) = \sum_{ij}(x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij})$. Yu *et al.* (2006) showed that this cost function is non-increasing under the update rule:

$$\tilde{h}_{ip} \propto h_{ip} \sum_j \frac{w_{ij}}{(H\Lambda H^T)_{ij}} \lambda_p h_{jp} \text{ s.t.} \sum_i \tilde{h}_{ip} = 1 \tag{4.2}$$

$$\tilde{\lambda}_p \propto \lambda_p \sum_j \frac{w_{ij}}{(H\Lambda H^T)_{ij}} h_{ip} h_{jp} \text{ s.t.} \sum_p \tilde{\lambda}_p = \sum_{ij} w_{ij} \tag{4.3}$$

w_{ij} can be interpreted as the probability of the direct transition between v_i and v_j: $w_{ij} = p(v_i, v_j)$, when $\sum_{ij} w_{ij} = 1$. b_{ip} can be interpreted as:

$$\begin{aligned} p(u_p, u_q) &= p(u_p)p(u_p|u_q) = \sum_{i=1}^{n} \frac{b_{ip}b_{iq}}{d_i} \\ &= (B^T D^{-1} B)_{pq} \\ D &= \text{diag}(d_1, ..., d_n) \text{ where } d_i = \sum_{p=0}^{m} b_{ip} \end{aligned} \tag{4.4}$$

$p(u_p, u_q)$ is the similarity between the clusters. It takes into account a weighted average of contributions from all the data. This is different from the linkage method, where only the data from two clusters are considered.

Given the cluster similarity $p(u_p, u_q)$, we can construct a new graph G_1 (Figure 4.1(d)) with the clusters U as vertices. The cluster algorithm can be applied again (Figure 4.1(e)). This process can go on iteratively, leading to a hierarchical graph. [2]

Additional steps need to be performed in order to extract a tree from the hierarchical graph. Yu *et al.* (2006) performs the extraction via a propagation of probabilities from the bottom level clusters. For a verb v_i, the probability of assigning it to cluster $v_p^{(l)}$ at level l is given by:

$$\begin{aligned} p(v_p^{(l)}|v_i) &= \sum_{V_{l-1}} ... \sum_{V_1} p(v_p^{(l)}|v^{(l-1)})...p(v^{(1)}|v_i) \\ &= (D_1^{(-1)} B_1 D_2^{-1} B_2 D_3^{-1} B_3 ... D_l^{-1} B_l)_{ip} \end{aligned} \tag{4.5}$$

[2] The hierarchical graph is constructed only for inducing the tree of clusters. It is not motivated by any linguistic phenomenon (e.g. verb polysemy). Although it is a soft clustering approach, we argue that soft clustering is not a proper model for polysemy (more details in section 8.2.2).

Algorithm 1 HGFC algorithm (Yu *et al.*, 2006)

```
Require: $N$ verbs $V$, number of clusters $m_l$ for $L$ levels
  Compute the similarity matrix $W_0$ from $V$
  Build the graph $G_0$ from $W_0$ , and $m_0 \leftarrow n$
  for $l = 1, 2$ to $L$ do
    Factorize $G_{l-1}$ to obtain bipartite graph $K_l$ with the adjacency matrix $B_l$ (eq. 4.1, 4.2 and 4.3)
    Build a graph $G_l$ with similarity matrix $W_l = B_l^T D_l^{-1} B_l$ according to equation 4.4
  end for
  return $B_L, B_{L-1}...B_1$
```

This method may not extract a consistent tree structure because the cluster membership at lower levels does not constrain the upper level membership. This prevented us from extracting a Levin style hierarchical classification in our initial experiments. For example, where two verbs were grouped together at a lower level, they could belong to separate clusters at an upper level. We therefore propose a new tree extraction algorithm (Algorithm 2).

The new algorithm starts from the top level bipartite graph, and generates consistent labels for each level by taking into account the tree constraints set at upper levels.

Algorithm 2 Tree extraction algorithm for HGFC

```
Require: Given $N$, $(B^l, m_l)$ on each level for $L$ levels
  On the top level $L$, collect the labels $T^L$ (eq. 4.5)
  Define $C$ to be a $(m_{L-1} \times m_L)$ zero matrix, $C_{ij} \leftarrow 1$, where $i, j = \arg\max_{i,j}\{B^L_{ij}\}$
  for $l = L - 1$ to 1 do
    for $i = 1$ to $N$ do
      Compute $p(v^l_p|v_i)$ for each cluster $p$ (eq. 4.5)
      $t^l_i = \text{argmax}_p\{p(v^l_p|v_i)|p = 1...m_l, C_{pt^{l+1}_i} \neq 0\}$
    end for
    Redefine $C$ to be a $(m_{l-1} \times m_l)$ zero matrix, $C_{ij} \leftarrow 1$, where $i, j = \arg\max_{i,j}\{B^l_{ij}\}$
  end for
  return Tree consistent labels $T^L, T^{L-1}...T^1$
```

Automatically determining the number of clusters for HGFC

HGFC needs the number of levels and clusters at each level as input. However, this information is not always available (e.g. when the goal is to actually learn this information automatically). We therefore propose a method for inferring the cluster structure from data. As shown in figure 1, a similarity matrix W models one-hop transitions that follow the links from vertices to neighbors. A walker can also go to other vertices via multi-hop transitions. According to the chain rule of the Markov process, the multi-hop transitions indicate a decaying similarity function on the graph (Yu *et al.*, 2006). After t transitions, the similarity matrix (W_t) becomes:

$$W_t = W_{t-1}D_0^{-1}W_0$$

Yu *et al.* (2006) proved the correspondence between the HGFC levels (l) and the random walk time: $t = 2^{l-1}$. So the vertices at level l induce a similarity matrix of verbs after t-hop transitions. The decaying similarity function captures the different scales of clustering structure in the data (Azran and Ghahramani, 2006b). The upper levels would have a smaller number of clusters which represent a more global structure. After several levels, all the verbs are expected to be grouped into one cluster. The number of levels and clusters at each level can thus be learned automatically.

We therefore propose a method that uses the decaying similarity function to learn the hierarchical clustering structure. One simple modification to algorithm 1 is to set the number of clusters at level l (m_l) to be $m_{l-1} - 1$. m is denoted as the number of clusters that have at least one member according to eq. 4.5. We start by treating each verb as a cluster at the bottom level. The algorithm stops when all the data points are merged into one cluster. The increasingly decaying similarity causes many clusters to have 0 members especially at lower levels, which are pruned in the tree extraction.

Adding constraints to HGFC

The basic version of HGFC makes no prior assumptions about the classification. It is useful for learning novel verb classifications from scratch. However, when wishing to extend an existing classification (e.g. VerbNet) it may be desirable to guide the clustering performance on the basis of information that is already known. We propose a constrained version of HGFC which makes uses of labels at the bottom level to learn upper level classifications. We do this by adding soft constraints to clustering, following Vlachos *et al.* (2009b).

We modify the similarity matrix W as follows: If two verbs have different labels ($l_i \neq l_j$), the similarity between them is decreased by a factor a, and $a < 1$. We set a to 0.5 in the experiments. The resulting tree is generally consistent with the original classification. The influence of the underlying data (domain or features) is reduced according to a.

As discussed in section 2.3.1, adding the constraints is a form of semi-supervised learning that enables us to make use of the kind of vague prior knowledge that we have available, e.g. knowledge about pairs of verbs that cannot be in the same class, even when we have no idea how they should be grouped.

4.4 Experimental evaluation

We applied the clustering methods introduced in section 4.3 to the test sets described in section 4.2 and evaluated them both quantitatively and qualitatively, as described in the subsequent sections.

4.4.1 Evaluation methods

We used ACC and R$_{adj}$ to evaluate the results on the flat test set T3 (see section 4.2 for details of T3-T5). Since NMI can compare clusterings with different numbers of clusters, and since we also want to compare to the F in previous experiments, we used NMI and F to evaluate hierarchical clustering results on T4 and T5. Finally, we supplemented quantitative evaluation with qualitative evaluation of clusters produced by different methods.

We used the McNemar's test (McNemar, 1947) as described in 3.3.4 to verify the statistical significance of the major findings.

4.4.2 Quantitative evaluation

We first evaluated AGG and the basic (unconstrained) HGFC on the small flat test set T3. The main purpose of this evaluation was to compare the results of our methods against previously published results on the same test set. The number of clusters (K) and levels (L) were inferred automatically for HGFC as described in section 4.3.2. However, to make the results comparable with previously published ones, we cut the resulting hierarchy at the level of closest match (12 clusters) to the K (13) in the gold standard. For AGG, we cut the hierarchy at 13 clusters.

Method	ACC	R$_{adj}$
HGFC	41.2 (+8.5)	17.4 (+7.5)
AGG (reproduced)	32.7	9.9
AGG (Stevenson and Joanis (2003)	31.0	9.0

Table 4.1: Comparison against Stevenson and Joanis (2003)'s result on T3 (using similar features).

Table 4.1 shows our results and the results of Stevenson and Joanis (2003) on T3 when employing AGG using Ward as the linkage criterion. In this experiment, we used the same feature set as Stevenson and Joanis (2003) (F-SCF+PP(B)) and were therefore able to reproduce their AGG result with a difference smaller than 2%. When using this simple feature set, HGFC outperforms the best performing AGG significantly: 8.5% in ACC and 7.3% in R$_{adj}$ ($p < 0.05$).

N_c	N_l	HGFC unconstrained		AGG	
		NMI	F	NMI	F
130	133	57.31	36.65	54.22	32.62
114	117	54.67	37.96	51.35	32.44
50	51	37.75	40.00	32.61	32.78

Table 4.2: Performance on T4 using a pre-defined tree structure.

N_c	N_l	HGFC unconstrained		HGFC constrained		AGG	
		NMI	F	NMI	F	NMI	F
31	32	51.65	42.01	91.47	92.07	49.70	40.30
15	14	42.75	47.70	82.16	82.80	39.19	43.69
11	11	38.91	51.17	71.69	75.00	34.88	44.80

Table 4.3: Performance on T5 using a pre-defined tree structure.

We also compared HGFC against the best reported clustering method on T3 to date – that of SPEC in section 3. We used the feature sets F-SCF+LP(A). HGFC obtains F of 49.93% on T3, which is 5% lower than the result in section 3 ($p < 0.05$). The difference comes from the tree consistency requirement. When HGFC is forced to produce a flat clustering (a one level tree only), it achieves F of 52.55%. This is very close to the performance of SPEC.

We then evaluated our methods on the hierarchical test sets T4 and T5. We used the best-performing feature sets F-SCF+LP(A) for these tasks. In the first set of experiments, we pre-defined the tree structure for HGFC by setting L to 3 and K at each level to be the K in the hierarchical gold standard. The hierarchy produced by AGG was cut into 3 levels according to the Ks in the gold standard. This enabled direct evaluation of the results against the 3-level gold standards using both NMI and F.

The results are reported in tables 4.2 and 4.3. In these tables, N_c is the number of clusters in HGFC clustering while N_l is the number of classes in the gold standard (the two do not always correspond perfectly because a few clusters have zero members).

Table 4.2 compares the results of the unconstrained version of HGFC against those of AGG on our largest test set T4. As with T3, HGFC outperforms AGG significantly ($p < 0.05$). The benefit can now be seen at three different levels of the hierarchy. On average, HGFC outperforms AGG 3.5% in NMI and 4.8% in F. The difference between the methods becomes clearer when moving towards the upper levels of the hierarchy.

Table 4.3 shows the results of both unconstrained and constrained versions of HGFC and those of AGG on the test set T5 (where singular classes are removed to enable proper evaluation of the constrained method). The results are generally better on this test set than on T4, which is to be expected since T5 is a refined subset of T4[3].

[3] NMI is higher on T4, however, because NMI has a higher baseline for a larger number of clusters (Vinh

T4			T5		
N_c	N_l	HGFC	N_c	N_l	HGFC
148	133	53.26	64	32	54.91
97	117	49.85	35	32	50.83
46	51	33.55	20	14	44.02
19	51	25.80	10	14	34.41
9	51	19.17	6	11	32.27
3	51	13.06			

Table 4.4: NMI of unconstrained HGFC when trees for T4 and T5 are inferred automatically.

Recall that the constrained version of HGFC learns the upper levels of classification on the basis of soft constraints set at the bottom level, as described earlier in section 4.3.2. As a consequence, NMI and F are both greater than 90% at the bottom level and the results at the top level are notably lower because the impact of the constraints degrades the further away one moves from the bottom level. Yet, the relatively good result across all levels ($p < 0.05$) shows that the constrained version of HGFC can be a useful method to extend the hierarchical structure of known classifications.

Finally, Table 4.4 shows the results for the unconstrained HGFC on T4 and T5. Here, the tree structure is not pre-defined, but inferred fully automatically as described in section 4.3.2. 6 levels are learned for T4 and 5 for T5. The number of clusters produced ranges from 3 to 148 for T4 and from 6 to 64 for T5. We can see that the automatically detected cluster numbers distribute evenly across different levels. The scale of the clustering structure is more complete here than in the gold standards.

In the table, N_c indicates the number of clusters in the inferred tree, while N_l indicates the closest match to the number of classes in the gold standard. This evaluation is not fully reliable because the match between the gold standard and the clustering is poor at some levels of the hierarchy. However, it is encouraging to see that the results do not drop dramatically until the match between the two is really poor.

4.4.3 Qualitative evaluation

To gain better insight into the performance of HGFC, we conducted further qualitative analysis of the clusters produced for T5 by the two versions of this method. We focussed on the top level of 11 clusters (in the evaluation against the hierarchical gold standard, see table 4.3), as the impact of soft constraints is the weakest for the constrained method at this level.

et al., 2009). NMI is not ideal for comparing the results of T4 and T5.

As expected, the constrained HGFC kept many individual verbs belonging to the same VerbNet subclass together (e.g. verbs *enjoy, hate, disdain, regret, love, despise, detest, dislike, fear* for the class 31.2.1) so that most clusters simply group lower level classes and their members together. Three nearly clean clusters were produced which only include sub-classes of the same class (e.g. 31.2.0 and 31.2.1 which both belong to 31.2 *Admire* verbs). However, the remaining 8 clusters group together sub-classes (and their members) belonging to unrelated parent classes. Interestingly, 6 of these make both syntactic and semantic sense. For example, several 37.7 *Say* verbs and 29.5 *Conjencture* verbs are found together. These verbs share the meaning of communication and take similar sentential complements.

In contrast, none of the clusters produced by the unconstrained HGFC represents a single VerbNet class. The majority represent a high number of classes and fewer members per class. Yet many of the clusters make syntactic and semantic sense. A good example is a cluster which includes member verbs from 9.7 *Spray/Load* verbs, 21.2 *Carve* verbs, 51.3.1 *Roll* verbs, and 10.4 *Wipe* verbs. The verbs included in this cluster share the meaning of a specific type of motion and show similar syntactic behaviour.

Thorough Levin-style investigation particularly of the unconstrained method, would require looking at shared diathesis alternations between cluster members. We left this for future work. However, the analysis we conducted confirmed that the constrained method could indeed be used for extending known classifications, while the unconstrained method is required for acquiring novel classifications from scratch. The errors in clusters produced by both methods were mostly due to syntactic idiosyncracy and the lack of semantic information in this clustering.

4.5 Discussion and conclusion

We have introduced a new graph-based method HGFC for hierarchical verb clustering which avoids some of the problems (e.g. error propagation, pairwise cluster merging) reported with the frequently used AGG method. We modified HGFC so that it can be used to automatically determine the tree structure for clustering, and proposed two extensions to it which make it even more suitable for our task. The first involves automatically determining the number of clusters to be produced, which is useful when this is not known in advance. The second involves adding soft constraints to guide the clustering performance, which is useful when aiming to extend an existing classification.

The results reported are promising. On a flat test set (T3), the unconstrained version of HGFC outperforms AGG and performs very similarly to the best current flat clustering method (SPEC) evaluated on the same dataset. On the hierarchical test sets (T4 and T5), the unconstrained and constrained versions of HGFC outperform AGG clearly at all levels of classification. The constrained version of HGFC detects the missing hierarchy from the existing gold standards with high accuracy. When the number of clusters and levels is learned automatically, the unconstrained method produces a multi-level hierarchy. Our evaluation against a 3-level gold standard shows that such a hierarchy is fairly accurate. Finally, the results from our qualitative evaluation show that both constrained and unconstrained versions of HGFC are capable of learning valuable novel information not included in the gold standards.

Previous work on Levin style verb classification has mostly focussed on flat classifications using methods suitable for flat clustering (Schulte im Walde, 2006; Joanis *et al.*, 2007; Sun *et al.*, 2008b; Li and Brew, 2008; Korhonen *et al.*, 2008; Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009b). However, some works have employed hierarchical clustering as a method to infer flat clustering.

For example, Schulte im Walde and Brew (2001) employed AGG to initialize the K-Means clustering for German verbs. This gave better results than random initialization. Stevenson and Joanis (2003) used AGG for flat clustering on T3. They cut the hierarchy at the number of classes in the gold standard and found that it is difficult to automatically determine a good cut-off. Our evaluation in the previous section shows that HGFC outperforms their implementation of AGG.

AGG was also used by Ferrer (2004) who performed hierarchical clustering of 514 Spanish verbs. The results were evaluated against a hierarchical gold standard resembling that of Levin's classification in English (Vázquez *et al.*, 2000). R_{adj} of 0.07 was reported for a 15-way classification, which is comparable to the result of Stevenson and Joanis (2003).

Hierarchical clustering has also been performed for the related task of semantic verb classification. For example, Basili *et al.* (1993) identified the problems of AGG, and applied a conceptual clustering algorithm (Fisher, 1987) to Italian verbs. They used semi-automatically acquired semantic roles and the concept types as features. No quantitative results were reported. The qualitative evaluation shows that the resulting clusters are very fine-grained.

Schulte im Walde (2008) performed hierarchical clustering of German verbs using human verb association as features and AGG as a method. They focussed on two small collections of 56 and 104 verbs and evaluated the result against flat gold standard extracted from GermaNet (Kunze and Lemnitzer, 2002) and German FrameNet (Erk *et al.*, 2003), respectively. They reported F of 62.69% for the 56 verbs, and F of 34.68% for the 104 verbs.

In the future, this research line could be extended in several directions. One would be to try to determine optimal features for different levels of clustering. For example, the general syntactic features (e.g. SCF) may perform the best at top levels of a hierarchy while more specific or refined features (e.g. SCF+pp) may be optimal at lower levels. Another idea is to investigate incorporating semantic features, like verb SPs, in our feature set. It is likely that different levels of clustering require more or less specific SPs. One way to obtain the latter is hierarchical clustering of relevant noun data.

In addition, unconstrained HGFC could be applied to specific domains to investigate its capability to learn novel, previously unknown classifications. As for the constrained version of HGFC, a larger scale experiment on the VerbNet data could be conducted to investigate what kind of upper level hierarchy it can propose for this resource (which currently has over 270 top level classes).

Finally, HGFC could be compared to other hierarchical clustering methods that are relatively new to NLP but have proved promising in other fields, including Bayesian Hierarchical Clustering (Heller and Ghahramani, 2005; Teh *et al.*, 2008) and the method of Azran and Ghahramani (2006a) based on SPEC.

Chapter 5

Verb classification in the biomedical domain

5.1 Introduction

In recent years, the application of NLP techniques to biomedicine has become increasingly popular due to the urgency to develop techniques which can automatically process information in the growing volume of literature in this field. Remarkable progress has been made in many areas of BIO-NLP, including information retrieval, information extraction and basic text processing (e.g. POS-tagging and parsing). The current challenge is to improve these techniques with richer and deeper analysis. Large scale lexical resources which specify the syntax and semantics of words are needed for this (Ananiadou and McNaught, 2005). General lexical resources such as WordNet, VerbNet and Comlex only provide limited coverage of biomedical words, and manually developed domain-specific lexical resources (e.g. the UMLS specialist lexicon (Browne *et al.*, 2003)) are insufficient in particular for verbs, and costly to extend.

Automatically acquiring or updating lexical information from corpora is a better approach for a rapidly developing domain such as biomedicine. The automatic approach is cost effective, it can handle sub-domain variation easily (i.e. the fact that the sub-domains of biomedicine differ in their lexical characteristics (Lippincott *et al.*, 2011)). It can also gather statistical information, which is highly useful for biomedical NLP applications, but difficult to collect via manual means.

A few large lexical resources exist for biomedicine such as the UMLS metathesaurus (Nelson *et al.*, 2002) (e.g. MeSH (Lipscomb, 2000)). Such resources mainly focus on biomedical concepts which are nouns. A few lexical resources cover verbs in biomedical texts. The UMLS SPECIALIST lexicon (Browne *et al.*, 2000) includes both scientific and biomedical vocabulary, with a particular emphasis on medical and health-related vocabulary. It was created manually by lexicographers. It contains a small number syntactic complementation patterns for verbs, but there is no statistical information on the usage of the patterns. The BioLexicon (Sasaki *et al.*, 2008) is a corpus-driven lexical resource which contains syntactic and semantic frame information for verbs. The grammatical frames of verbs are acquired from the output of the Enju (Miyao *et al.*, 2008) deep syntactic parser in the E. Coli subdomain, so only a limited number of verbs and frames are represented. There are also smaller corpus-driven lexical resources: BioFrameNet (Dolbey *et al.*, 2006) and PASBio (Wattarujeekrit *et al.*, 2004), but e.g. PASBio covers 30 verbs only.

The work presented in this chapter is most related to the work of Korhonen et. al 2006 and 2008, because it deals with the same task: Levin-style biomedical verb classification. Biomedical texts have domain-specific verb classes. Some of them capture senses that only or mainly appear in biomedical texts. For example, verbs in the ACTIVATE class (e.g. activate / up-regulate / induce / stimulate) (Korhonen *et al.*, 2008) take similar SCFs and SPs, and share similar underlying predicate argument structure (e.g. PROTEINS: TP53 ACTIVATE GENES: CIP1). Others capture general or general-scientific senses (e.g. DEMONSTRATE verbs).

Korhonen *et al.* (2006b, 2008) created a three level gold standard containing 192 verbs and 50 fine-grained classes for biomedicine and using biomedical journals as corpus data, performed a preliminary verb clustering experiment using pairwise clustering (Puzicha *et al.*, 2000) as a method. Korhonen *et al.* (2008) took this work further and systematically compared a range of syntactic and semantic features for the task. These works showed that despite the challenging nature of biomedical texts (e.g. frequent use of passive, anaphora and long, embedded sentences) and the fact that the unsupervised lexical acquisition did not rely on the best performing, lexicalised parsers in biomedicine, it was still possible to obtain good results because biomedical texts tend to be quite uniform in terms of word senses. The sense uniformity results in cleaner features, e.g. clearer selectional preferences, which can aid clustering.

Using the same data, feature types and gold standard as those employed by Korhonen *et al.* (2006b, 2008), we investigate how our clustering methods perform on the biomedical data in comparison with those used in previous works. In the previous work of Korhonen *et al.* (2008), PC was the best performing clustering method. In our previous experiment on general English data (in chapter 3), SPEC outperformed PC. In this experiment, we investigate the performance of SPEC and HGFC on the biomedical data. Korhonen *et al.* (2008) produced a three-level hierarchical result for biomedical data. However, the resulting hierarchy is not a tree but a graph. In order words, verbs in the same cluster can be divided into different clusters at upper levels. This does not match the tree-consistent standard of Levin's style verb classification, and it also makes the resulting hierarchy difficult for humans to interpret. We showed in chapter 4 how our HGFC method produces a tree-consistent hierarchical verb classification for general language; here we investigate how it performs in the biomedical domain.

5.2 Classification methods

We used the SPEC and HGFC clustering methods described earlier in section 3.2 and 4.3.2 respectively. The JSD similarity measure is employed for calculating the similarity matrix. The details of JSD with comparison to other similarity measures are listed in appendix B.

5.3 Data

5.3.1 Test Verbs and Gold Standard

We employed in our experiments the same gold standard as earlier employed by Korhonen *et al.* (2006b, 2008). The gold standard includes 192 test verbs (typically frequent verbs in biomedical journal articles) classified into 16, 34 and 50 classes, respectively. This three level gold standard was created by 4 domain experts and 2 linguists who were asked to examine whether the test verbs are similar in terms of their syntactical properties (i.e. verbs with similar SCF distribution) and also similar in term of semantics (i.e. they share a common meaning). If a group of verbs match the criteria, a verb class was identified and named. The classes created by domain experts were labeled as BIO and those created by linguists as GEN. BIO classes include 116 verbs whose analysis required domain knowledge (e.g. *activate, solubilize, harvest*). GEN classes include 76 general or scientific text verbs (e.g. *demonstrate, hypothesize and appear*). Each class is associated with 1-30 member verbs.

The linguists used Levin (1993) classes as gold standard classes whenever possible and created new ones when needed. The domain experts used two semantic classifications of biomedical verbs (Friedman *et al.*, 2002; Spasic *et al.*, 2005) as a starting point. Only those classes/memberships which all experts agreed on were included. Table 5.1 shows all the gold standard classes with member verbs.

5.3.2 Test Data

We used the same corpus data as that used by Korhonen *et al.* (2008). The data was downloaded from the MEDLINE database, from eight journals covering various areas of biomedicine (rather than aiming to focus on any particular sub-domain of biomedicine). The first column in table 5.2 lists each journal, the second shows the years from which the articles were downloaded, and the third indicates the size of the data. According to Korhonen *et al.* (2008), the data is sufficient in size for the various feature sets to have good quality frequency distributions.

1	HAVE AN EFFECT ON ACTIVITY	
1.1	Activate / Inactivate	
1.1.1	Change activity	abolish accelerate activate arrest block disrupt enhance inactivate inhibit
1.1.2	Suppress	repress suppress
1.1.3	Stimulate	stimulate
1.1.4	Inactivate	compromise delay diminish
1.2	Affect	
1.2.1	Modulate	alter modulate stabilize
1.2.2	Regulate	affect control induce influence regulate support
1.3	Increase / decrease	decrease elevate increase
1.4	Modify	catalyze modify
2	BIOCHEMICAL EVENTS	
2.1	Express	express overexpress
2,2	Modification	
2.2.1	Biochemical modification	dephosphorylate phosphorylate
2.2.2	Cleave	cleave
2.3	Interact	coincide colocalize cooperate correlate interact interfere react
3.1	Omit	deplete displace omit
3.2	Subtract	dissect draw subtract
4	EXPERIMENTAL PROCEDURES	
4.1	Prepare	
4.1.1	Wash	rinse wash
4.1.2	Mix	mix
4.1.3	Label	fix immunoblot label probe stain
4.1.4	Incubate	incubate preincubate
4.1.5	Elute	elute
4.2	Precipitate	coimmunoprecipitate coprecipitate precipitate
4.3	Solubilize	lyse solubilize
4.4	Dissolve	dissolve freeze homogenize resuspend suspend
4.5	Place	deposit embed load locate mount place plate seed spot
5	PROCESS	align cut fill linearize overlap
6	TRANSFECT	cotransfect inject microinject transfect
7	COLLECT	
7.1	Collect	collect harvest select
7.2	Process	centrifuge process recover
8	PHYSICAL RELATION BETWEEN MOLECULES	
8.1.0	Binding	attach bind conjugate couple fuse hybridize tether
8.2	Translocate and Segregate	
8.2.1	Translocate	redistribute shift switch translocate
8.2.2	Segregate	export segregate
8.3	Transmit	
8.3.1	Transport	deliver extend transmit transport
8.3.2	Link	connect link map
9	REPORT	
9.1	Investigate	
9.1.1	Examine	analyze assess estimate evaluate examine explore
9.1.2	Establish	establish investigate test
9.1.3	Confirm	confirm determine verify
9.2	Suggest	
9.2.1	Presentational	argue assume conclude hypothesize note reason speculate
9.2.2	Cognitive	believe consider mean postulate predict propose think
9.3	Indicate	demonstrate imply indicate suggest
10	PERFORM	
10.1	Quantify	
10.1.1	Quantitate	measure monitor quantify quantitate
10.1.2	Calculate	calculate record
10.1.3	Conduct	conduct perform
10.2	Score	count score
11	RELEASE	detach dissociate excise release
12	USE	employ exploit use utilize
13	INCLUDE	
13.1	Encompass	bear comprise encompass harbor possess span
13.2	Include	carry constitute contain include underlie
14	CALL	call designate name
15	MOVE	
15.1	Proceed	move pass point proceed progress recycle traffic
15.2	Emerge	arise come disappear emerge originate
16	APPEAR	appear become occur prove remain seem

Table 5.1: The gold standard classification

Journal	Years	Words
Genes & Development	2003-5	4.7M
Journal of Biological Chemistry	2004 (Vol.1-9)	5.2M
The Journal of Cell Biology	2003-5	5.6M
Cancer Research	2005	6.5M
Carcinogenesis	2003-5	3.4M
Nature Immunology	2003-5	2.3M
Drug Metabolism and Disposition	2003-5	2.3M
Toxicological Sciences	2003-5	3.1M
Total:		33.1M

Table 5.2: Data from Korhonen *et al.* (2008) downloaded from MEDLINE

5.4 Features

Korhonen *et al.* (2008) investigated optimal features for biomedical verb classification, including lexical, syntactic and semantic ones. Their experiments showed that feature sets containing both syntactic and semantic information perform the best on this task. To facilitate direct comparison of our results against those of Korhonen *et al.* (2008), we adopted the same feature sets for our experiments. We briefly describe each feature set below.

Table 5.3 provides the mapping of our features to the features shown earlier in table 2.4. These features along with other features used in this thesis, are listed in appendix A. The basic SCF features were extracted using Preiss *et al.* (2007)'s system as in one earlier experiment. The system tags, lemmatizes and parses the corpus data using the RASP toolkit (Briscoe *et al.*, 2006). RASP is a domain independent parser. We did not use the bio-tuned Enju parser and GENIA tagger as used for creating BioLexicon, because we wanted to evaluate new clustering methods and keep the feature sets similar to those used in the previous works. However, the biomedical version of Enju does not outperform general parsers (GENIA retrained) in biomedical event extraction tasks (Miyao *et al.*, 2008), so using a bio-tuned Enju parser might not give optimal performance anyway.

The SPs are acquired using the method described in section 3.1.2.

5.5 Experimental evaluation

5.5.1 Experimental settings

We set the number of clusters (Ks) to be the number of clusters in the 3-level gold standard: 16, 34, 50. Both SPEC and the HGFC have a random element. We therefore

F1:	F-SCF
F2:	F-SCF+PP(A)
F3:	F-SCF+PP(B)
F4:	F-SCF+TENSE(A)
F5:	F-SCF+TENSE(B)
F6:	F-SCF+VOICE(A)
F7:	F-SCF+VOICE(B)
F8:	F-SCF+SP(A)
F9:	F-SCF+SP(B)
F10:	F-SCF+SP(C)

Table 5.3: The mapping to the features in table 2.4

repeat the K-Means module of the SPEC 100 times with random initialization and the result that minimizes average distance to the centroid is used. Also the HGFC algorithm is run 100 times and the result that minimizes the objective function of HGFC is used.

5.5.2 Measures

We employed the same measures (*m*PUR, ACC and F) as previously employed by Korhonen *et al.* (2008) in order to facilitate the meaningful comparison of results.

We performed McNemar's statistical significance test (McNemar, 1947) for the major findings. The details of the test are described in section 3.3.4.

5.5.3 Results

Tables 5.4 and 5.5 show the results for SPEC and HGFC, respectively. The results (first given for each individual feature set) are directly comparable with the PC clustering results in Korhonen *et al.* (2008) where the same corpus and gold standard was used. Recall that F1-F3 include the basic SCF features (F1) refined with information about prepositions (F2-F3). As in Korhonen *et al.* (2008), F2-F3 mostly outperforms F1. However, while in Korhonen *et al.* F3 did not outperform F2, in our case, the improvement is significant ($p < 0.05$). Parameterizing additional frames with prepositions is clearly helpful, yielding useful class distinctions (for example, the SCF NP-P-NP-ING has different variations, depending on the preposition in question, e.g. he **attributed** his failure **to** buying his books; he **told** her **about** climbing the mountain).

Recall that F4-F10 refine F3 further with additional information. In our case, the performance is not improved with F4 and F5 (the verb tense features). This contrasts with the result of Korhonen *et al.*, but is in line with our result on general language verb classification in chapter 3. As in Korhonen *et al.*'s experiment, F6-F7 (verb voice features) do not prove helpful. F4 combines verb tense with the SCF feature by simple concatenation. Since a concatenated feature cannot be seen as a probabilistic distribution anymore, it is expected that a distributional similarity measure like JSD will underperform. F5 parameterizes the SCF feature with verb tense information. According to Merlo and Stevenson (2001), verb tense and voice features are related to the transitive/intransitive alternation, since the use of a past participle or a passive voice implies a transitive use of the verb. If the voice feature does not improve the result over F3, the verb tense feature should not provide any improvement as they are providing similar information. In addition, the SCFs already indicate the transitive/intransitive use of a verb, so the verb tense information provides no new information.

Recall that F8-F10 parameterize F3 with additional information about LPs and SPs. F8 and F9 parameterize F3 with LPs, while F10(a-c) supplements F3 with automatically acquired information about SPs. Like in chapter 3, these feature sets are most useful. When SPEC is used, F8-F10 outperform other features sets by a large margin. In the case of HGFC, the improvement is smaller, but still consistent. F10a is the best-performing feature with SPEC. It performs at F-Measure of 82.2, 76.4 and 74.0 at the three levels of the gold standard (16, 34 and 50 classes), respectively. This shows that SP information can perform very well with SPEC – a finding we reported earlier also on general English (see section 3). With HGFC, the best performing feature is also a SP-based feature, but F10c gives the best performance at 16 and 50 classes, while F10a performs best at 34 classes. The largest improvement over F8 is found at the coarse-grained level of 16 classes (from 72.8 to 76.1). This indicates that SPs can offer a larger improvement over LP features on the coarse-grained level than the other levels.

Comparing our own two methods (SPEC to HGFC), SPEC outperforms HGFC on almost every feature set except at 50 classes using F5 and at 16 classes using F10c. The differences in the best performances at the 3 levels (16, 34 and 56 classes) are 10.8, 3.5 and 4.1 in F respectively. This is expected since HGFC produces a tree-consistent hierarchy, but SPEC and other flat clustering algorithms (e.g. K-Means and PC) produce 3 flat clustering results for 3 levels which are not tree-consistent. Tree-consistency is important as all the existing Levin-style verb classification gold standards are tree-consistent. Tree-consistency is also important for humans to understand and interpret the resulting hierarchy. For example, if two verbs *v1* and *v2* are assigned into the same cluster at the 50 classes level, *v1* and *v2* can't belong to different clusters on the upper levels (34 and 16 classes). In chapter 4, we showed that without the tree-consistency constraint, HGFC performs similarly to SPEC.

When comparing our results against those of Korhonen *et al.*'s, SPEC outperforms PC ($p < 0.05$) on almost all feature sets and class number settings (except on F4 and F5 at 34 and 50 classes where the results worsen slightly or the improvement is not significant ($p > 0.05$)). The biggest improvement can be observed at the level of 16 classes (7.9 on average over all feature sets). On the level of 34 and 50 classes, the average improvement over all feature sets is 3.2 and 3.8 in F respectively. The best performance is improved over Korhonen *et al.*'s best result by 11.9, 4 and 5.2 in F for 16, 34 and 50 classes using SPEC. For HGFC, the performance is not always improved. Small improvements and declines can be found in F1-F9. However, consistent improvements can be observed using F10a-c (2.26 on average, $p < 0.05$ except F10b and F10c at 34 classes). The best performance at 16, 34 and 50 classes is then improved by 4.4, 3 and 2.9, respectively. As discussed above, the tree-consistency requirement is the main reason that the HGFC does not always perform better than PC.

We further investigated whether combining the best features might yield higher overall performance. We constructed F8+F10(a-c) which combine LP and SP information. For SPEC, the performance at 16 classes is further improved by 4.7 in F when compared to the best performance with F8 and F10(a-c). For 30 and 54 classes, the best performance stays the same. For HGFC, the performance is not further improved by the combination of LP and SP. For both methods, F8+F10c outperforms F8 consistently ($p < 0.05$ except at 50 classes using SPEC), showing that adding SP information on top of LP is useful. However, F8+F10a-c do not consistently outperform F10a-c, demonstrating that adding LP information over SP is not always helpful. A possible explanation is that since SPs were acquired automatically and are therefore noisy, the actual LPs can sometime provide more detailed information which the SPs lack. At the level of 16 classes, SPs are much less clear than at the level of 34 and 50 classes, and therefore LPs can make a real difference.

		16 Classes			34 Classes			50 Classes		
		*m*PUR	ACC	F	*m*PUR	ACC	F	*m*PUR	ACC	F
SCF	F1	76.6	74.5	75.6(+15)	65.8	65.1	65.6(+9.7)	61.6	61.5	61.6(+8.6)
	F2	75.6	76.0	75.9(+4.8)	68.2	67.2	67.7(+2.1)	65.4	63.5	64.5(+1.3)
	F3	77.9	78.6	78.3(+7.3)	74.7	75.0	74.8(+5)	69.9	69.8	69.8(+5.7)
F3 + tense	F4	74.7	75.0	74.8(+3.8)	68.7	69.3	69.0(-0.8)	64.5	64.6	64.6(+0.5)
	F5	76.6	76.0	76.3(+5.3)	70.1	70.3	70.2(+0.4)	64.1	63.5	63.8(-0.3)
F3+voice	F6	73.5	77.1	77.2(+6.8)	70.3	69.8	70.0(+4.9)	67.5	68.2	67.9(+3.5)
	F7	79.2	78.0	78.6(+10.4)	70.3	70.4	70.4(+5.2)	66.7	67.1	66.9(+2.2)
F3+SP	F8	79.6	79.2	79.4(+9.5)	74.9	72.9	73.9(+3.7)	73.7	71.9	72.7(+6.1)
	F9	78.6	76.6	77.4(+7.5)	76.7	74.5	75.5(+3.7)	73.7	72.4	72.9(+4.1)
	F10A	83.7	80.7	82.2(+11.6)	78.7	74.0	**76.4**(+6.5)	74.4	71.9	**74.0**(+5.2)
	F10B	84.2	79.7	81.9(+9)	74.5	68.8	71.5(+1.3)	71.6	71.4	71.5(+4.9)
	F10C	79.4	72.9	75.9(+4.2)	74.2	69.8	71.9(+1.2)	72.8	70.8	71.8 (+4.8)
	F8 + F10a	84.2	82.8	83.5	78.1	71.9	74.9	74.9	72.9	**73.9**
	F8 + F10b	86.6	84.8	85.7	78.2	73.4	75.7	73.3	71.9	72.6
	F8 + F10c	89.6	84.3	**86.9**	78.4	74.5	**76.4**	74.0	72.5	73.2

Table 5.4: SPEC clustering results in comparison to the results in Korhonen *et al.* (2008)

We performed further qualitative analysis of clusters produced by SPEC and HGFC with their best features. The most common error type was misclustering due to purely syn-

		16 Classes			34 Classes			50 Classes		
		*m*PUR	ACC	F	*m*PUR	ACC	F	*m*PUR	ACC	F
SCF	F1	70.0	66.7	68.3$_{(+7.7)}$	59.7	59.9	59.8$_{(+4)}$	59.5	57.8	58.7$_{(+4.8)}$
	F2	69.4	65.1	67.2$_{(-3.9)}$	65.8	64.1	64.9$_{(-0.7)}$	59.2	59.4	59.3$_{(-3.9)}$
	F3	75.9	68.8	72.1$_{(-0.1)}$	68.4	67.2	67.8$_{(+2.8)}$	65.6	66.7	66.1$_{(+3.2)}$
F3+tense	F4	74.0	70.8	72.4$_{(+1.4)}$	69.3	67.7	68.5$_{(-1.3)}$	60.3	59.9	60.1$_{(-4)}$
	F5	72.4	71.4	71.9$_{(-1.5)}$	67.9	66.1	67.0$_{(-0.1)}$	66.6	65.6	66.1$_{(-0.7)}$
F3+voice	F6	73.2	68.8	70.9$_{(+0.5)}$	69.0	66.7	67.8$_{(+1.9)}$	66.5	66.7	66.6$_{(+2.2)}$
	F7	76.7	69.8	73.1$_{(+4.9)}$	68.8	67.2	68.0$_{(+2.8)}$	66.1	65.6	65.9$_{(+1.2)}$
F3+SP	F8	76.1	69.8	72.8$_{(+3.9)}$	71.4	69.8	70.6$_{(+0.4)}$	67.6	67.9	67.8$_{(+2.1)}$
	F9	82.5	53.1	64.6$_{(-5.3)}$	70.9	63.0	66.7$_{(-5.1)}$	65.9	61.5	63.6$_{(-5.2)}$
	F10a	76.3	73.9	75.1$_{(+4.5)}$	74.7	71.4	**72.9**$_{(+3)}$	68.1	68.8	68.4$_{(+2.2)}$
	F10b	77.1	70.8	73.8$_{(+0.9)}$	70.7	67.7	69.2$_{(+0.07)}$	69.3	67.2	68.2$_{(+2.3)}$
	F10c	77.7	74.5	**76.1**$_{(+4.4)}$	72.8	69.8	71.3$_{(+0.06)}$	70.7	69.3	**69.9**$_{(+2.9)}$
	F8+F10a	69.0	66.7	67.8	69.5	67.7	68.6	65.3	65.1	65.2
	F8+F10b	77.1	70.8	73.8	70.7	67.7	69.2	69.3	67.2	68.2
	F8+F10c	77.7	74.5	76.0	72.8	69.8	71.3	70.7	69.3	70.0

Table 5.5: HGFC clustering results in comparison to the results in Korhonen *et al.* (2008)

tactic similarity. For example, the APPEAR verb *occur* showed up in the same cluster with MOVE verbs *pass, proceed* and *progress* because it frequently appears with locative arguments, like MOVE verbs tend to do. However, an equally common error type was misclassification due to genuine semantic similarity, which was not apparent in the gold standard. Some such cases were due to polysemy. The gold standard was not created like previous similar gold standards where verbs were classified on the basis of their predominant sense in language. This was not possible because no sense-annotated data was available. However, Korhonen *et al.* (2006b) relied on the assumption that verbs tend to have one predominating sense in a domain. As the annotators were given syntactically similar verb distributions to consider as a starting point, it is indeed possible that if a predominating sense exists, the gold standard captures it in most cases. Yet we found several examples which showed that polysemy plays a factor in this data as well. For example, while the verb *diminish* belongs to the gold standard class of INACTIVATE it is also a perfectly valid member of the class INCREASE/DECREASE, and indeed gets clustered together with its member verbs. Another type of semantic ”error” was due to the members of distinct, but semantically related gold standard classes being clustered together.

Although the gold standard is hierarchical, it fails to relate some classes together which share aspects of meaning. Even looking at the clustering output at the finest-grained level of 50 classes, our methods confuse classes such as EXPERIMENTAL PROCEDURES and COLLECT which are clearly both related to experiments but distinct in the gold standard. Other such examples include the INVESTIGATE and QUANTIFY classes, which include very similar general scientific verbs (e.g. *determine, estimate, measure*). Although the clustering methods tend to respect the basic division between biomedical and general verbs made by annotators, usually assigning them in different classes, in some cases both general and biomedical verbs are found in the same cluster. Sometimes this confusion is due to the genuine relatedness of the classes, e.g. the biomedical class TRANSPORT is related in meaning to the general class MOVE. Looking at the clusterings at the higher levels of the hierarchy reveals further, larger groupings of semantically related classes which are distinct in the gold standard. This demonstrates how clustering can be used to hypothesise both flat and hierarchical verb classifications.

5.6 Summary

In this work, we applied our flat and hierarchical clustering methods (SPEC and HGFC) to the verb classification task in the biomedical domain. Using the same feature and dataset as in previous work (Korhonen *et al.*, 2008), both of our methods outperformed the best previous method (PC). In particular, HGFC produced a tree-consistent hierarchy while PC can only output three levels of flat clusterings. Tree-consistency is important, as all the manual classifications are tree-consistent. As in the general domain, our experiments demonstrated that the feature sets containing both syntactic and semantic information performed the best. We analyzed the errors in a qualitative analysis, and discovered that our classification includes some genuine classes that are missing from the gold standard. In sum, this experiment showed that our clustering approach can be applied to the biomedical domain without any change, indicating that the approach could be used to construct large-scale verb resource for biomedical domain.

5.7 Future Work

Future work could investigate using a bio-tuned tagger and parser for more accurate processing of data. The currently available Enju parser does not outperform other parsers that are retrained on the same biomedical corpus (e.g. Charniak (2000) and Sagae and Tsujii (2007)). But according to Miyao *et al.* (2008), retrained parsing models outperform their original models. Because GENIA contains only molecular biology publications related to *E.Coli*, lexicalized parser models might not be able to generalize well to other biomedical subdomains. Therefore, one could first perform domain-specific tuning of the unlexicalized RASP parser on the GENIA corpus, and then select the best performing parser among the retrained RASP parser, Enju parser and Charniak's Parser for processing of the data.

In addition, the resulting verb classification could be evaluated in the context of BIO-NLP tasks. For example, semantic role labelling of biomedical text has employed VerbNet classes as features in the classifier (Chou *et al.*, 2006). Verb clustering could improve this approach further by yielding classes specific to biomedicine that are missing in VerbNet.

Chapter 6

Cross-linguistic potential of verb classification

6.1 Introduction

[1]Real-world exploitation of Levin style classes has been limited because for most languages, no such classes are available. To date most work on Levin type classification has focussed on English. Large-scale research on other languages such as German (Schulte im Walde, 2006) and Japanese (Suzuki and Fukumoto, 2009) has focussed on semantic classification. Although there are similarities between the two classification systems, studies comparing the overlap between VerbNet and WordNet (Miller, 1995) have reported that the mapping is only partial and many to many due to the fine-grained nature of classes based on synonymy (Kipper-Schuler, 2005; Shi and Mihalcea, 2005; Abend *et al.*, 2008).

Only few studies have been conducted on Levin style classification for languages other than English. In their experiment involving 59 verbs and three classes, Merlo *et al.* (2002) applied a supervised approach developed for English to Italian, obtaining high accuracy (86.3%). In another experiment with 60 verbs and three classes, they showed that features extracted from Chinese translations of English verbs can improve English classification. These results are promising, but those from a later experiment by Ferrer (2004) are not. Ferrer applied a clustering approach developed for English to Spanish, and evaluated it against the manual classification of Vázquez *et al.* (2000), constructed using criteria similar (but not identical) to Levin's. This experiment involving 514 verbs and 31 classes produced results only slightly better than the random baseline.

In this chapter, we investigate the cross-linguistic potential of Levin style classification further. In past years, verb classification techniques – in particular unsupervised ones – have improved considerably, making investigations for a new language more feasible. We

[1] The research reported in this chapter was published in Sun *et al.* (2010).

take the SPEC verb clustering method developed for English (see section 3.2) and apply it to French – a major language for which no such experiment has been conducted yet. Basic NLP resources (corpora, taggers, parsers and subcategorization acquisition systems) are now sufficiently developed for this language for the application of a state-of-the-art verb clustering approach to be realistic.

Our investigation reveals similarities between the English and French classifications, supporting the linguistic hypothesis (Jackendoff, 1990) and the earlier result of Merlo *et al.* (2002) that Levin classes have a strong cross-linguistic basis. Not only the general methodology but also the best performing features are transferable between the languages, making it possible to learn useful classes for French automatically and without the need for language-specific tuning.

This is joint work with Thierry Poibeau and Cedric Messiant. The French gold standard was constructed by Thierry Poibeau. The SCF lexicon LexSchem was provided by Cedric Messiant. The rest of the work was performed by the author of this thesis.

6.2 French verb classes and the gold standard

The development of an automatic verb classification approach requires at least an initial gold standard. Some syntactic (Gross, 1975) and semantic (Vossen, 1998) verb classifications exist for French, along with ones which aim to integrate aspects of both (Saint-Dizier, 1998). Although such resources could be combined to hypothesise Levin-style classes for French (using e.g. an approach similar to that employed by Kipper *et al.* (2008)), we adopted a more direct approach: following the idea of Merlo *et al.* (2002), we translated a number of Levin classes from English to French.

We chose an English gold standard which has been used to evaluate several recent clustering works – that of Sun *et al.* (2008b). It includes 17 fine-grained Levin classes. Each class has 12 member verbs whose predominant sense (according to the WordNet frequency data) belongs to that class. We evaluated each class in this resource as follows:

1. Member verbs were first translated to French. Where several relevant translations were identified, each of them was considered.

2. For each candidate verb, SCFs were identified and possible diathesis alternations were considered using the criteria of Levin (1993): alternations must result in the same or extended verb sense. Only verbs sharing diathesis alternations were kept in the class – others were discarded.

For example, the gold standard class 31.1 AMUSE includes the following English verbs: *stimulate, threaten, shock, confuse, upset, overwhelm, scare, disappoint, delight, exhaust, intimidate* and *frighten*. Relevant French translations were identified for all of

Class No	Class	Verbs
9.1	PUT	accrocher, déposer, mettre, placer, répartir, réintégrer, empiler, emporter, enfermer, insérer, installer
10.1	REMOVE	ôter, enlever, retirer, supprimer, retrancher, débarrasser, soustraire, décompter, éliminer
11.1	SEND	envoyer, lancer, transmettre, adresser, porter, expédier, transporter, jeter, renvoyer, livrer
13.5.1	GET	acheter, prendre, saisir, réserver, conserver, garder, préserver, maintenir, retenir, louer, affréter
18.1	HIT	cogner, heurter, battre, frapper, fouetter, taper, rosser, brutaliser, éreinter, maltraiter, corriger,
22.2	AMALGAMATE	incorporer, associer, réunir, mélanger, mêler, unir, assembler, combiner, lier, fusionner
29.2	CHARACTERIZE	appréhender, concevoir, considérer, décrire, définir, dépeindre, désigner, envisager, identifier, montrer, percevoir, représenter, ressentir
30.3	PEER	regarder, écouter, examiner, considérer, voir, scruter, dévisager
31.1	AMUSE	abattre, accabler, briser, déprimer, consterner, anéantir, épuiser, exténuer, écraser, ennuyer, éreinter, inonder,
36.1	CORRESPOND	coopérer, participer, collaborer, concourir, contribuer, prendre part, s'associer, travaille
37.3	MANNER OF SPEAKING	râler, gronder, crier, ronchonner, grogner, bougonner, maugréer, rouspéter, grommeler, larmoyer, gémir, geindre, hurler, gueuler, brailler, chuchoter
37.7	SAY	dire, révéler, déclarer, signaler, indiquer, montrer, annoncer, répondre, affirmer, certifier, répliquer
43.1	LIGHT EMISSION	briller, étinceler, flamboyer, luire, resplendir, pétiller, rutiler, rayonner., scintiller
45.4	CHANGE OF STATE	mélanger, fusionner, consolider, renforcer, fortifier, adoucir, polir, atténuer, tempérer, pétrir, façonner, former
47.3	MODES OF BEING	trembler, frémir, osciller, vaciller, vibrer, tressaillir, frissonner, palpiter, grésiller, trembloter, palpiter
51.3.2	RUN	voyager, aller, se promener, errer, circuler, se déplacer, courir, bouger, naviguer, passer

Table 6.1: A Levin style gold standard for French

them: *abattre, accabler, briser, déprimer, consterner, anéantir, épuiser, exténuer, écraser, ennuyer, éreinter, inonder*. The majority of these verbs take similar SCFs and diathesis alternations, e.g. *Cette affaire écrase Marie (de chagrin), Marie est écrasée par le chagrin, Le chagrin écrase Marie*. However, *stimuler* (*stimulate*) and *menacer* (*threaten*) do not, and they were therefore removed.

40% of translations were discarded from classes after step 2 was applied. The final version of the gold standard (shown in table 6.1) includes 171 verbs in 16 classes. Each class is named according to the original Levin class. The smallest class (30.3) includes 7 verbs and the largest (37.3) 16. The average number of verbs per class is 10.7.

6.3 Verb clustering

We performed an experiment where we

- took a French corpus and a SCF lexicon automatically extracted from that corpus using French NLP technology,
- extracted from these resources a range of features (lexical, syntactic and semantic) – a representative sample of those employed in recent English experiments (Joanis *et al.*, 2007; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009b) and also those in section 3.
- clustered the features using a method which has proved promising in both English and German experiments: SPEC,
- evaluated the clusters both quantitatively (using the gold standard) and qualitatively,
- and finally, compared the performance of individual features to that recently obtained for English in order to gain a better understanding of the cross-linguistic and language-specific properties of verb classification

6.3.1 Data: the LexSchem lexicon

We extracted the features for clustering from LexSchem (Messiant *et al.*, 2008). This large subcategorization lexicon provides SCF frequency information for 3,297 French verbs. It was acquired fully automatically from Le Monde newspaper corpus (200M words from the period 1991-2000) using ASSCI – a recent subcategorization acquisition system for French (Messiant, 2008).

Systems similar to ASSCI have been used in recent verb classification works e.g. (Schulte im Walde, 2006; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008). Like these other systems, ASSCI takes raw corpus data as input. The data is first tagged and lemmatized using the Tree-Tagger and then parsed using the Syntex parser (Bourigault *et al.*, 2005). Syntex is a shallow parser which employs a combination of statistics and heuristics to identify grammatical relations (GRs) in sentences. It is a state-of-the-art French parser: it obtained the best precision and F-measure for written texts in the recent EASY evaluation[2].

ASSCI considers those GRs where the target verbs occur and constructs SCFs from nominal, prepositional and adjectival phrases, and infinitival and subordinate clauses. When a verb has no dependency, its SCF is considered as intransitive. Otherwise, ASSCI assumes no pre-defined list of SCFs, but almost any combination of permitted constructions can appear as a candidate SCF. The number of automatically generated SCF types in LexSchem is 336.

Many of the candidate SCFs are noisy due to processing errors and the difficulty of argument-adjunct distinction. Most SCF systems operate on the basis of the assumption that true arguments occur in argument positions more frequently than adjuncts. Many of them also integrate sophisticated filters for removing noise from the system output. When LexSchem was evaluated using a relative frequency and heuristics-based filter its F-measure was 69 – which is similar to those of other current SCF systems (Messiant *et al.*, 2008) However, we used the unfiltered version of LexSchem because previous work on English verb classification has showed that information about adjuncts can actually help verb clustering (Sun *et al.*, 2008b).

6.4 Features

Lexical entries in LexSchem provide a variety of material for verb clustering, including (statistical) information related to the POS tags, SCFs, argument heads, and adjuncts of verbs. Using this material, we constructed a range of features for experimentation. The first three include basic information about SCFs:

F1: F-SCF

F2: F-SCF+TENSE(B)

F3: F2, with SCFs parameterized for prepositions (PP).

The following six features include information about the lexical context (co-occurrences) of verbs. We adopt the best method of Li and Brew (2008) where collocations (COs) are extracted from the window of words immediately preceding and following a lemmatized verb. Stop words are removed prior to extraction.

[2] See `http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/syntex.html` for details.

F4, F6, F8: COs are extracted from the window of 4, 6 and 8 words, respectively. The relative word position is ignored.

F5, F7, F9: F4, F6 and F8 with the relative word position recorded.

The next four features include information about lexical preferences (LP) of verbs in argument head positions of specific GRs associated with the verb:

F10: LP(PREP): the type and frequency of prepositions in the preposition (PREP) relation.

F11: LP(SUBJ): the type and frequency of nouns in the subject (SUBJ) relation.

F12: LP(IOBJ): the type and frequency of nouns in the object (OBJ) and indirect object (IOBJ) relation.

F13: LP(ALL): the combination of F10-F13.

The final two features refine SCF features with LPs and semantic information about verb SP:

F14-F16: F1-F3 parameterized for LPs.

F17: F3 refined with SPs.

We adopt the fully unsupervised approach to SP acquisition described in chapter 3, with the difference that we determine the optimal number of SP clusters automatically following Zelnik-Manor and Perona (2004). The method is introduced in the following section. The approach involves (i) taking the GRs (subj, obj, iobj) associated with verbs, (ii) extracting all the argument heads in these GRs, and (iii) clustering the resulting N most frequent argument heads into M classes. An empirically determined N of 200 was used. The method produced 40 SP clusters.

Features F1, F2, F4-F9, F10-F13 and F14 are used in previous verb classification experiments. More details and the feature extraction methods are listed in table 2.4. All the above features and other features used in this thesis are summarized in appendix A.

6.5 Clustering methods

We used the SPEC method described in chapter 3 with JSD as the similarity measure. The reason why JSD is used and details on other similarity measures can be found in appendix B. As the number of clusters is not known beforehand, we use Zelnik-Manor and Perona (2004)'s method to estimate it. This method finds the optimal value by minimizing a cost function based on the eigenvector structure of the similarity matrix.

Like Brew and Schulte im Walde (2002), we compare SPEC against a K-Means baseline. We used the Matlab implementation with Euclidean distance as the distance measure.

6.6 Experimental evaluation

6.6.1 Data and pre-processing

Our initial plan was to experiment with all the 171 verbs in the gold standard (see Table 6.1). However, we decided to exclude phrasal verbs (e.g. *faire disparaître*) and drop one class (40.2 NON-VERBAL EXPRESSION) which included reflexive verbs in French (e.g. *s'amuser, se moquer*) since multiword units would have been challenging for our method. Also verbs assigned to several classes due to polysemy were excluded. This left us with 147 verbs in 15 classes (10 verbs per class on average).

The SCF-based features (F1-F3 and F14-F17) were extracted directly from LexSchem. The CO (F4-F9) and LP features (F10-F13) were extracted from the raw and parsed corpus sentences, respectively, which were used for creating the lexicon. Features that only appeared once were removed. Feature vectors were normalized by the sum of the feature values before clustering. Since our clustering algorithms have an element of randomness, we repeated clustering multiple times. We report the results that minimize the distortion (the distance to cluster centroid).

6.6.2 Evaluation measures

We employed the same measures (*m*PUR, ACC and F) for evaluation as in our experiments on English (chapter 3).

We performed McNemar's statistical significance test (McNemar, 1947) to verify the major findings. The details of the test are described in section 3.3.4.

6.7 Evaluation

6.7.1 Quantitative evaluation

In our first experiment, we evaluated 116 verbs – those which appeared in LexSchem a minimum of 150 times. We did this because English experiments had shown that due to the Zipfian nature of SCF distributions, 150 corpus occurrences are typically needed to obtain a sufficient number of frames for clustering (Sun *et al.*, 2008b). The amount of corpus evidence needed for reliable clustering is discussed in section 8.2.3.

Table 6.2 shows F-measure results for all the features. The 4th column of the table shows, for comparison, the results (in the section 3.4) obtained for English. The results for English are obtained using SPEC with same features, which are evaluated against the English version of the same gold standard, also using F-measure[3].

As expected, SPEC (the 2nd column) outperforms K-Means (the 3rd column) throughout the feature sets with an average improvement of 14.1 on F ($p < 0.05$). Looking at the basic SCF features F1-F3, we can see that they perform substantially better than the BL method. F3 performs the best among the three features both in French (50.6 F, $p < 0.05$) and in English (63.3 F). We therefore use F3 as the SCF feature in F14-F17 (the same was done for English).

In French, most CO features (F6-F9) outperform SCF features ($p < 0.05$). The best result is obtained with F7: 55.1 F. This is clearly better than the best SCF result 50.6 (F3). This result is interesting since SCFs correspond better than COs to features used in the manual Levin classification. Also, SCFs perform considerably better than COs in the English experiment (we only have the result for F4 available, but it is considerably lower than the result for F3). However, earlier English studies have reported contradictory results (e.g. Li and Brew (2008) showed that CO performs better than SCF in supervised verb classification), indicating that the role of CO features in verb classification requires further investigation.

Looking at the LP features, F13 produces the best F (52.7) for French which is slightly better than the best SCF result for the language. Also in English, F13 performs the best in this feature group and yields a higher result than the best SCF-based feature F3.

Parameterizing the best SCF feature F3 with LPs (F14-16) and SPs (F17) yields better performance in French. F15 and F17 have an F-measure of 54.5 and 54.6, respectively. These results are so close to the result of the best CO feature F7 (55.1 – which is the highest result in this experiment) that the differences are not statistically significant ($p > 0.05$). In English, the results of F14-F17 are similarly good; however, only F17 beats the already high performance of F13.

On the basis of this experiment, it is difficult to tell whether shallow CO features or more sophisticated SCF-based features are better for French. In the English experiment sophisticated features performed better, and the SCF-SP feature F17 was the best one. However, the English experiment employed a much larger dataset. These more sophisticated features may suffer from data sparseness in our French experiment since although we required the minimum of 150 occurrences per verb in LexSchem, verb clustering performance tends to improve when more data is available, and given the fine-grained nature of LexShem SCFs it is likely that more data is required for optimal performance.

We therefore performed another experiment with French on the full set of 147 verbs, using SPEC, where we investigated the effect of instance filtering on the performance of

[3] Note that the results for the two languages are not mutually comparable due to differences in test sets, data sizes, and feature extraction systems. The results for English are included so that we can compare the relative performance of individual features in the two languages in question.

the best features from each feature group: F3, F7, F13 and F17. The results shown in Table 6.3 reveal that the performance of the features remains fairly similar until the instance threshold of 1000. When 2000 occurrences per verb are used, the differences become clearer, until at the threshold of 4000, it is obvious that the most sophisticated SCF-SP feature F17 is by far the best feature for French (65.4 F, $p < 0.05$ when compared to F3) and the SCF feature F3 the second best (60.5 F). The CO-feature F7 and the LP feature F13 are not nearly as good (53.4 and 51.0 F).

Although the results at different thresholds are not comparable due to the different number of verbs and classes (see columns 2-3), the results for features at the same threshold are. Those results suggest that when 2000 or more occurrences per verb are used, most features perform like they performed for English in the experiment (chapter 3) with CO being the least informative[4] and SCF-SP being the most informative feature. The only exception is the LP feature which performed relatively better than CO in English.

		French		English
		SPEC	K-Means	SPEC
BL		6.7	6.7	6.7
F1	SCF	42.4	39.3	57.8
F2	SCF(POS)	45.9	40.3	46.7
F3	SCF(PP)	**50.6**	36.9	63.3
F4	CO(4)	50.3	38.2	40.9
F5	CO(4+loc)	48.8	26.3	-
F6	CO(6)	52.7	29.2	-
F7	CO(6+loc)	**55.1**	33.8	-
F8	CO(8)	54.2	36.4	-
F9	CO(8+loc)	54.6	37.2	-
F10	LP(PREP)	35.5	32.8	49.0
F11	LP(SUBJ)	33.7	23.6	-
F12	LP(OBJ)	50.1	33.3	-
F13	LP(ALL)	**52.7**	40.1	74.6
F14	SCF+LP(SUBJ)	50.3	40.1	71.7
F15	SCF+LP(OBJ)	**54.5**	35.6	74.0
F16	SCF+LP(SUBJ+OBJ)	53.4	36.2	73.0
F17	SCF+SP	54.6	39.8	80.4

Table 6.2: Results for all features for French (SPEC and K-means) and English (SPEC)

[4] However, it is worth noting that CO is not a useless feature. As table 6.3 shows, when 150 or fewer occurrences are available for a verb, CO outperforms all the other features in French, compensating for data sparseness.

THR	Verbs	Cls	F3	F7	F13	F17
0	147	15	43.7	57.5	43.3	50.1
50	137	15	47.9	56.1	44.8	49.1
100	125	15	49.2	54.3	44.8	49.5
150	116	15	50.6	55.1	52.7	54.6
200	110	15	54.9	52.9	49.7	52.5
400	96	15	52.7	52.9	43.9	53.2
1000	71	15	51.4	54.0	44.8	54.5
2000	59	12	52.3	45.9	42.7	53.5
3000	51	12	55.7	49.0	46.8	59.2
4000	43	10	60.5	53.4	51.0	**65.4**

Table 6.3: The effect of verb frequency on performance

6.7.2 Qualitative evaluation

We conducted qualitative analysis of the clusters for French, focusing on those created using SPEC with F17 and F3.

Verbs in the gold standard classes 29.2, 36.1, 37.3, 37.7 and 47.3 (Table 6.1) performed particularly well, with the majority of member verbs found in the same cluster. These verbs are ideal for clustering because they have distinctive syntactic-semantic characteristics. For example, verbs in 29.2 CHARACTERIZE class (e.g. *concevoir, considérer, dépeindre*) not only have a very specific meaning but they also take high frequency SCFs involving the preposition *comme* (Eng. *as*) which is not typical to many other classes. Interestingly, Levin classes 29.2, 36.1, 37.3, and 37.7 were among the best performing classes also in the English supervised verb classification experiment of Sun *et al.* (2008b) (which employed the English version of our gold standard) because these classes have distinctive characteristics also in English.

The benefit of sophisticated features which integrate also semantic (SP) information (F17) is particularly evident for classes with non-distinctive syntactic characteristics. For example, the intransitive verbs in 43.1 LIGHT EMISSION class (e.g. *briller, étinceler, flamboyer*) are difficult to cluster based on syntax only, but semantic features work because the verbs pose strong SPs on their subjects (entities capable of light emission). In the experiment of Sun *et al.* (2008b), 43.1 was the worst performing class for English, possibly because no semantic features were used in the experiment.

The most frequent source of error is syntactic idiosyncracy. This is particularly evident for classes 10.1 REMOVE and 45.4 CHANGE OF STATE. Although verbs in these classes can take similar SCFs and alternations, only some of them are frequent in data. For example, the SCF *ôter X à Y* is frequent for verbs in 10.1, but not *ôter X de Y*. Although class 10.1 did not suffer from this problem in the English experiment of Sun *et al.* (2008b), class 45.4 did. Class 45.4 performs particularly bad in French also because its member verbs are low in frequency.

Some errors are due to polysemy, caused partly by the fact that the French version of the gold standard was not controlled for this factor. Some verbs have their predominant senses in classes which are missing in the gold standard, e.g. the most frequent sense of *retenir* is *memorize*, not *keep* as in the gold standard class 13.5.1. GET.

Finally, some errors are not true errors but demonstrate the capability of clustering to learn novel information. For example, the CHANGE OF STATE class 45.4 includes many antonyms (e.g. *weaken* vs. *strengthen*). Clustering (using F17) separates these antonyms, so that verbs *adoucir, atténuer* and *tempérer* appear in one cluster and *consolider* and *renforcer* in another. Although these verbs share the same alternations, their SPs are different. For the same reason, verbs in LIGHT EMISSION class 43.1 end up in different clusters, depending on whether they describe abstract or concrete light emission.

The opposite effect can be observed when clustering maps together classes which are actually semantically and syntactically related (e.g. 36.1 CORRESPOND and 37.7 SPEAK). Such classes are distinct in Levin and VerbNet, because these resources do not to draw links between semantically similar classes belonging to different main classes.

Cases such as these show the potential of clustering in discovering novel valuable information in data. It is encouraging that we have observed this effect in this first clustering experiment in French.

6.8 Discussion and conclusion

We have seen that when sufficient corpus data is available, there is a strong correlation between the types of features which perform the best in English and French. Interestingly, we have also seen that when the best features are used, many individual Levin classes have similar performance in the two languages.

Due to differences in language-specific datasets and sizes, a direct comparison of the actual performance figures for English and French is not possible. When considering the general level of performance, our best performance for French (65.4 F) is clearly lower than the best performance for English in the section 3. However, it compares favourably to the performance of other state-of-the-art (even supervised) systems for English verb classification (Joanis *et al.*, 2007; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009b). This is impressive considering that we experimented with a fully unsupervised method originally developed for another language.

Our experiment suggests that when aiming to improve performance further, employing larger data is critical. Most recent experiments on English have employed bigger test and datasets, and unlike us, some of them have only considered the predominant senses of medium-high frequency verbs (Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009b). As seen in section 6.7.1, such differences in data can have significant impact on performance.

However, parser and feature extraction performance can also play a big role in overall accuracy, and should therefore be investigated further. When we evaluated our basic SCF feature (equivalent to F1) using the same corpus data and gold standard but an older version of the RASP parser and the SCF extraction system in section 3.4, the F dropped dramatically: from 57.8 to 38.3. The relatively low performance of basic LP features in French suggests that at least some of the current errors are due to parsing. Future research should therefore investigate the source of error at different stages of processing.

In the future, it would also be interesting to investigate whether performance on French can be further enhanced by language-specific tuning (e.g. by experimenting with language specific features such as auxiliary classes).

Methodology similar to ours has yielded promising results on semantic verb classification in German (Schulte im Walde, 2006) and Japanese (Suzuki and Fukumoto, 2009). However, these studies have not focussed on Levin style classes, and have not explored cross-linguistic transfer. The works most related to ours are those of Merlo *et al.* (2002) and Ferrer (2004). Our results contrast with those of Ferrer who showed that a clustering approach does not transfer well from English to Spanish. However, her experiment used basic SCF and named entity features only, and a clustering algorithm less suitable for high dimensional data.

Like us, Merlo *et al.* (2002) created a gold standard by translating Levin classes to another language (Italian). They also applied a classification approach developed for English to Italian, and reported good overall performance using features developed for English. Although the experiment was very small in scale (involving three classes and a few features only), and although it involved a use of a supervised classification technique, the results are in agreement with our results from this larger, unsupervised experiment with French.

In their recent experiment, Falk *et al.* (2012) built on some of the work we have described in this chapter (Sun *et al.*, 2010). They made use of existing syntactic and semantic lexical resources to cluster 2183 French verbs in our gold standard classes (a superset of our gold standard). They experimented with a new clustering method and new feature sets. They obtained better result (70 F), but this result is not comparable with ours because the gold standard was not identical. Also, manually specified rather than automatically acquired features were used in the experiment. In addition, we found that there are two potential flaws with the experiment which can affect the results. [5]:

[5] The second point was confirmed with the first author via personal communication. We were not able to get a clarification regarding the first point.

1. In order to obtain the thematic grid feature from VerbNet, a classifier was trained to map French verbs to VerbNet classes. The gold standard verbs and classes were used to train the classifier (see footnote 3 on page 2 in their paper). In other words, the gold standard was used for feature extraction. This makes the clustering result higher than in fully automatic work, as the thematic grid feature is already implicitly encoded in the class label.

2. F-Measure was used to select the number of clusters for K-Means and IGNG (see page 4 in their paper). This means that the gold standard was used as help in clustering. This also makes the result unrealistically high from the perspective of automatic acquisition, as the reported best F-Measure cannot be extracted when the gold standard is unknown.

In sum, the experiments reported in this chapter further support the linguistic hypothesis that Levin-style classification can be cross-linguistically applicable or overlapping (Levin, 1993). A clustering technique such as the one presented here could be used as a helpful tool to investigate this hypothesis further, and to find out whether classifications are similar across a wider range of more diverse languages. From the NLP perspective, the fact that an unsupervised technique developed for one language can be applied to another language without substantial changes in the methodology means that automatic techniques can be used to hypothesise useful Levin-style classes in a cost-effective manner (Kipper *et al.*, 2008). This, in turn, can facilitate the creation of VerbNets for new languages.

Chapter 7

Task-based evaluation of verb classification

VerbNet has proved useful for many practical NLP tasks including automatic verb acquisition (Swift, 2005), semantic role labelling (Swier and Stevenson, 2004), robust semantic parsing (Shi and Mihalcea, 2005), word sense disambiguation (Dang, 2004), building conceptual graphs (Hensman and Dunnion, 2004), and creating a unified lexical resource for knowledge extraction (Croch and King, 2005). According to our knowledge, *automatically* acquired classification has not been evaluated in the context of an NLP task yet, although such an evaluation would be important. We apply our automatically acquired verb and noun classes (SPs) to two NLP tasks: metaphor identification and argumentative zoning. We did this work in collaboration with Ekaterina Shutova and Yufan Guo. The project plan, system design, experiment and evaluation were carried out by Ekaterina Shutova and Yufan Guo respectively. The author's contribution was to provide the lexical classifications and the related statistics for the two tasks. We summarise the resulting work in this chapter. Details of the work can be found in the following publications: Shutova *et al.* (2010); Guo *et al.* (2010, 2011b). All the examples, figures and tables in this chapter were originally authored by Ekaterina Shutova and Yufan Guo for the publications above.

7.1 Use of verb and noun classification in metaphor identification

7.1.1 Introduction

Shutova (2011) proposed an approach for automatic metaphor identification based on noun and verb classification. According to Shutova and Teufel (2010), the phenomenon of metaphor is frequent in all types of discourse. A *metaphor identification* system that is capable of distinguishing between literal and metaphorical expressions in unrestricted text would be useful for many NLP applications.

Shutova followed the hypothesis proposed by Wilks (1978): metaphor demonstrates a violation of selectional restrictions in a given context. For example:

(1) My car *drinks* gasoline. (Wilks, 1978)

The verb *drink* normally takes an *animate* subject and a *liquid* object. Therefore, *drink* taking a *car* as a subject is a violation of the normal selectional restriction, which also indicates that *drink* is used metaphorically. This approach was automated by a few systems, but they are limited in the following aspects: 1) Fass (1991)'s system overgenerates with respect to metaphor, as it detects any kind of non-literalness or anomaly in language (metaphors, metonymies and others). 2) Fass' system and Krishnakumaran and Zhu (2007)'s system are mainly based on hand-coded knowledge which affect the coverage of the systems. 3) Approaches like Gedigan *et al.* (2006) only identify metaphors for a specific domain in a specific type of discourse.

In contrast to the previous works, the scope of this experiment is the whole of the BNC (Leech, 1992) and the domain of the expressions the system can identify is unrestricted. The motivation of using the clustering methods for metaphor identification lies in the nature of metaphorical reasoning based on association. In the following examples, both of the *marriage* and *political regime* target concepts are mapped to the source domain of *mechanism*, despite of having quite different meanings.

(2) Our relationship is not really *working*.

(3) Diana and Charles did not succeed in *mending* their marriage.

(4) The *wheels* of Stalin's regime were *well oiled* and already *turning*.

Shutova's expectation is that such relatedness of mappings from distinct target concepts to the sample source concept should appear in similar lexico-syntactic environments. Thus, clustering concepts using GRs and lexical features would allow the system to capture their relatedness **by association** and acquire metaphorical expressions beyond the seed set. For example, if the sentence in (2) is in the seed set, the system should be able to identify metaphors in both (3) and (4).

In summary, the system (1) requires an initial seed set of metaphorical expressions (source–target domain mappings); (2) performs unsupervised noun clustering in order to acquire target concepts associated with the same source domain; (3) creates a source domain verb lexicon using unsupervised verb clustering; (4) identifies metaphorical expressions from BNC that describes the target domain concepts using the verbs from the source domain lexicon. The noun and verb clusters used for step 2 and 3 were supplied by the author of this thesis.

7.1.2 Data

Shutova used Shutova (2010)'s dataset as a seed set. The dataset is composed of 62 phrases that are single-word metaphors representing verb-subject and verb-object relations, where a verb is used metaphorically. The seed phrases include e.g. verb - direct object constructions: *stir excitement, reflect enthusiasm, accelerate change, grasp theory* and subject - verb constructions: *campaign surged, factor shaped.*

The system carried out the metaphor identification on the BNC that was parsed using the RASP parser of Briscoe *et al.* (2006). The GRs output of RASP for BNC created by Andersen *et al.* (2008) was used.

7.1.3 Method

The system generalizes over the metaphorical expressions in the seed set by means of unsupervised verb and noun clustering. As demonstrated in example 2, 3 and 4 in section 7.1.1, Shutova expects that the obtained clusters represent potential source and target concepts between which metaphorical associations hold. The system searches the source and target domain terms within object and subject relations from the BNC parsed by RASP. These GRs can be classified as metaphorical or non-metaphorical according to the source and target domain vocabulary in the related clusters. In addition, Shutova employs a selectional preference strength filter to remove the candidate expressions that are associated with verbs that are less likely to be used metaphorically. The hypothesis of this approach is that only the verbs that exhibiting strong SP would be prone to metaphoricity. The SP strength measure proposed by Resnik (1993) was used.

We discuss the works performed by the author of this thesis in the rest part of this section.

Verb and noun clustering

We adopt the verb clustering approach as described in chapter 3, which uses rich syntactic and semantic features extracted using a shallow parser and a clustering method suitable for the resulting high dimensional feature space. JSD was used as the similarity measure. The reason that JSD is used instead of skew divergence is described in appendix B. Recall that when we evaluated our approach on 204 verbs from 17 Levin classes in chapter 3, we obtained 80.4 F-measure (which is high in particular for an unsupervised approach). In this experiment, we apply this approach to a much larger set of 1610 verbs: all the verb forms appearing in VerbNet (Kipper *et al.*, 2006a) with the exception of highly infrequent ones. In addition, we applied the approach to noun clustering.

Feature Extraction Our verb dataset is a subset of VerbNet compiled as follows. For all the verbs in VerbNet we extracted their occurrences (up to 10,000) from the raw corpus

Source: MECHANISM
Target Cluster: consensus relation tradition partnership resistance foundation alliance friendship contact reserve unity link peace bond myth identity hierarchy relationship connection balance marriage democracy defense faith empire distinction coalition regime division
Source: STORY; JOURNEY
Target Cluster: politics practice trading reading occupation profession sport pursuit affair career thinking life
Source: LOCATION; CONTAINER
Target Cluster: lifetime quarter period century succession stage generation decade phase interval future
Source: LIVING BEING; END
Target Cluster: defeat fall death tragedy loss collapse decline disaster destruction fate

Figure 7.1: Clustered target concepts

data collected originally by Korhonen *et al.* (2006a) for construction of VALEX lexicon. Only the verbs found in this data more than 150 times were included in the experiment.

For verb clustering, we adopted the best performing features in chapter 3: automatically acquired verb SCFs parameterized by their SPs and prepositions (section 3.1.2 for details of the feature extraction).

Our noun dataset consists of 2000 most frequent nouns in the BNC. Following previous works on semantic noun classification (Pantel and Lin, 2002; Bergsma *et al.*, 2008), we used GRs as features for noun clustering. We employed all the argument heads and verb lemmas appearing in the subject, direct object and indirect object relations in the RASP-parsed BNC. The feature vectors were first constructed from the corpus counts, and subsequently normalized by the sum of the feature values before applying clustering.

Clustering Algorithm We used the SPEC method for clustering. See chapter 3 for the details of how this was done for both verbs and nouns. We experimented with different number of clusters settings (50, 100, 200, 300, 400) for both noun and verb clusterings. Shutova found that 200 is the most suitable setting for both nouns and verbs by means of qualitative analysis of clusters as representations of source and target domains. Some of the clusters obtained as a result of applying the algorithm to our noun and verb datasets are shown in Figures 7.1 and 7.2, respectively. The noun clusters represent target concepts that Shutova expects to be associated with the same source concept (some suggested source concepts are given in Figure 7.1, although the system only captures those implicitly). The verb clusters contain coherent lists of source domain vocabulary.

Source Cluster: sparkle glow widen flash flare gleam darken narrow flicker shine blaze bulge
Source Cluster: gulp drain stir empty pour sip spill swallow drink pollute seep flow drip purify ooze pump bubble splash ripple simmer boil tread
Source Cluster: polish clean scrape scrub soak
Source Cluster: kick hurl push fling throw pull drag haul
Source Cluster: rise fall shrink drop double fluctuate dwindle decline plunge decrease soar tumble surge spiral boom

Figure 7.2: Clustered verbs (source domains)

7.1.4 Evaluation and discussion

Shutova compared the system's output to that of a baseline using WordNet synsets as source and target domains. The precision of the two systems is determined with help of human judges. The coverage is calculated based on the number of metaphorical expressions found by the system.

Comparison against WordNet baseline

The baseline system employs the synonymy information from WordNet to expand on the seed set. It assumes that the source and target vocabularies are represented by the synonyms of the verbs and nouns in seed expressions. The system then searches for phrases composed of lexical items belonging to those vocabularies. The **coverage** of the two systems is compared by estimating the number of WordNet synsets in the metaphorical expressions captured by the two systems. The baseline covers only 13% of the metaphors identified using clustering. When compared to the output of the baseline, the metaphors tagged by the system represent a considerably wider range of meanings, e.g. given the seed metaphors *stir excitement, throw remark, cast doubt* the system identifies previously unseen expressions *swallow anger, hurl comment, spark enthusiasm* etc. as metaphorical.

Comparison with human judgments

Shutova used the help of annotators in order to assess the precision of metaphor identification by both systems. The annotators were presented with 78 randomly sampled sentences containing metaphorical expressions as annotated by the system and by the baseline. They were asked to decide whether the tagged expressions were metaphorical or not. The *kappa* score (Siegel and Castellan, 1988) of this annotation task is 0.63 (n=2, N=78, k=5). The precision is calculated as the percentage of metaphorical expressions that were tagged correctly out of the ones that were tagged. A tagged metaphorical expression is considered to be correct if at least three annotators agree that this is the case.

The resulting precision of the system is 0.63, and the baseline obtains 0.44. For the details on the annotation process and error analysis of the two systems, please consult Shutova (2011).

7.1.5 Conclusion

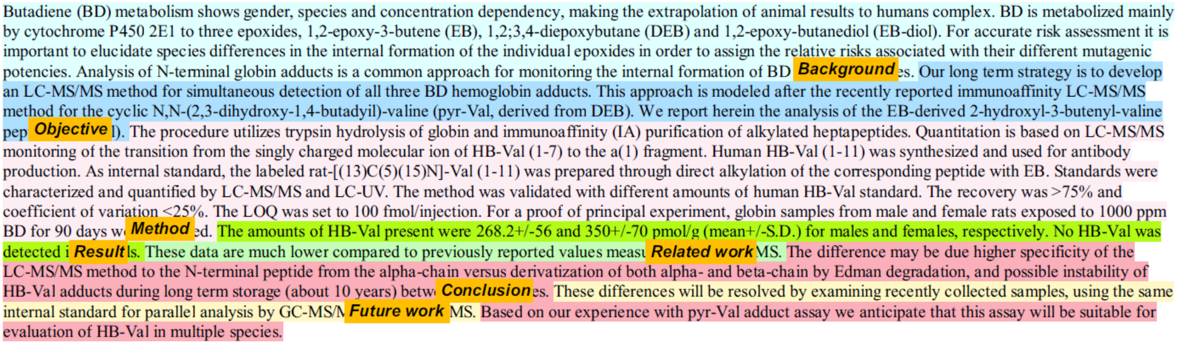
We presented Shutova (2011)'s novel metaphor identification system for unrestricted text using verb and noun clusterings. The system has a better coverage and precision than a baseline using WordNet synsets. The resulting metaphors go far beyond synonymy and generalize well over the source and target domains. In this task, we demonstrated that our automatically acquired Levin-style verb classifications can be very useful for an NLP task. To our knowledge, this is the first task-based evaluation of automatic Levin-style classification. VerbNet lacks the statistical information required by the experiment we have reported, so it could not have been used for it anyway. Since our clustering methods and features can transfer to new languages and domains, the metaphor identification system can be applied to a new task at a low cost.

7.2 Use of verb classes in argumentative zoning

7.2.1 Introduction

Argumentative zoning (AZ) is the analysis of the argumentative structure of a scientific paper. It has proved useful for many NLP tasks (Teufel and Moens, 2002; Mizuta *et al.*, 2006; Tbahriti *et al.*, 2005; Ruch *et al.*, 2007). Most approaches to AZ are fully supervised which means they are difficult to apply to new scientific domains. Also many of their features (e.g. individual verbs) tend to be sparse, so large data annotation is needed to obtain good features.

We experimented with features with more generalization power (including verb classes) and also employed weakly-supervised ML, which only relies on a small amount of labelled data. This work is joint work with Yufan Guo. The author of this thesis is responsible for providing verb classes. The work summarised here is published in two papers (Guo *et al.*, 2010, 2011b). The 2010 paper focusses on fully supervised learning, and the recent 2011 paper focusses on weakly supervised learning. We will only discuss latter work here because it aims to minimise the need for data annotation and therefore relies more heavily on less sparse features. We also focus mainly on the parts relevant to the verb class feature. See the relevant paper for further details of the experiment.



Butadiene (BD) metabolism shows gender, species and concentration dependency, making the extrapolation of animal results to humans complex. BD is metabolized mainly by cytochrome P450 2E1 to three epoxides, 1,2-epoxy-3-butene (EB), 1,2;3,4-diepoxybutane (DEB) and 1,2-epoxy-butanediol (EB-diol). For accurate risk assessment it is important to elucidate species differences in the internal formation of the individual epoxides in order to assign the relative risks associated with their different mutagenic potencies. Analysis of N-terminal globin adducts is a common approach for monitoring the internal formation of BD es. Our long term strategy is to develop an LC-MS/MS method for simultaneous detection of all three BD hemoglobin adducts. This approach is modeled after the recently reported immunoaffinity LC-MS/MS method for the cyclic N,N-(2,3-dihydroxy-1,4-butadyil)-valine (pyr-Val, derived from DEB). We report herein the analysis of the EB-derived 2-hydroxyl-3-butenyl-valine pep l). The procedure utilizes trypsin hydrolysis of globin and immunoaffinity (IA) purification of alkylated heptapeptides. Quantitation is based on LC-MS/MS monitoring of the transition from the singly charged molecular ion of HB-Val (1-7) to the a(1) fragment. Human HB-Val (1-11) was synthesized and used for antibody production. As internal standard, the labeled rat-[(13)C(5)(15)N]-Val (1-11) was prepared through direct alkylation of the corresponding peptide with EB. Standards were characterized and quantified by LC-MS/MS and LC-UV. The method was validated with different amounts of human HB-Val standard. The recovery was >75% and coefficient of variation <25%. The LOQ was set to 100 fmol/injection. For a proof of principal experiment, globin samples from male and female rats exposed to 1000 ppm BD for 90 days w ed. The amounts of HB-Val present were 268.2+/-56 and 350+/-70 pmol/g (mean+/-S.D.) for males and females, respectively. No HB-Val was detected i ls. These data are much lower compared to previously reported values measu MS. The difference may be due higher specificity of the LC-MS/MS method to the N-terminal peptide from the alpha-chain versus derivatization of both alpha- and beta-chain by Edman degradation, and possible instability of HB-Val adducts during long term storage (about 10 years) betw es. These differences will be resolved by examining recently collected samples, using the same internal standard for parallel analysis by GC-MS/N MS. Based on our experience with pyr-Val adduct assay we anticipate that this assay will be suitable for evaluation of HB-Val in multiple species.

Figure 7.3: Example of annotated abstract in Guo *et al.* (2010)'s corpus

7.2.2 Data and experiments

AZ provides a sentence based classification of scientific text into categories that capture information structure or scientific discourse of the paper. Originally developed by Simone Teufel (Teufel, 1999), we employed a version of the AZ scheme modified for biomedical papers (Mizuta *et al.*, 2006). The AZ corpus built by Guo *et al.* (2010) contains 1000 biomedical abstracts. Sentences in each abstract are annotated according to one of the seven categories of biomedical AZ appearing in abstracts: Background (BKG), Objective (OBJ), Method (METH), Result (RES), Conclusion (CON), Related work and Future work. Figure 7.3 shows an example of an annotated abstract.

To obtain the verb class features, we extracted 60 verb classes by clustering all verbs with frequency greater than 150 in Korhonen *et al.* (2008)'s biomedical dataset (see section 5.3 for details). The acquisition method is same as in section 7.1.3: the SPEC method with the SCF parameterized with preposition and SP was employed. The feature extraction method is described in section 2.3.2. The other features used for AZ include location, word, bi-gram, POS, GR, voice and verb. In table 7.1, *All* refers to all of these features, and the verb feature is compared to the verb class feature.

We experimented with fully supervised classification methods including SVM and CRF, and weakly supervised methods (active SVM, active SVM with self-training (ASSVM, transductive SVM) and semi-supervised CRF.

The results were evaluated using ACC, precision (P), recall(R) and F.

$$\text{ACC} = \frac{\text{no. of correctly classified sentences}}{\text{total no. of sentences in the corpus}}$$

$$\text{P} = \frac{\text{no. of sentences correctly identified as the class}}{\text{total no. of sentences identified as the class}}$$

$$\text{R} = \frac{\text{no. of sentences correctly identified as the class}}{\text{total no. of sentences in the class}}$$

Features	Acc	MF	BKG	OBJ	METH	RES	CON
All	0.81	0.76	0.86	0.56	0.76	0.88	0.76
¬Verb	0.81 (-0%)	0.79	0.84	0.77	0.73	0.87	0.75
¬Verb class	0.79 (-2%)	0.75	0.86	0.62	0.72	0.84	0.70

Table 7.1: Result on leaving one feature out for ASSVM. The difference of the accuracy to that of *all* is also labelled.

$$\text{F} = \frac{2 \times p \times r}{p + r}$$

The MF is the Marco average of F across five high frequency zone categories. All the results are produced using 10-fold cross validation in order to avoid the bias on the training data. In addition, only 10% of the labelled data are used as training data, and the rest are used as the unlabelled training data to the classifier.

With all the features, ASSVM produced the best result at 0.81 in ACC and 0.76 in MF. It outperforms the best fully supervised method SVM by 0.04 in ACC and 0.02 in MF. Next, we conducted an analysis to investigate the usefulness of the verb class feature. We took the best performing ASSVM method and conducted leave-one-out of the features on the 10% left out data. The results are compared to the results of using all the features (table 7.1). By excluding the verb feature, the ACC stays the same, but the MF is actually improved by 3%. However, raw verb feature was regarded as a useful feature for AZ as shown in previous experiments (Guo *et al.*, 2010; Liakata *et al.*, 2012). This interesting result shows that when combined with the verb class feature, the raw verb does not offer additional valuable information. We can also observe that ACC and MF are decreased when the verb class feature is left out. The verb class information is particularly useful for the Method, Result and Conclusion categories: a 4-6% decrease can be observed when the verb class feature is removed. In addition, the verb class feature is the third most important feature among all features. Only the location and POS feature are more important than it. It is more important than many features commonly used in AZ, e.g. words, bigrams.

7.2.3 Summary

The results reported here show that the weakly supervised classifier (ASSVM) outperforms the fully supervised classifier by making use of both labelled and unlabelled data. The verb class feature was shown to be very useful for weakly supervised learning. In contrast, the raw verb feature actually decreases the performance when the verb class feature is used. The verb class feature alleviates the data sparseness problem of the raw verb feature, which lowers the requirement of the amount of training data that need to be annotated. This makes it a particularly suitable feature for a weakly-supervised classifier when the labelled training data is small.

7.3 Release of large-scale verb classifications

To enable the use of automatically induced verb classes in further NLP tasks, we took our best performing clustering method and feature combination and created two large scale classifications: one for general English, and one for the biomedical domain.

For both clusterings, we used the SPEC method. The features we used were SCFs parameterized with prepositions and SPs. Chapter 3 provides details on the method, features and feature extraction techniques.

We constructed the general English classification from the English Gigaword corpus (Graff *et al.*, 2003). We clustered all the verbs appearing in VerbNet which occurred at least 1500 times in Gigaword. For the verbs with more than one class, we assigned each verb to the class which, according to VerbNet, corresponds to its predominant sense in WordNet. The verbs that are in the class which have less than three member verbs are discarded. The resulting gold-standard hierarchy has three levels with 170, 144 and 60 classes on each level. These verbs were clustered according to the number of clusters on each level in the gold standard. When evaluated against the three top levels of VerbNet, the results are 59.1, 54.8 and 52.3 in F respectively for the three levels. On the similar scale dataset, the following results were reported: 36.7, 37.9 and 40.0 in F in hierarchical clustering (with tree constraint) in chapter 4 and Sun and Korhonen (2011), as well as 52.8 in macro-averaged recall (48 classes) for supervised learning in Li and Brew (2008). These results are not directly comparable to each other, as the evaluation measure, features and corpora used are different.

For the biomedical classification, we employed the biomedical corpus in Korhonen *et al.* (2008). Korhonen *et al.* (2008)'s gold-standard was extended with additional verbs suggested by a linguist and domain experts. We only kept the verbs that occur at least 150 times in the corpus. These 399 verbs were clustered into three levels. We set the number of clusters to be 78, 46 and 17 respectively for each level, same as the number in the gold-standard. The results against the gold standard on the three levels are 68.8, 64.7 and 62.5 in F respectively.

We make these two resources publicly available in order to allow the researchers to use automatically acquired classes for their tasks. The classification can be tuned to different tasks or extended easily by re-running the clustering with different settings (e.g. features, corpora, frequency cut). The resources are available at:

`http://www.cl.cam.ac.uk/users/ls418/resource_release/`

The clustering code is available on request.

7.4 Summary

In this chapter, we evaluated the automatically acquired verb classes on two NLP tasks: AZ and metaphor identification. To our knowledge, these are the first task-based evaluations of automatic verb classification. In the first task, the automatically acquired verb and noun clusters enable a new approach for identifying metaphor expressions in language. The clusters were used to represent the source and target concepts. The experimental results are promising in terms of recall and precision. This is an example of a task that requires an automatically acquired rather than manually built classification because it makes use of statistical information. In the AZ task, we used our verb classes as an additional feature for AZ of biomedical abstracts. The result shows that verb classes can improve the performance over the raw verb feature. Among all features, it is the third most important feature. Automatically acquired classification is needed because no manually developed biomedical verb classification is available. Finally, using our best clustering methods, we created and evaluated two large verb classifications: one for general English and another one for the biomedical domain. We have made these classifications publicly available so that researchers wishing to use them for NLP tasks and applications can do so.

Chapter 8

Conclusions

In this chapter, we summarize the contributions of this thesis (section 8.1) and outline directions for future research (section 8.2).

8.1 Contributions of this thesis

The main contribution of this thesis is to advance the state of the art of automatic verb classification by improving its accuracy and applicability across domains and languages. We improved the accuracy of verb clustering by introducing novel methods and semantic features that improved performance. We evaluated the classification methodology on established datasets. We also applied and evaluated the methods to a different language (French) and to a different domain (biomedical). For the first time, we performed task-based evaluation of the automatic verb classification on two NLP tasks. The results show that the automatically acquired classification can be very useful for NLP applications. Our research resulted in several experimental findings and methodological proposals which we discuss in the following sections:

8.1.1 Clustering methods

We introduced two new clustering methods to the task and to the NLP field.

SPEC was previously used for verb classification by e.g. Brew and Schulte im Walde (2002). However, we used an alternative version: the MNCut (Maila and Shi, 2001) algorithm, which has a probabilistic interpretation. This algorithm is particularly useful for handling high-dimensional feature space of verb classification (Brew and Schulte im Walde, 2002). In addition, we used this algorithm to acquire SPs. We employed Zelnik-Manor and Perona (2004)'s method for automatically detecting the number of clusters. In our experiments, the method outperformed clustering methods that have been used for verb classification in previous works (K-Means and PC).

HGFC is a graph-based hierarchical clustering algorithm (Yu *et al.*, 2006). Most previous verb classification work has focused on acquiring and evaluating flat classifications. Levin's classification is not flat, but taxonomic in nature, which is practical for NLP purposes since applications may differ in terms of the granularity they require from a classification. Additionally, all the previous works using hierarchical clustering had used linkage (AGG) based hierarchical clustering. We addressed two problems of the linkage method: *error propagation* - when a verb is misclassified at a level, the error propagates to all other levels and *local pairwise merging* - only two clusters can be combined at any level. We demonstrated that HGFC can avoid both problems. We also modified HGFC so that it can be used to automatically determine the tree structure for clustering, and proposed two extensions: automatically determining the number of clusters and adding soft constraints. In the experiments we reported, HGFC greatly outperformed AGG on the all test sets, and it performed similarly with the current best flat clustering method SPEC. The constrained version of HGFC detects the missing hierarchy from the existing gold standards with high accuracy. In addition, HGFC produces a fairly accurate multi-level hierarchy, when the number of clusters and levels are detected automatically. Our qualitative evaluation showed that both constrained and unconstrained versions of HGFC are capable of learning valuable novel information not included in the gold standards.

8.1.2 Semantic features

We investigated the role of the semantic feature SP for verb classification. In previous works, SP acquired from WordNet/GermaNet offered no significant improvement over syntactic features (Schulte im Walde, 2006; Joanis, 2002). However, in manual verb classification (e.g. VerbNet), detailed verb selectional restrictions were assigned to verbs in many classes. We introduced a method for acquiring SPs from corpus data automatically. We demonstrated that SP can be very useful for verb classification. Using the SPEC method for clustering, the verb classification performance was greatly improved over the mere use of syntactic features on two established datasets and also in the biomedical domain.

8.1.3 Cross-lingual and cross-domain study

We applied our methods and features to a different language (French) and the biomedical domain in order to investigate the portability of our verb clustering approach.

Cross-linguistic study We took the SPEC verb clustering method and features (including SP) developed for English and applied it to French which had no verb classification available before the experiment was performed. Our investigation revealed similarities between the English and French classifications supporting the argument that Levin classes have a strong cross-linguistic basis. We demonstrated that both our best methods and features are transferable between the two languages.

Cross-domain study We applied our best clustering methods (SPEC and HGFC) to Korhonen *et al.* (2008)'s biomedical dataset using the same features as Korhonen *et al.*'s. The gold standard contains both general scientific and biomedical verbs. Both our methods outperformed the PC method used by Korhonen *et al.* (2008). In all the experiments, the SP feature was the best feature (as in the general domain). We demonstrated that our methods can achieve good performance in the biomedical domain without any change.

8.1.4 Task-based evaluation

We applied our verb classification to two NLP tasks. Although the manual verb classification in VerbNet has proved useful for different application tasks, automatically acquired verb classifications had not been used for any NLP task.

Metaphor identification We used the automatically acquired noun and verb clusters (obtained with SPEC) to identify metaphorical expressions in language. The clusters were used to represent the source and target concepts. We obtained promising experimental results. This is an example of a task that requires an automatically acquired rather than manually built classification because it makes use of statistical information.

Argumentative zoning We used our verb classes as an additional feature for AZ of scientific abstracts. The result shows that verb classes can improve the performance over the raw verb feature. Using the raw verb feature over the verb class feature actually decreases the performance. This demonstrates that the verb class feature provides a generalisation over the raw verb feature. Also for this task automatically obtained classification is needed because no manually developed one is available.

8.2 Directions for future research

Future work could further improve verb classification performance. We mention below some ideas for on future research. Some extensions of our existing research were already described in earlier sections:

Section 3.6 (Verb clustering using selectional preferences)

Section 4.5 (Hierarchical verb clustering using graph factorization)

Section 6.8 (Cross-linguistic potential of verb classification)

8.2.1 Diathesis alternation as a new feature for verb classification

Encouraged by the good results obtained using the SP feature, we conducted a preliminary experiment with a semantic feature important for manual classification, but previously not used in automatic verb classification: diathesis alternations (DAs).

DAs are the regular alternations of the syntactic expression of verbal arguments, sometimes accompanied by a change in meaning. For example:

- The man broke the window.
- The window broke.

In Levin's classification, a verb class is usually characterized in terms of DAs. For example, COOK verbs (e.g. bake, cook, fry, toast ...) can take DAs such as the causative alternation, middle alternation and instrument subject alternation.

There have been a few works on automatic DA detection (McCarthy and Korhonen, 1998; Lapata, 1999; McCarthy, 2000; Tsang and Stevenson, 2004), but they all rely on WordNet. There is no prior work on incorporating automatically acquired DAs to aid verb classification.

We can define two approaches to DA acquisition: *detection* and *approximation*. Detection is similar to the previous work (McCarthy and Korhonen, 1998; Lapata, 1999; McCarthy, 2000; Tsang and Stevenson, 2004): SCFs are first acquired, and then supplemented with semantic features (e.g. SPs) in order to detect whether we have a DA. We can replace the WordNet SPs with automatically acquired SPs which have the benefit that they can be ported across tasks. Specific *approximations* of features related to DA have been attempted in an earlier verb classification experiment; the *causativity* feature in Merlo and Stevenson (2001) (section 2.3.2) is one example. Since our goal is to improve verb classification, we do not have to know which sentences are alternating in order to make use of DAs. We only need to make DAs as part of the classification model. One way is to model DAs as correlations between frames. If we observe that two types of frames co-occur frequently enough, we assume a potential occurrence of DA. One drawback of the approximation approach is false positives (pairs of frames co-occur frequently, but they are not DA). In what follows, we will evaluate the potential usefulness of approximation (with false positives) for verb classification. We will discuss the two approaches in the subsequent sections.

Frame	Example sentence	Example frequency
NP+PP(on)	Jessica sprayed paint on the wall.	40
NP+PP(with)	Jessica sprayed the wall with paint.	30
PP(with)	*The wall sprayed with paint.	0
PP(on)	Jessica sprayed paint on the wall	30

Table 8.1: Example frames for verb spray

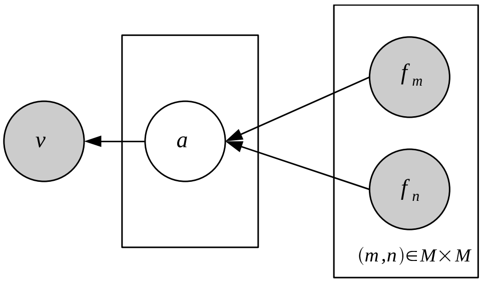


Figure 8.1: Graphical model for the joint probability of pairs of frames. v represents a verb, a represents a DA and f represents a specific frame in total of M possible frames

Diathesis alternation approximation

A DA can be approximated by a pair of frames. We define a frame as SCF parameterized for the preposition. Example frames for the verb "spray" are shown in table 8.1.

The feature value of a single frame feature is the frequency of the frame. Given two frames $f_v(i), f_v(j)$ of a verb v, they can be transformed into a feature pair $(f_v(i), f_v(j))$ as an approximation to a DA. The feature value of the DA feature $(f_v(i), f_v(j))$ is approximated by the joint probability of the pair of frames $p(f_v(i), f_v(j)|v)$, obtained by integrating all the possible DAs. In other words, the key assumption is that *the joint probability of two frames has a strong correlation with the* DAs*, if the joint probability is properly modelled by taking account of the hidden* DAs. We use the DA feature $(f_v(i), f_v(j))$ with its value $p(f_v(i), f_v(j)|v)$ as a new feature for verb clustering.

As a comparison point, we can ignore the DA and make a frame independence assumption. The joint probability is decomposed as:

$$p(f_v(i), f_v(j)|v)' \triangleq p(f_v(i)|v) \cdot p(f_v(j)|v) \tag{8.1}$$

Since SCFs are generated by the underlying meaning components (Levin and Hovav, 2006), they are dependent. The dependency of the frames is represented by a simple graphical model shown in figure 8.1. The verb v and frames f are observed, and alternation a is hidden. The aim is to approximate but not to detect a DA, so a is summed out:

$$p(f_v(i), f_v(j)|v) = \sum_a p(f_v(i), f_v(j)|a) \cdot p(a|v) \tag{8.2}$$

In order to evaluate this sum, we make a relaxation [1]: the *sum* in equation 8.1 is replaced with the maximum (*max*). This is a reasonable relaxation, as a pair of frames rarely participates in more than one type of a DA.

$$p(f_v(i), f_v(j)|v) \approx \max(p(f_v(i), f_v(j)|a) \cdot p(a|v)) \tag{8.3}$$

The second relaxation is to further relax the first relaxation by replacing the *max* with the least upper bound (*sup*): If $f_v(i)$ occurs a times, $f_v(j)$ occurs b times and $b < a$, the number of times that a DA occurs between $f_v(i)$ and $f_v(j)$ must be smaller or equal to b.

$$\begin{aligned} p(f_v(i), f_v(j)|v) &\approx \sup\{p(f_v(i), f_v(j)|a)\} \cdot \sup\{p(a|v)\} \qquad (8.4) \\ \sup\{p(f_v(i), f_v(j)|a)\} &= Z^{-1} \cdot \min(f_v(i), f_v(j)) \\ \sup\{p(a|v)\} &= 1 \\ Z &= \sum_m \sum_n \min(f_v(m), f_v(n)) \end{aligned}$$

So we end up with a simple form:

$$p(f_v(i), f_v(j)|v) \approx Z^{-1} \cdot \min(f_v(i), f_v(j)) \tag{8.5}$$

The equation is intuitive: If $f_v(i)$ occurs 40 times and $f_v(j)$ occur 30 times, the DA between $f_v(i)$ and $f_v(j) \leq 30$ times. This upper bound value is used as the feature value of the DA feature.

The original feature vector $\mathbf{f}$ of dimension M is transformed into M^2 dimensions feature vector $\tilde{\mathbf{f}}$. Table 8.2 shows the transformed feature space for the example verb *spray*. We can see that the feature space matches our expectation well: the valid DA has a value greater than 0 and the wrong DA is assigned the value of 0.

Preliminary experiment In order to evaluate the usefulness of this model, a preliminary verb clustering experiment was performed using three feature sets:

- **F1**: F-SCF+PP(B) (See table 2.4)
- **F2**: The frame pair features built from F1 with frame independence assumption (equation 8.1). This feature is not a proper DA feature as it ignores the interdependency of the frames which are produced by the underlying DA.
- **F3**: The frame pair features (DAs) built from F1 with the frame dependency assumption (equation 8.4).

[1] A relaxation is a method used in mathematical optimization for relaxing the strict requirement, by either substituting for it another easier requirement or else dropping it completely.

Frame pair	Possible alternation	Occurrence	Feature value
NP+PP(on) NP+PP(with)	Locative	30	0.158
NP+PP(on) PP(with)	Causative(with)	0	0
NP+PP(on) PP(on)	Causative(on)	30	0.158
NP+PP(with) PP(with)	?	0	0
NP+PP (with) PP(on)	?	30	0.158
PP(with) PP(on)	?	0	0
NP+PP(on) PP(on)	?	40	0.211
NP+PP(with) NP+PP(with)	?	0	0
PP(with) PP(with)	?	30	0.158
PP(on) PP(on)	?	30	0.158

Table 8.2: Example frame pair features for the verb spray

The datasets are the test sets 7-11 (3-14 classes) in Joanis *et al.* (2007), and the Sun *et al.* (2008b)'s 17 classes test set (T2) introduced in chapter 3.

The SPEC clustering algorithm was used. We used a divergence-based distributional similarity measure in the works described in chapter 3. Due to the high dimensionality of the quadratic feature space, the computational cost of the divergence similarity measure (e.g. equation 3.1) is prohibitive. So we use the Bhattacharyya kernel (Jebara and Kondor, 2003) to improve the computational efficiency.

$$w_b(v, v') = \sum_{d=1}^{D} (v_d v'_d)^{1/2} \tag{8.6}$$

The mean-filed bound of the Bhattacharyya kernel is very similar to the KL divergence kernel (Jebara *et al.*, 2004). The form of the Bhattacharyya kernel is relatively simple, which also helps the theoretical analysis in the next section.

To further reduce the computational complexity, a set of high frequency features over instances was used. For 3-6 way classifications (Joanis et al.'s test set 7-9), 50 features are used and 7-17 way classifications employ 100 features. In the next section, we will show that F3 outperforms F1 regardless of the feature number setting.

The results are shown in table 8.3. The result of F2 is lower than that of F3, and even lower than that of F1 for 3-6 way classification. This indicates that the frames independence assumption is a poor assumption. F3 yields substantially better result than F2 and F1. This experiment shows that DA features are clearly more effective than the frame features on these two datasets, even when relaxations are used.

Analysis with the Bhattacharyya kernel In this section, we examine the effect of the DA features by investigating their impact on the kernel, especially the correlation with the feature frequency.

Feature set	Joanis et al.					Sun et al.
	7	8	9	10	11	
F1	54.54	49.97	35.77	46.61	38.81	60.03
F2	50.00	49.50	32.79	54.13	40.61	64.00
F3	**56.36**	**53.79**	**52.90**	**66.32**	**50.97**	**69.62**

Table 8.3: Results when using F3 (DA), F2 (pair of independent frames) and F1 (single frame) features with Bhattacharyya kernel

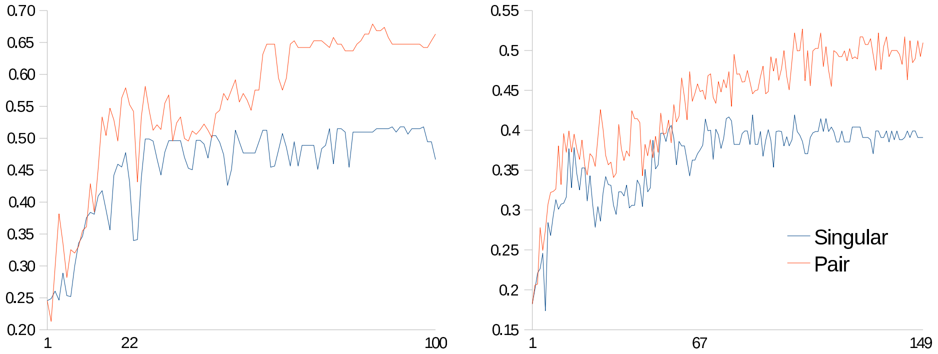


Figure 8.2: Comparison between frame features (in blue) and DA features (in red) with different feature number settings. DA features clearly outperform frame features. The left figure is the result on test set 10 (8 ways). The right figure is the result on test set 11 (14 ways). The x axis is the number of features. The y axis is the F-Measure result.

We prove that the DA feature increases the impact of the middle-range frequency frames on the Bhattacharyya kernel. The high frequency features have yet a larger impact in general than low frequency ones. This is a nice property as the high frequency features are often considered to be more reliable than the low frequency ones. The details of the mathematical proof are shown in the Appendix C.

An experiment was carried out using F1 and F3 features on Joanis *et al.* (2007)'s test set 10 and 11. The frequency ranked frames were added to the clustering one at a time, starting from the most frequent one. The results are shown in figure 8.2. F3 (in red) clearly outperforms F1 (in blue) on all the feature number settings. After adding some highly frequent frames (30 for test set 10 and 60 for test set 11), the performance for F1 is not further improved. This is in line with the mathematical proof in equation C.2: the kernel value is dominated by the top frequency frames in F1. The performance of F3, in contrast, is generally improved for almost all the frames including the mid-range frequency frames. However, the improvement becomes less significant for the frames with relatively low frequency.

In conclusion, this experiment demonstrates that the performance of using frame features is dominated by the high frequency frames, whereas the DA features reduce the dominance by enabling the mid-range frequency frames to further improve the performance.

Future work Our preliminary experiment shows, for the first time, that automatically acquired DA can provide a useful feature for verb classification. In the future, we plan to evaluate the performance of DA features in a larger scale experiment. We were not able to perform large scale experiments yet, because the dimensionality of the transformed feature space is too high (quadratic of the original feature space). An unsupervised dimensionality reduction technique (e.g. Zhao and Liu (2007)) will need to be used in order to improve the computational efficiency. Moreover, we plan to integrate the DA feature with other features (e.g. SPs) in order to further improve the accuracy of verb clustering.

Detecting diathesis alternations from selectional preferences

A few studies including McCarthy and Korhonen (1998); Lapata (1999); McCarthy (2000) have attempted DA detection using SPs. WordNet (Miller, 1995) classes have been employed as SP classes. We plan to investigate whether SPs acquisition using our new unsupervised technique (section 3.2) could be used for DA detection. Comparing to the latent variable model, this approach aims to actually detect DAs and find the participating instances instead of just approximating DAs.

We will investigate the best approaches to DA detection using automatically acquired SPs. The method needs to be general enough to cover most types of DAs and efficient enough for a large scale experiment.

One of the main problems in previous work on DA detection has been the sparse data problem in syntactic slots for which SPs are acquired. In order to reliably detect DAs, we plan to experiment with a very large corpus (e.g. Gigaword corpus (Graff *et al.*, 2003)). We will compare the resulting DAs to the DAs listed in Levin (1993). We will also evaluate the usefulness of DA features in the verb classification task.

8.2.2 Partial membership model for verb polysemy

Polysemy is a pervasive phenomenon in language, particularly among high frequency verbs. For example, in Levin's verb classification (Levin, 1993), the verb *cut* belongs to the CUT class and SPLIT class, among others. Earlier work on automatic verb classification has largely ignored polysemy by assuming a single class for each verb (e.g. Sun *et al.* (2008b); Li and Brew (2008); Joanis *et al.* (2007)). Few attempts have been made to address the problem. Multi-label classification was used for supervised adjective classification (Boleda *et al.*, 2007). For verbs, the use of an unsupervised soft clustering method

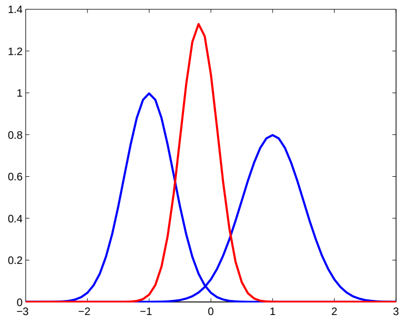


Figure 8.3: The product of two blue Gaussians with mean (-1,1) and variance (0.16, 0.25) is a new Gaussian in red with mean -0.219 and variance 0.097

was investigated by Korhonen *et al.* (2003), but the output was still based on a single class per verb.

Soft clustering has a few issues for modelling the overlapping clusters (Heller *et al.*, 2008). If two senses are modelled as two clusters, a soft clustering method would assign probabilities to the latent cluster membership, e.g. 0.5 for the sense 1 and 0.5 for sense 2. But what is the *probability* of the word having both senses? Soft clustering does not give a probabilistic interpretation of it. The traditional method is to apply a threshold value on the probability. For example, if the threshold value is set to 0.1, a word with sense probabilities (0.95, 0.05) would have sense 1 only, but word with probabilities (0.8, 0.2) would have both senses. The problem is that the threshold value is highly data dependent and we lose the probability interpretation of the sense membership. Ideally, for the word with two senses, the method should not only show the probability of having sense 1 or sense 2, but also the probability of having both senses.

A novel unsupervised clustering method can be applied for modelling the overlap between lexical categories. A few clustering algorithms (Heller *et al.*, 2008; Fu and Banerjee, 2008) have been proposed for modelling the overlapping clusters. They are all based on the *products of experts* model (Hinton, 2002). In this model, each cluster is modelled as an exponential family distribution. The product of the distributions can represent a cluster overlap. Figure 8.3 shows the product of two Gaussian distributions. We can use the single exponential family distribution for each class and use the products of distributions to model verb polysemy. The inference is achieved using the Markov Chain Monte Carlo method. We will investigate the use of efficient inference methods (e.g. Hybrid Markov chain Monte Carlo (Bonet-Cunha *et al.*, 1998) or Variational Bayesian (Attias and Ar, 1999)). We have already implemented a version of Hybrid Markov chain Monte Carlo in MATLAB which is more efficient than the Gibbs sampling. The next step would be to evaluate and tune the method on existing datasets.

8.2.3 Data requirement for reliable clustering

Frequency-based filtering is needed in almost every verb classification experiment. Two types of filtering are needed: feature filtering to remove the features which have a low number of occurrences and verb filtering to ignore highly infrequent verbs. The purpose of filtering is to improve clustering accuracy by removing noisy features and verbs which have insufficient evidence in corpus data. Table 6.3 shows the effect of verb frequency on clustering performance. Another purpose of filtering is to improve computational efficiency. However, the computational efficiency requirement depends on the complexity of the clustering algorithm and the available computational resources which differ among experiments.

One interesting question is how to set the frequency cut value. The value can be set automatically for supervised classification. In Li and Brew (2008), it is determined automatically by running cross-validation on the labelled training data with different cut settings. It can be applied to both features and verbs. Unsupervised learning cannot use this method as there is no labelled data available. Future work could carry out an empirical investigation of the optimal frequency cut for large-scale verb clustering. We can observe the relation between the clustering accuracy and the frequency cut value on both features and verbs. Experiments should be carried out across domains, languages and clustering methods. One alternative approach on filtering features is to use unsupervised feature selection methods (Zhao and Liu, 2007) to select the most useful features. Previous experiments (e.g. Sun *et al.* (2008b)) showed that the low frequency features can also benefit verb classification. It would be interesting to compare the performance of the feature selection methods to the frequency based filtering.

8.2.4 Further task and application based evaluations

We applied automatically acquired verb classifications to two NLP tasks: metaphor identification and argumentative zoning. There are many other important NLP tasks that could benefit from such classifications.

One of such task is POS tagging. According to ACL wiki[2], the fully supervised tagger performance on Penn TreeBank sections 22-24 was not improved since Toutanova *et al.* (2003). However, Manning (2011) was able to improve over Toutanova *et al.*'s results. Among the other improvements, the unknown-word error is reduced by 13% by using the word classes induced using Clark and Tim (2003)'s method. Their clustering features are shallow: co-occurrences and morphological features. Future work could investigate the use of Levin's style verb clusters and noun clusters acquired using more sophisticated or deeper features (e.g. SCFs and GRs).

[2] http://aclweb.org/aclwiki/index.php?title=POS_Tagging_%28State_of_the_art%29

We have just released an automatically induced verb classification for the biomedical domain (see section 7.3). This classification could be used to help event extraction, like in the work by Kolya *et al.* (2011), where VerbNet was used to identify the event actors. In this work, all the thematic roles and frames of the verbs in VerbNet were extracted. For a parsed sentence, the verb's argument structure was compared to its frames in VerbNet. If a match was found, the event actor corresponding to each event verb was tagged with the actor information in the appropriate slot in the sentence. Here is an example from Kolya *et al.* (2011):

Sentence: Ram killed Shyam with a knife.

Parser output: (ROOT (S (**NP** (NNP Ram)) (VP (VBD killed) (NP (NNS Shyam)) (PP (IN with) (NP (DT a) (NN knife)))) (. .)))

Acquired argument structure: [NP VP NP PP-with]

Matching frame from VerbNet:[<**NP value="Actor"**> <VERB/><NP patient><PREP value="with">]

The system was compared against a strong baseline - the noun in the subject relation in the output of the Stanford parser (De Marneffe *et al.*, 2006) (the approach is similar to that employed in Vlachos *et al.* (2009a)). The F of the baseline is 65.98, and the F of the approach using VerbNet is 67.99. Kolya *et al.*'s method could be extended with automatically acquired verb classes, which can cover verb types frequent in biomedical texts that are missing in VerbNet. In addition, the statistical information related to the frames in clustering input can make the frame matching process probabilistic (e.g. we know the probability of a verb taking a certain frame).

The work presented in this thesis has taken research on automatic verb classification much closer to the situation where it can be realistically used to benefit this and many other NLP tasks benefiting from Levin style classes that are tuned to the domain, language or task in question.

Appendix A

Features used in this thesis

Chapter	Num	Reference	Description
3	F1	F-CO	Co-occurrence (Li and Brew, 2008)
	F2	F-PP	Prepositional preference, a subset of F-LP which only include the type and frequency of prepositions in the indirect object relation
	F3	F-LP	Lexical preference, extracted as in Korhonen et al. (2008) using RASP parser
	F4	F-SCF	Basic SCF, extracted using Preiss *et al.* (2007)'s system
	F5	F-SCF+CO	The concatenation of the F-SCF and F-CO
	F6	F-SCF+TENSE(B)	F-SCF with the tense of the verb, the frequency of verbal POS tags is calculated specific to each SCF.
	F7	F-SCF+PP(B)	F-SCF with all PP frames parameterized for prepositions (Korhonen *et al.*, 2008).
	F8	F-SCF(B)	Basic SCF feature, as in Sun *et al.* (2008b), extracted from the VALEX lexicon.
	F9		F7 with F3 (subject only)
	F10		F7 with F3 (object only)
	F11		F7 with F3 (subject, object, indirect object)
	F12-14		F9-F11 with 20 clusters from 200 argument heads
	F15-17		F9-F11 with 30 clusters from 500 argument heads
4	F1	F-SCF	Basic SCF, extracted using Preiss *et al.* (2007)'s system
	F2	F-SCF+PP(B)	F-SCF with all PP frames parameterized for prepositions (Korhonen *et al.*, 2008).
	F3	F-SCF+LP(A)	F-SCF is parameterized by the F-LP in all argument slots (Korhonen *et al.*, 2008).
5	F1	F-SCF	Basic SCF, extracted using Preiss *et al.* (2007)'s system
	F2	F-SCF+PP(A)	F-SCF with two high frequency PP frames parameterized for prepositions: the PP and NP-PP frames (Korhonen *et al.*, 2008).
	F3	F-SCF+PP(B)	F-SCF with all PP frames parameterized for prepositions (Korhonen *et al.*, 2008).
	F4	F-SCF+TENSE(A)	F-SCF with the tense of the verb. The frequency of verbal POS tags is calculated over all SCFs (Korhonen *et al.*, 2008).
	F5	F-SCF+TENSE(B)	Same as above, but the frequency of verbal POS tags is calculated specific to each SCF.
	F6	F-SCF+VOICE(A)	F-SCF with the active/passive voice of the verb. The frequency of the voice is calculated over all SCFs (Korhonen *et al.*, 2008)
	F7	F-SCF+VOICE(B)	Same as above, but the frequency of voice is calculated specific to each SCF.
	F8	F-SCF+SP(A)	F-SCF is parameterized by the F-SP in all argument slots. As in Korhonen *et al.* (2008), the SPs are acquired automatically by clustering the argument head. The number of clusters was set to 10.
	F9	F-SCF+SP(B)	The number of clusters was set to 20.
	F10	F-SCF+SP(C)	The number of clusters was set to 50.
6	F1	F-SCF	Basic SCF, extracted using Preiss *et al.* (2007)'s system
	F2	F-SCF+TENSE(B)	F-SCF with the tense of the verb, the frequency of verbal POS tags is calculated specific to each SCF.
	F3		F2, with SCFs parameterized for prepositions.
	F4-F8	F-CO	COs extracted from the window of 4, 6 and 8 words, respectively. The relative word position is ignored.
	F5-F9		F4, F6 and F8 with the relative word position recorded.
	F10	F-PP	the type and frequency of preposition in the preposition relation.
	F11		the type and frequency of nouns in the subject relation.
	F12		the type and frequency of nouns in the object and indirect object relation.
	F13	F-LP	the combination of F10-F12
	F14-F16		F1-F3 parameterized for LPs.
	F17		F3 refined with SPs.

Table A.1: Summary of all the features that are used in this thesis. The references refer to the features used in previous research (table 2.4). The features without references are our new features.

Appendix B

Similarity measures used in this thesis

Name	Description and Formula
Kullback–Leibler divergence	$D_{kl}(v, v') = \sum_i^D \ln(\frac{v(i)}{v'(i)})v(i)$. The measure is not symmetric, namely $D_{kl}(v, v') \neq D_{kl}(v', v)$. An asymmetric measure cannot be used for building an undirected graph, so it cannot be used for SPEC and HGFC.
Skew divergence (Lee, 2001)	The input vector v is smoothed with v'. The level of smoothing is controlled by a whose value is set to a value close to 1. The formula is $D_{skew}(v, v') = D_{kl}(v' \| a \cdot v + (1 - a) \cdot v')$. This measure is asymmetric. In our experiments, we symmetrize the skew divergence as: $D(v, v')_{sskew} = \frac{1}{2}(D_{skew}(v, v') + D_{skew}(v', v))$. The measure was only used in experiments described in chapter 3. We employed JSD for all other experiments. When compared to JSD, this measure has an extra smoothing parameter. The value of the parameter is difficult to set automatically without labelled training data.
Jensen-Shannon divergence (Lin, 1991)	The average between v and v' is denoted as m. The formula is $D_{jsd}(v, v') = \frac{1}{2}D_{kl}(v\|m) + \frac{1}{2}D_{kl}(v'\|m)$. This measure is symmetric.
Bhattacharyya kernel (Jebara and Kondor, 2003)	Bhattacharyya kernel is a simple symmetric similarity measure. Its mean-filed bound is very similar to KL divergence kernel (Jebara *et al.*, 2004). The measure was only used for theoretical analysis and fast computation for the DA experiments. The formula is $w_b(v, v') = \sum_{i=1}^D \sqrt{(v(i)v'(i))}$.

Table B.1: A list of all the similarity measures that are used in this thesis. v and v' are two input vectors with D dimensions. The first three measures are actually distance measures. These distance measures can be converted to a similarity measure by $w(v, v') = \exp(-D(v, v'))$

Appendix C

A proof of the impact of diathesis alternation features on the kernel

Assume we have two verbs v_1 and v_2 with feature vector dimensionality D. Let v_1's features $f_{11} \ldots f_{1D}$ has rank $r_{11} \ldots r_{1D}$ according to their frequency. Suppose a low frequency feature f_{11} and a high frequency feature f_{12}, $r_{11} > r_{12}$ and denote $\phi_{12}^1 = \frac{f_{11}}{f_{12}}$. The impact of f_{ij} on the DA feature space can be written as

$$\sum_{\{m,n \in D | f_{mn} > f_{ij}\}} \min(f_{mn}, f_{ij}) = a_{ij} \times f_{ij} \ , a_{ij} = 2r_{ij} - 1 \tag{C.1}$$

Let verb v_2 has features f_{21} and f_{22}, and we assume $f_{21} = f_{22}$ for simplicity because the focus is on f_{11} and f_{12}.

According to equation 8.6, the ratio of f_{11} and f_{12}'s contribution to the Bhattacharyya kernel is:

$$\frac{\sqrt{f_{11} f_{21}}}{\sqrt{f_{12} f_{22}}} = \sqrt{\phi_{12}^1} \text{ (given } f_{21} = f_{22}) \tag{C.2}$$

This shows that ratio of the impact of the frame feature is proportional to the ratio of the frequency rank ϕ .

The ratio for the DA feature's contribution is:

$$\frac{\sqrt{(2r_{11} - 1)} \sqrt{f_{11} f_{21}}}{\sqrt{(2r_{12} - 1)} \sqrt{f_{12} f_{22}}} = \sqrt{\frac{a_{11}}{a_{12}}} \sqrt{\phi_{12}^1} \tag{C.3}$$

$\frac{a}{a'}$ is denoted as A. A is related to the feature frequency rank of the frame features. It is invariant to the underlying frequency distribution given the ordering. It *balances* the impact of feature frequency ratio ϕ: if $\phi < 1$ then $A > 1$ and vice versa. Therefore, the impact of low frequency features like f_{11} is increased. The value range of A is relatively small when compared to ϕ, so the high frequency features, in general, still have a larger impact than low frequency features. This is a nice property as the high frequency features are often considered to be more reliable than the low frequency ones.

Appendix D

The list of verbs in T1 and T3

The verb selection criteria used in T1 and T3 were originally proposed in Joanis *et al.* (2007) and Stevenson and Joanis (2003). The actual verbs are however not found in those two papers. We thus include all the verbs used in T1 and T3.

Class name	Levin class	Verbs
Putting	9.1-6	dip install lay lean lodge lower mount place put raise rest sit situate suspend tuck
Spray/Load	9.7	cram inject load pack pile plant pump scatter settle smear spray spread stick stuff wrap
Fill	9.8	bind block choke cover decorate edge endow face frame infect line pave spot staff surround
Wipe	10.4.1-2	comb erase filter flush lick pluck polish prune scour scrub shave skim strain suck wear
Steal and Remove	10.1, 10.5	capture discharge dismiss eliminate expel extract grab recover remove rescue seize separate snatch steal withdraw
Cheat	10.6	absolve acquit burgle cheat con cure defraud deprive free rob
Recipient	13.1, 13.3	allocate award extend feed give grant issue leave lend offer owe pay sell serve vote
Object drop	26.1, 26.3, 26.7	assemble build cast compose dance develop direct fix make mix perform play prepare produce write
Amuse	31.1	affect concern encourage engage impress inspire move relax satisfy threaten throw touch transport try worry
Admire	31.2	admire appreciate enjoy fancy fear hate like love miss respect support tolerate trust value worship
Light and Substance	43.1, 43.4	beam blink emanate flare flash flicker glare gleam glow leak puff radiate seep shed shine
Sound	43.2	bellow buzz clash clatter cling cry groan hiss moan murmur rattle roar scream shriek thump
Change of State	45.1-4	burst change close collapse decrease divide double expand improve increase operate sink strengthen stretch vary
Run	51.3.2	charge crawl creep drift hurry jump leap march race rush slide stumble travel walk wander

Table D.1: Verbs and Levin verb classes in T1

Class name	Levin class	Verbs
Putting	9.1-6	dip drip drop funnel lean lodge mount place position put rest scoop sit spew spill stash tuck twist wedge wind
Spray/Load	9.7	cram daub drape heap inject load pack pile plant pump smudge sow sprinkle stack stick stock strew string stuff wrap
Fill	9.8	bandage blot contaminate dam dapple deck douse drench emblazon endow face festoon garnish mask pave plug saturate soil swathe taint
Steal and Remove	10.1, 10.5	abduct confiscate delete disgorge dislodge eject evict excise excommunicate expel kidnap liberate lop ostracize pirate purloin reap rescue shoo uproot
Wipe	10.4.1-2	bail buff comb expunge filter flush hose leach mop plow pluck polish scour scratch scrub shave shear strain suck wear
Cheat	10.6	acquit bereave bilk bleed burgle cheat cleanse defraud deplete deprive dispossess divest free gull milk pardon plunder rob sap swindle
Recipient	13.1, 13.3	allocate allot assign cede concede extend grant issue leave lend loan offer owe peddle refund rent sell serve trade will
Object Drop	26.1, 26.3, 26.7	blow carve chant chisel choreograph compile direct fashion grind intone knit mix paint perform play prepare recite sculpt toss write
Amuse	31.1	antagonize bewilder dismay dumbfound humble hypnotize infuriate nauseate outrage peeve reassure repel ruffle satisfy spellbind stun tempt terrorize threaten wound
Admire	31.2	adore cherish deplore detest dislike distrust dread hate lament loathe love prize regret relish resent respect revere treasure venerate worship
Sound	43.2	beep blare buzz chime clank clink clunk creak fizzle gurgle peal ping putter rattle sputter swish tinkle toll toot whir
Change of State	45.1-4	ameliorate blunt braise coddle degrade heat heighten lessen loose narrow perk quadruple quicken scallop sink slow solidify steepen thin toughen
Run	51.3.2	bowl clamber flit inch jog limp lumber parade romp rove scamper scuttle sidle slither streak toddle tramp trek vault wade

Table D.2: Verbs and Levin verb classes in T3

Bibliography

O. Abend, R. Reichart, and A. Rappoport. A supervised algorithm for verb disambiguation into verbnet classes. In *Proceedings of COLING*, pages 9–16, Stroudsburg, PA, USA, 2008. ACL.

S. Abney and S. P. Abney. Parsing by chunks. In *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers, 1991.

S. Ananiadou and J. McNaught. *Text Mining for Biology And Biomedicine*. Artech House, Inc., Norwood, MA, USA, 2005.

O. E. Andersen, J. Nioche, E. Briscoe, and J. Carroll. The BNC parsed with RASP4UIMA. In *Proceedings of LREC*, Marrakech, Morocco, 2008.

H. Attias and L. Ar. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of Uncertainty in Artificial Intelligence*, volume 30, 1999.

A. Azran and Z. Ghahramani. A new approach to data driven clustering. In *Proceedings of ICML*, pages 57–64. ACM New York, NY, USA, 2006.

A. Azran and Z. Ghahramani. Spectral methods for automatic multiscale data clustering. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 1*, pages 190–197. IEEE Computer Society Washington, DC, USA, 2006.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet project. In *Proceedings of COLING*, pages 86–90, Montreal, 1998.

R. Basili, M. T. Pazienza, and P. Velardi. Hierarchical clustering of verbs. In *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, 1993.

N. Bassiou and C. Kotropoulos. Long distance bigram models applied to word clustering. *Pattern Recogn.*, 44:145–158, January 2011.

S. Bergsma, D. Lin, and R. Goebel. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the EMNLP*, 2008.

B. Boguraev and E. Briscoe. Large lexicons for natural language processing: utilising the grammar coding system of the *Longman Dictionary of Contemporary English*. *Computational Linguistics*, 13(4):219–240, 1987.

B. Boguraev, T. Briscoe, J. Carroll, D. Carter, and C. Grover. The derivation of a grammatically indexed lexicon from the longman dictionary of contemporary english. In *Proceedings of ACL*, pages 193–200, Stroudsburg, PA, USA, 1987. ACL.

G. Boleda, S. Schulte im Walde, and T. Badia. Modelling polysemy in adjective classes by multi-label classification. In *Proceedings of EMNLP-CoNLL*, pages 171–180, Prague, Czech Republic, June 2007. ACL.

L. Bonet-Cunha, D. Oliver, R. Redner, and A. Reynolds. A hybrid Markov chain Monte Carlo method for generating permeability fields conditioned to multiwell pressure data and prior information. *SPE Journal*, 3(3):261–271, 1998.

D. Bourigault, M.-P. Jacques, C. Fabre, C. Frérot, and S. Ozdowska. Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, 2005.

C. Brew and S. Schulte im Walde. Spectral clustering for german verbs. In *Proceedings of EMNLP*, 2002.

T. Briscoe and J. Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the fifth conference on Applied natural language processing*, pages 356–363, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

E. Briscoe, J. Carroll, and R. Watson. The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80, 2006.

C. Brockmann and M. Lapata. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of EACL*, 2003.

A. C. Browne, A. T. McCray, and S. Srinivasan. The SPECIALIST Lexicon. *National Library of Medicine Technical Reports*, 2000.

A. Browne, G. Divita, C. Lu, L. McCreedy, and D. Nace. TECHNICAL REPORT LHNCBC-TR-2003–003, Lexical Systems: A report to the Board of Scientific Counselors. *Lister Hill National Center for Biomedical Communications, National Library of Medicine*, 2003.

N. Chambers, D. Cer, T. Grenager, D. Hall, C. Kiddon, B. MacCartney, M. de Marneffe, D. Ramage, E. Yeh, and C. Manning. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 165–170. ACL, 2007.

E. Charniak. A maximum-entropy-inspired parser. In *Proceedings of NAACL*, pages 132–139, Stroudsburg, PA, USA, 2000. ACL.

J. Chen, D.-H. Ji, C. L. Tan, and Z.-Y. Niu. Unsupervised relation disambiguation using spectral clustering. In *Proceedings of COLING-ACL*, 2006.

W. Chou, R. Tsai, Y. Su, W. Ku, T. Sung, and W. Hsu. A semi-automatic method for annotating a biomedical proposition bank. In *Proceedings of the workshop on frontiers in linguistically annotated corpora 2006*, pages 5–12. ACL, 2006.

S. Clark and J. Curran. Formalism-independent parser evaluation with ccg and depbank. In *Proceedings of ACL*, pages 248–255, Prague, Czech Republic, June 2007. ACL.

A. Clark and I. Tim. Combining distributional and morphological information for part of speech induction. In *Proceedings of EACL*, pages 59–66, 2003.

W. Cochran. The comparison of percentages in matched samples. *Biometrika*, pages 256–266, 1950.

D. Croch and T. H. King. Unifying lexical resources. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany, 2005.

H. T. Dang. *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation*. PhD thesis, CIS, University of Pennsylvania, 2004.

M. De Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

C. Ding, T. Li, and M. Jordan. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):45–55, 2010.

R. Dixon. *Ergativity*. Cambridge Studies in Linguistics. Cambridge University Press, 1994.

A. Dolbey, M. Ellsworth, and J. Scheffczyk. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In *Proceedings of the "Biomedical Ontology in Action" Workshop at KR-MED*, pages 87–94, 2006.

K. Erk, A. Kowalski, S. Padó, and M. Pinkal. Towards a resource for lexical semantics: a large german corpus with extensive semantic annotation. In *Proceedings of ACL*, pages 537–544, Stroudsburg, PA, USA, 2003. ACL.

K. Erk. A simple, similarity-based model for selectional preferences. In *Proceedings of ACL*, 2007.

G. Escudero, L. Màrquez, and G. Rigau. A comparison between supervised learning algorithms for word sense disambiguation. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 31–36. ACL, 2000.

I. Falk, C. Gardent, and J.-C. Lamirel. Classifying french verbs using french and english lexical resources. In *Proceedings of ACL*. ACL, 2012.

D. Fass. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90, 1991.

C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (ISBN: 0-262-06197-X)*. MIT Press, first edition, 1998.

E. E. Ferrer. Towards a semantic classification of spanish verbs based on subcategorisation information. In *Proceedings of the ACL 2004 workshop on Student research*, Stroudsburg, PA, USA, 2004. ACL.

D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

C. Friedman, P. Kra, and A. Rzhetsky. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of Biomedical informatics*, 35(4):222–235, 2002.

Q. Fu and A. Banerjee. Multiplicative Mixture Models for Overlapping Clustering. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 791–796. IEEE Computer Society Washington, DC, USA, 2008.

M. Gedigan, J. Bryant, S. Narayanan, and B. Ciric. Catching metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York, 2006.

J. Geiß. Latent semantic sentence clustering for multi-document summarization. Technical Report UCAM-CL-TR-802, University of Cambridge, Computer Laboratory, July 2011.

A. Genkin, D. Lewis, and D. Madigan. BMR: Bayesian Multinomial Regression Software. *DIMACS, http://www. stat. rutgers. edu/madigan/BMR/, accessed on*, 14(02), 2008.

G. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970.

D. Graff, J. Kong, K. Chen, and K. Maeda. English Gigaword. *Linguistic Data Consortium, Philadelphia*, 2003.

D. Graff. North american news text corpus. *Linguistic Data Consortium*, 1995.

R. Grishman, C. Macleod, and A. Meyers. COMLEX syntax: building a computational lexicon. In *Proceedings of the International Conference on Computational Linguistics*, Kyoto, Japan, 1994.

M. Gross. *Méthodes en syntaxe*. Hermann, Paris, 1975.

Y. Guo, A. Korhonen, M. Liakata, I. S. Karolinska, L. Sun, and U. Stenius. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107, Stroudsburg, PA, USA, 2010. ACL.

Y. Guo, A. Korhonen, I. Silins, and U. Stenius. Weakly supervised learning of information structure of scientific abstracts—is it accurate enough to benefit real-world tasks in biomedicine? *Bioinformatics*, 27(22):3179–3185, 2011.

Y. Guo, A. Korhonen, and T. Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of EMNLP*, pages 273–283, Stroudsburg, PA, USA, 2011. ACL.

K. Hale and J. Keyser. A view from the middle, lexicon project work paper 10, 1987.

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

K. A. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *Proceedings of ICML*, pages 297–304. ACM, 2005.

K. Heller, S. Williamson, and Z. Ghahramani. Statistical models for partial membership. In *Proceedings of ICML*, pages 392–399. ACM New York, NY, USA, 2008.

S. Hensman and J. Dunnion. Automatically building conceptual graphs using VerbNet and WordNet. In *Proceedings of the 3rd International Symposium on Information and Communication Technologies (ISICT)*, pages 115–120, Las Vegas, NV, 2004.

G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

R. Jackendoff. *Semantic Structures*. The M.I.T. Press, Cambridge, MA, 1990.

A. Jain, M. Murty, and P. Flynn. Data clustering: a review. *ACM computing surveys*, 31(3), 1999.

T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop*, page 57. Springer, 2003.

T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *The Journal of Machine Learning Research*, 5:819–844, 2004.

E. Joanis, S. Stevenson, and D. James. A general feature space for automatic verb classification. *Natural Language Engineering*, 2007.

E. Joanis. Automatic Verb Classification Using a General Feature Space. Master's thesis, University of Toronto, 2002.

R. Jones. Semi-supervised learning on small worlds. In *Link Discovery Workshop at KDD*, 2004.

P. Kingsbury and M. Palmer. From TreeBank to PropBank. In *Proceedings of LREC-2002*, Gran Canaria, Canary Islands, Spain, 2002.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. Extensive classifications of english verbs. In *Proceedings of the 12th EURALEX International Congress*, 2006.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. Extending VerbNet with novel verb classes. In *Proceedings of LREC*, Genova, Italy, 2006.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42:21–40, 2008.

K. Kipper-Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA, June 2005.

T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

A. K. Kolya, A. Ekbal, and S. Bandyopadhyay. A hybrid approach for event extraction and event actor identification. In *Proceedings of RANLP*, pages 592–597, Hissar, Bulgaria, September 2011.

A. Korhonen, Y. Krymolowski, and Z. Marx. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of ACL*, pages 64–71, Morristown, NJ, USA, 2003. ACL.

A. Korhonen, Y. Krymolowski, and T. Briscoe. A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC*, 2006.

A. Korhonen, Y. Krymolowski, and N. Collier. Automatic classification of verbs in biomedical texts. In *Proceedings of ACL*, pages 345–352, Morristown, NJ, USA, 2006. ACL.

A. Korhonen, Y. Krymolowski, and N. Collier. The Choice of Features for Classification of Verbs in Biomedical Texts. In *Proceedings of COLING*, 2008.

A. Korhonen. *Subcategorization Acquisition*. PhD thesis, University of Cambridge, UK, 2002.

S. Krishnakumaran and X. Zhu. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Rochester, NY, 2007.

C. Kunze and L. Lemnitzer. GermaNet-representation, visualization, application. In *Proceedings of LREC*, 2002.

J.-C. Lamirel, R. Mall, P. Cuxac, and G. Safi. Variations to incremental growing neural gas algorithm based on label maximization. In *International Joint Conference on Neural Networks - IJCNN 2011*, San Jose, United States, July 2011.

M. Lapata. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of ACL*, pages 397–404. ACL Morristown, NJ, USA, 1999.

J. H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. Word sense induction for novel sense detection. In *Proceedings of EACL*, pages 591–601, 2012.

L. Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*, pages 65–72, 2001.

G. Leech. 100 million words of english: the british national corpus. *Language Research*, 28(1):1–13, 1992.

B. Levin and M. Hovav. Argument realization. *Computational Linguistics*, 32(3):447–450, 2006.

B. Levin. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London, 1993.

J. Li and C. Brew. Which Are the Best Features for Automatic Verb Classification. In *Proceedings of ACL*, 2008.

M. Liakata, S. Saha, S. Dobnik, C. Batchelor, and D. Rebholz-Schuhmann. Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Journal of Bioinformatics*, 2012.

D. Lin and P. Pantel. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360, 2001.

Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In *Proceeding of the 17th international conference on World Wide Web*, pages 685–694, New York, NY, USA, 2008. ACM.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

T. Lippincott, D. Ó Séaghdha, and A. Korhonen. Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, 12(1):212, 2011.

T. Lippincott, D. Ó Séaghdha, and A. Korhonen. Learning syntactic verb frames using graphical models. In *Proceedings of ACL*. ACL, 2012.

C. E. Lipscomb. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 2000. 88(3): 265–266.

M. Maila and J. Shi. A random walks view of spectral segmentation. In *Proceedings of AISTATS*, 2001.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

C. D. Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of CICLING*, pages 171–189, Berlin, Heidelberg, 2011. Springer-Verlag.

Y. Matsuo, T. Sakaki, K. Uchiyama, and M. Ishizuka. Graph-based word clustering using a web search engine. In *Proceedings of the EMNLP*, pages 542–550, 2006.

D. McCarthy and A. Korhonen. Detecting verbal participation in diathesis alternations. In *Proceedings of ACL*, volume 36, pages 1493–1495. ACL, 1998.

D. McCarthy. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of NAACL*, pages 256–263. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2000.

D. McCarthy. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, UK, 2001.

Q. McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947.

P. Merlo and S. Stevenson. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408, 2001.

P. Merlo, S. Stevenson, V. Tsang, and G. Allaria. A multilingual paradigm for automatic verb classification. In *Proceedings of ACL*, 2002.

C. Messiant, T. Poibeau, and A. Korhonen. Lexschem: a large subcategorization lexicon for french verbs. In *Proceedings of LREC*, 2008.

C. Messiant. ASSCI : A subcategorization frames acquisition system for French. In *Proceedings of ACL Student Research Workshop*, pages 55–60, 2008.

G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Y. Miyao, R. Sætre, K. Sagae, T. Matsuzaki, and J. Tsujii. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of ACL*, pages 46–54, Columbus, Ohio, June 2008. ACL.

Y. Mizuta, A. Korhonen, T. Mullen, and N. Collier. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6):468–487, 2006.

P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Proceedings of NIPS*, volume 16, pages 1385–1392, 2004.

R. Navigli and G. Crisafulli. Inducing word senses to improve web search result clustering. In *Proceedings of EMNLP*, pages 116–126. ACL, 2010.

S. J. Nelson, T. Powell, and B. L. Humphreys. The Unified Medical Language System (UMLS) Project. In *Encyclopedia of Library and Information Science*, pages 369–378. Marcel Dekker, 2002.

A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of NIPS*, page 849. MIT Press, 2002.

T. E. Oliphant. Python for scientific computing. *Computing in Science and Engineering*, 9:10–20, 2007.

D. Ó Séaghdha and A. Copestake. Semantic classification with distributional kernels. In *Proceedings of COLING*, 2008.

D. Ó Séaghdha and A. Korhonen. Modelling selectional preferences in a lexical hierarchy. In *Joint Conference on Lexical and Computational Semantics*, 2012.

D. Ó Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the ACL*, pages 435–444. ACL, 2010.

M. Palmer. Consistent criteria for sense distinctions. *Computers and the Humanities*, pages 217–222, 2000.

P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of SIGKDD*, pages 613–619. ACM, 2002.

T. Pham, H. Ng, and W. Lee. Word sense disambiguation with semi-supervised learning. In *Proceedings of AAAI*, volume 20, page 1093, 2005.

S. Pinker. *Learnability and Cognition: The acquisition of Argument Structure*. Cambridge, Mass.: MIT Press, 1989.

S. P. Ponzetto and R. Navigli. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of ACL*, pages 1522–1531, 2010.

J. Preiss, T. Briscoe, and A. Korhonen. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL*, volume 45, page 912, 2007.

J. Puzicha, T. Hofmann, and J. M. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 2000.

P. Resnik. *Selection and Information: A Class-based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA, 1993.

M. Rios, W. Aziz, and L. Specia. Tine: A metric to assess mt adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 116–122. ACL, 2011.

D. L. T. Rohde, L. M. Gonnerman, and D. C. Plaut. An improved method for deriving word meaning from lexical co-occurrence. *Cognitive Psychology*, 7:573–605, 2004.

T. Rose, M. Stevenson, and M. Whitehead. The Reuters Corpus Volume 1 from yesterday's news to tomorrow's language resources. In *Proceedings of LREC*, pages 29–31. ACL, 2002.

A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of EMNLP-CONLL*, 2007.

P. Ruch, C. Boyer, C. Chichester, I. Tbahriti, A. Geissbühler, P. Fabry, J. Gobeill, V. Pillet, D. Rebholz-Schuhmann, C. Lovis, and A.-L. Veuthey. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2-3):195–200, 2007.

K. Sagae and J. Tsujii. Dependency parsing and domain adaptation with lr models and parser ensembles. In *Proceedings of EMNLP-CoNLL'07 shared task*, pages 1044–1050, 2007. Prague, Czech Republic.

P. Saint-Dizier. Verb Semantic Classes Based on 'alternations' and WordNet-like criteria . In P. Saint-Dizier, editor, *Predicative Forms in Natural language and lexical Knowledge Bases* , pages 247–279. Kluwer Academic, 1998.

J. Salojärvi, K. Puolamäki, and S. Kaski. On discriminative joint density modeling. *Machine Learning: ECML 2005*, pages 341–352, 2005.

Y. Sasaki, S. Montemagni, P. Pezik, D. Rebholz-Schuhmann, J. McNaught, and S. Ananiadou. BioLexicon: A lexical resource for the biology domain. In T. Salakoski and S. Pyysalo, editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland*, pages 109–116. Turku Centre for Computer Science (TUCS), 2008.

S. Schulte im Walde and C. Brew. Inducing german semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of ACL*, pages 223–230, Morristown, NJ, USA, 2001. ACL.

S. Schulte im Walde, C. Hying, C. Scheible, and H. Schmid. Combining EM Training and the MDL Principle for an Automatic Verb Classification incorporating Selectional Preferences. In *Proceedings of ACL*, pages 496–504, 2008.

S. Schulte im Walde. Experiments on the choice of features for learning verb classes. In *Proceedings of EACL*, pages 315–322, Morristown, NJ, USA, 2003. ACL.

S. Schulte im Walde. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194, 2006.

S. Schulte im Walde. Human associations and the choice of features for semantic verb classification. *Research on Language and Computation*, 6:79–111, 2008.

D. Seung and L. Lee. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS*, volume 13, pages 556–562, 2001.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

L. Shi and R. Mihalcea. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Proceedings of CICLING*, 2005.

E. Shutova and S. Teufel. Metaphor corpus annotated for source - target domain mappings. In *Proceedings of LREC 2010*, Malta, 2010.

E. Shutova, L. Sun, and A. Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of COLING*, pages 1002–1010. ACL, 2010.

E. Shutova. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of NAACL*, Los Angeles, USA, 2010.

E. V. Shutova. Computational approaches to figurative language. Technical report, University of Cambridge, Computer Laboratory, August 2011.

S. Siegel and N. J. Castellan. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Book Company, New York, USA, 1988.

S. Siegel. *Nichtparametrische statistische Methoden*. Klotz, 5., unveränd. aufl. edition, 2001.

M. Silverstein. *Hierarchy of Features and Ergativity*. Humanities Press, 1976.

A. Søgaard. Semisupervised condensed nearest neighbor for part-of-speech tagging. In *Proceedings of ACL: short papers*, volume 2, pages 48–52, 2011.

I. Spasic, S. Ananiadou, and J. Tsujii. Masterclass: a case-based reasoning system for the classification of biomedical terms. *Journal of Bioinformatics*, 21(11):2748–2758, 2005.

S. Stevenson and E. Joanis. Semi-supervised verb class discovery using noisy features. In *Proceedings of HLT-NAACL*, pages 71–78, Morristown, NJ, USA, 2003. ACL.

L. Sun and A. Korhonen. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of EMNLP*, pages 638–647, 2009.

L. Sun and A. Korhonen. Hierarchical verb clustering using graph factorization. In *Proceedings of EMNLP*, pages 1023–1033, Edinburgh, Scotland, UK., July 2011. ACL.

L. Sun, A. Korhonen, and Y. Krymolowski. Automatic classification of english verbs using rich syntactic features. In *Proceedings of IJCNLP*, Hyderabad,India, 2008.

L. Sun, A. Korhonen, and Y. Krymolowski. Verb class discovery from rich syntactic data. *Lecture Notes in Computer Science*, 4919:16, 2008.

L. Sun, A. Korhonen, T. Poibeau, and C. Messiant. Investigating the cross-linguistic potential of verbnet: style classification. In *Proceedings of COLING*, pages 1056–1064. ACL, 2010.

Y. Suzuki and F. Fukumoto. Classifying japanese polysemous verbs based on fuzzy c-means clustering. In *Proceedings of TextGraphs-4*, pages 32–40, 2009.

R. Swier and S. Stevenson. Unsupervised semantic role labelling. In *Proceedings of EMNLP*, pages 95–102, 2004.

M. Swift. Towards automatic verb acquisition from VerbNet for spoken dialog processing. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany, 2005.

M. Tate and S. Brown. Note on the cochran q test. *Journal of the American Statistical Association*, 65(329):155–160, 1970.

I. Tbahriti, C. Chichester, F. Lisacek, and P. Ruch. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75, 2005.

Y. W. Teh, H. D. III, and D. Roy. Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, December 2002.

S. Teufel. *Argumentative Zoning: Information Extraction from Scientific Text*. PhD thesis, University of Edinburgh, 1999.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL*, pages 173–180, Stroudsburg, PA, USA, 2003. ACL.

V. Tsang and S. Stevenson. Using selectional profile distance to detect verb alternations. In *HLT/NAACL 2004 Workshop on Computational Lexical Semantics*, 2004.

T. Tuells. The derivation of a large computational lexicon for a two-level morphological analyzer for catalan from a machine-readable dictionary. 1997.

A. Ushioda. Hierarchical clustering of words. In *Proceedings of COLING*, pages 1159–1162. ACL, 1996.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

G. Vázquez, A. Fernández, I. Castellón, and M. A. Martí. Clasificación verbal: Alternancias de diátesis. In *Quaderns de Sintagma*. Universitat de Lleida, 2000.

D. Verma and M. Meila. Comparison of spectral clustering methods. *Advances in Neural Information Processing Systems (NIPS 15)*, 2003.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of ICML*, pages 1073–1080, New York, NY, USA, 2009. ACM.

A. Vlachos, P. Buttery, D. O. Séaghdha, and T. Briscoe. Biomedical event extraction without training data. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 37–40, Stroudsburg, PA, USA, 2009. ACL.

A. Vlachos, A. Korhonen, and Z. Ghahramani. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 74–82, 2009.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395 – 416, 2007.

P. Vossen. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.

J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

T. Wattarujeekrit, P. K. Shah, and N. Collier. Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5:155, 2004.

Y. Wilks. Making preferences more active. *Artificial Intelligence*, 11(3):197–223, 1978.

S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987.

Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, pages 1101–1113, 1993.

Z. Yang and E. Oja. Clustering by low-rank doubly stochastic matrix decomposition. In *Proceedings of ICML*, 2012.

K. Yu, S. Yu, and V. Tresp. Soft clustering on graphs. In *Proceedings of NIPS*, volume 18, page 1553. MIT; 1998, 2006.

B. Zapirain, E. Agirre, and L. Màrquez. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proceedings of ACL*, pages 550–558, 2008.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Proceedings of NIPS*, volume 17, page 16. MIT Press, 2004.

X. Zhang, Y. Wang, and P. Li. Word alignment based recognizing textural entailment. In *Progress in Informatics and Computing (PIC), 2010 IEEE International Conference on*, volume 1, pages 352–355. IEEE, 2010.

Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of ICML*, pages 1151–1157, New York, NY, USA, 2007. ACM.