Statistical Modeling of Multiword Expressions

A thesis presented
by

Su Nam Kim

to
The Department of Computer Science and Software Engineering
in total fulfillment of the requirements
for the degree of
Doctor of Philosophy

University of Melbourne
Melbourne, Australia
December 2008

Thesis advisor
Timothy Baldwin

Author
Su Nam Kim

Statistical Modeling of Multiword Expressions

Abstract

In natural languages, words can occur in single units called simplex words or in a group of simplex words that function as a single unit, called multiword expressions (MWEs). Although MWEs are similar to simplex words in their syntax and semantics, they pose their own sets of challenges (Sag *et al.* 2002). MWEs are arguably one of the biggest roadblocks in computational linguistics due to the bewildering range of syntactic, semantic, pragmatic and statistical idiomaticity they are associated with, and their high productivity. In addition, the large numbers in which they occur demand specialized handling. Moreover, dealing with MWEs has a broad range of applications, from syntactic disambiguation to semantic analysis in natural language processing (NLP) (Wacholder and Song 2003; Piao *et al.* 2003; Baldwin *et al.* 2004; Venkatapathy and Joshi 2006).

Our goals in this research are: to use computational techniques to shed light on the underlying linguistic processes giving rise to MWEs across constructions and languages; to generalize existing techniques by abstracting away from individual MWE types; and finally to exemplify the utility of MWE interpretation within general NLP tasks.

In this thesis, we target English MWEs due to resource availability. In particular, we focus on noun compounds (NCs) and verb-particle constructions (VPCs) due to their high productivity and frequency.

Challenges in processing noun compounds are: (1) interpreting the semantic relation (SR) that represents the underlying connection between the head noun and modifier(s); (2) resolving syntactic ambiguity in NCs comprising three or more terms;

and (3) analyzing the impact of word sense on noun compound interpretation. Our basic approach to interpreting NCs relies on the semantic similarity of the NC components using firstly a nearest-neighbor method (Chapter 5), then verb semantics based on the observation that it is often an underlying verb that relates the nouns in NCs (Chapter 6), and finally semantic variation within NC sense collocations, in combination with bootstrapping (Chapter 7).

Challenges in dealing with verb-particle constructions are: (1) identifying VPCs in raw text data (Chapter 8); and (2) modeling the semantic compositionality of VPCs (Chapter 5). We place particular focus on identifying VPCs in context, and measuring the compositionality of unseen VPCs in order to predict their meaning. Our primary approach to the identification task is to adapt localized context information derived from linguistic features of VPCs to distinguish between VPCs and simple verb-PP combinations. To measure the compositionality of VPCs, we use semantic similarity among VPCs by testing the semantic contribution of each component.

Finally, we conclude the thesis with a chapter-by-chapter summary and outline of the findings of our work, suggestions of potential NLP applications, and a presentation of further research directions (Chapter 9).

Contents

Title Page i
Abstract iii
Table of Contents v
List of Figures x
List of Tables xii
Citations to Previously Published Work xvi
Acknowledgments xviii
Dedication xix

1 Introduction 1
1.1 Research Motivation 1
1.2 Research Issues, Related Work and Focus 7
1.2.1 Identification 7
1.2.2 Extraction 9
1.2.3 Modeling/Measuring Compositionality 10
1.2.4 Semantic Classification 12
1.2.5 Semantic Interpretation 13
1.2.6 Cross-over and Cross-lingual study 15
1.3 Focus of the Thesis 16
1.3.1 Our Aims and Approaches 16
1.3.2 Scope of Research 18
1.4 Thesis Outline and Our Contribution 18

2 The Linguistics of Multiword Expressions 23
2.1 Overview of Multiword Expressions 24
2.1.1 Linguistic Properties of Multiword Expressions 24
2.1.2 Types of Multiword Expression 32
2.2 Noun Compound 34
2.3 Verb-Particle Constructions 35
2.4 Light-Verb Constructions 38
2.5 Idioms 40
2.6 Determinerless-Prepositional Phrases 41

2.7 Human annotation experiments on Multiword Expressions 43
2.8 Chapter Summary 44

3 Statistical Frameworks and Related Work 46
3.1 Introduction 46
3.2 Co-occurrence Properties 46
3.2.1 Overview of Co-occurrence Properties 46
3.2.2 Co-occurrence Properties for Extraction 48
3.2.3 Co-occurrence Properties for Compositionality 50
3.2.4 Co-occurrence Properties for Semantics 52
3.3 Substitutability 52
3.3.1 Overview of Substitutability 52
3.3.2 Substitutability for Extraction 54
3.3.3 Substitutability for Semantic Classification 55
3.4 Distributional Similarity 57
3.4.1 Overview of Distributional Similarity 57
3.4.2 Distributional Similarity for Compositionality 60
3.4.3 Distributional Similarity for Identification 61
3.5 Semantic Similarity 62
3.5.1 Overview of Semantic Similarity 62
3.5.2 Semantic Similarity for Compositionality 63
3.5.3 Semantic Similarity for Semantic Relations 64
3.6 Recovering Ellipsed Predicates 65
3.6.1 Overview of Ellipsed Predicates 65
3.6.2 Ellipsed Predicates for Semantic Relation Interpretation . . . 66
3.7 Linguistic Properties 69
3.7.1 Overview of Linguistic Properties 69
3.7.2 Linguistic Properties for Extraction 71
3.7.3 Linguistic Properties for Semantics 73
3.8 Chapter Summary 73

4 Resources 76
4.1 Corpus 76
4.1.1 British National Corpus 76
4.1.2 Wall Street Journal 77
4.1.3 Brown Corpus 77
4.2 Lexical Resources 78
4.2.1 WordNet 78
4.2.2 CoreLex 79
4.2.3 The Moby Thesaurus 80
4.3 Tools 80
4.3.1 WordNet::Similarity 80

4.3.2 TiMBL 83
4.3.3 Robust Accurate Statistical Parsing (RASP) parser 83
4.4 Chapter Summary 84

5 MWEs and Semantic Similarity 85
5.1 Noun Compound Interpretation via Constituent Similarity 85
5.1.1 Past Research on Noun Compound Interpretation 86
5.1.2 Motivation for the Constituent Similarity Method 92
5.1.3 Details of the Constituent Similarity Method 94
5.1.4 Data Collection for Constituent Similarity Method 95
5.1.5 Experiment (I): 2-term NCs 96
5.1.6 Experiment (II): 3-term NCs 98
5.1.7 Experiment (III): The Relative Contribution of the Components 99
5.1.8 Experiment (IV): Combining the Basic Method with Bootstrapping 102
5.1.9 Experiment (V): Combining the Basic Method with the k-nearest Neighbor Algorithm 105
5.1.10 Experiment (VI): Bracketing using Semantic Relations 107
5.1.11 Summary of the Constituent Similarity Method 112
5.2 Modeling the Compositionality of English Verb-Particle Constructions using Semantic Similarity 113
5.2.1 Past Research on Modeling the Compositionality of MWEs . . 113
5.2.2 Motivation for our VPC Compositionality Method 115
5.2.3 Classifier Design 117
5.2.4 Data Collection for Modeling VPC Compositionality 118
5.2.5 Evaluation of VPC Compositionality Modeling 121
5.2.6 Summary of VPC Compositionality Modeling 127
5.3 Chapter Summary 128

6 MWEs and Ellipsed Predicates 131
6.1 Motivation and Method for Interpreting Noun Compounds via Ellipsed Predicates 132
6.2 System Architecture for the Ellipsed Predicate Method 135
6.3 Data 137
6.3.1 Data Processing 137
6.3.2 Verb-mapping method 139
6.4 Experiment (I): Ellipsed Predicate Method 141
6.5 Experiment (II): Comparison with Constituent Similarity Method . . 142
6.6 Chapter Summary 145

7 MWEs and Substitutability 146

7.1 Constituent Substitution for Interpreting Noun Compounds 146

7.1.1 Motivation for Constituent Substitution 147

7.1.2 System Architecture for the Constituent Substitution Method 149

7.1.3 Data Collection 153

7.1.4 Experiment (I): Interpreting NCs by the Constituent Substitution Method 156

7.1.5 Experiment (II): Constituent Substitution Co-training Method 159

7.1.6 Experiment (III): Benchmarking NC Interpretation 162

7.1.7 Summary of the Constituent Substitution Method 167

7.2 Disambiguating Noun Compounds 169

7.2.1 Background of Word Sense Disambiguation 170

7.2.2 Motivation for Disambiguating Nouns in NCs 171

7.2.3 System Architecture to Disambiguate Word Sense 175

7.2.4 Data Collection for Disambiguating Word Sense 176

7.2.5 Experiment (I): Word Sense Disambiguation 179

7.2.6 Experiment (II): NC interpretation using WSD 181

7.2.7 Summary of Word Sense Disambiguation over NCs 183

7.3 Chapter Summary 184

8 MWEs and Linguistic Properties 187

8.1 Using Linguistic Properties to Identify Verb-Particle Constructions . 188

8.2 Past Research on Syntactic Disambiguation of Verb-Particle Constructions 189

8.3 Selectional preferences 191

8.4 Approach and Architecture for Identifying VPCs 192

8.5 Data for VPC identification 194

8.5.1 Data Classification 194

8.5.2 Data Collection 196

8.6 Evaluation 199

8.6.1 Experiment (I): Fully Resolved Pronouns and Proper Nouns . 200

8.6.2 Experiment (II): No Resolved Pronouns or Proper Nouns . . . 202

8.6.3 Experiment (III): Partially Resolved Pronouns and Proper Nouns using WordNet 204

8.6.4 Experiment (IV): 4 Word Senses vs. 1 Word Sense 205

8.7 Discussion 206

8.7.1 Performance Analysis 206

8.7.2 Performance of Resources 207

8.7.3 Reflections on Compositionality 207

8.7.4 Factors for Further Study 209

8.8 Chapter Summary 210

9 Conclusions **212**
9.1 Summary and Findings of this Thesis 212
9.1.1 Interpreting Noun Compounds 213
9.1.2 Disambiguating the Word Sense of Noun Compounds 216
9.1.3 Modeling the Compositionality of VPCs 217
9.1.4 Identifying VPCs 217
9.2 Directions for Further Research 218
9.2.1 Future Research on Noun Compounds 219
9.2.2 Future Research on Verb-Particle Constructions 222

A Abbreviations **241**

B Semantic Relations from SemEval-2007 **243**

C Results from SemEval-2007 **244**

D Results of all Experiments to Model the Compositionality of English Verb-Particle Constructions **246**

List of Figures

1.1 The semantic classification task 12
1.2 The semantic interpretation task 14

2.1 Examples of syntactic non-markedness vs. markedness 26
2.2 Examples of semantic idiomaticity 27
2.3 MWE types (Sag et al. 2002) 32

3.1 Co-occurrence-based feature representation (Uchiyama et al. 2005) . . 58
3.2 Distributional semantics of the German idiom *den loffel Abgeben* (Katz and Giesbrecht 2006) 62
3.3 Interpretation via verb semantics (Vanderwende 1994) 67

4.1 Classification of methods in WordNet::Similarity 81
4.2 RASP pipeline 83

5.1 Example of constituent similarity 92
5.2 Similarity between the i_{th} NC in the test data and j_{th} NC in the training data 94
5.3 Classifier accuracy at different α values 101
5.4 Classification accuracy for each semantic relation at different α values 102
5.5 Performance with $\theta = 0.6$ 105
5.6 Performance with $\theta = 0.7$ 105
5.7 Performance with $\theta = 0.8$ 105
5.8 Performance with $\theta = 0.9$ 105
5.9 Performance of k-NN with WUP, for simple and weighted voting . . . 106
5.10 Performance of k-NN with LCH, for simple and weighted voting . . . 106
5.11 Performance of k-NN with JCN, for simple and weighted voting . . . 107
5.12 Performance of k-NN with LIN, for simple and weighted voting 107
5.13 Performance of k-NN with VECTOR, for simple and weighted voting . 107
5.14 Performance of k-NN with LESK, for simple and weighted voting . . . 107
5.15 Overview of bracketing 108
5.16 Performance and coverage of bracketing by semantic relations 109

5.17 Verb classes in predicting the compositionality of verb-particle constructions 115
5.18 2-dimensional matrix classifier 118

6.1 An example of the NC interpretation of Vanderwende (1994) 133
6.2 Interpreting compound nominalizations (Lapata 2002) 133
6.3 Interpreting NCs via underlying predicates in our method 134
6.4 System Architecture 136
6.5 Verb-mapping method 139
6.6 Expanding verbs via synonyms 140
6.7 Performance with 57 vs. 84 seed verbs over the two data representations143

7.1 Architecture of the constituent substitution method 150
7.2 Range of similarity values over different ontological relation types . . 151
7.3 Architecture of Hybrid Method I 152
7.4 Architecture of Hybrid Method II: the constituent similarity co-training method 153
7.5 Performance across the individual SRs for the 7-way classification task 166
7.6 Positive task 169
7.7 Negative task 169
7.8 Sense restriction due to the semantics of the disambiguated element in NCs 173
7.9 Word sense of *art* in different sense collocations 178

8.1 System architecture for VPC identification 193
8.2 Data groups in terms of the POS tags from RASP 194
8.3 Senses of *apple* and *orange* 198
8.4 Error rate reduction for VPCs of varying compositionality 208

List of Tables

2.1 Examples of statistical idiomaticity ("+" = strong lexical affinity, "?" = marginal lexical affinity, "−" = negative lexical affinity) (Cruse 1986) 30
2.2 Classification of MWEs in terms of different forms of idiomaticity . . 31
2.3 Classification of the compositionality of VPCs (Bannard et al. 2003 vs. McCarthy et al. 2003) 38
2.4 A semantic classification of D-PPs 42

3.1 MWEs and Non-MWEs based on substitution 53
3.2 Examples of MWEs and anti-collocations (Pearce 2001) 55

4.1 Composition of WordNet 2.1 79

5.1 The semantic relations used in this research to interpret noun compounds (n_1 = modifier, n_2 = head noun) 87
5.2 WordNet-based similarities for the component nouns in the training and test instances 92
5.3 The effects of polysemy on the similarities between nouns in the training 93
5.4 The distribution of semantic relations in 2-term and 3-term noun compounds 97
5.5 Accuracy of NC interpretation for the different WordNet-based similarity measures over 2-term NCs 98
5.6 Accuracy of NC interpretation for the different WordNet-based similarity measures over 3-term NCs 99
5.7 Accuracy of NC interpretation for the different WordNet-based similarity measures over 3-term NCs 100
5.8 Relative contribution of head noun and modifier for different semantic relations 100
5.9 Bootstrapping with $\theta = 0.6$ 103
5.10 Bootstrapping with $\theta = 0.7$ 103
5.11 Bootstrapping with $\theta = 0.8$ 104
5.12 Bootstrapping with $\theta = 0.9$ 104
5.13 Accuracy with the k-nearest neighbor algorithm 106

5.14 Coverage and accuracy of the strictly probabilistic vs. combined methods 110
5.15 Analysis of bracketing using SRs and the probabilistic models 111
5.16 VPCs with compositionality judgements, based on the semantic contribution of the components 116
5.17 Breakdown of the compositionality data from Bannard (2006) 119
5.18 Breakdown of the compositionality data from McCarthy *et al.* (2003) 120
5.19 Semantics of particles in the data 121
5.20 Different data sets used in our experiments 122
5.21 Results of experiments with TiMBL v5.1 123
5.22 Results of experiments with Maximum Entropy 124
5.23 Number of instances with verb and/or particle semantic contribution where V is verb in isolation 125
5.24 Result of the verb and particle contribution to compositionality . . . 126
5.25 Accuracy and correlation over the compositionality task 130

6.1 Data for the ellipsed predicate interpretation method: sentences and NCs 137
6.2 Coverage of direct (D) and indirect (I) synonyms in WordNet . . . 141
6.3 Results with 17 SRs 141
6.4 Results with 19 SRs 142
6.5 Results in combination with the constituent similarity method with 17 SRs 143
6.6 Results in combination with the constituent similarity method with 19 SRs 144

7.1 Average similarity and selected threshold value for each ontological relation type 151
7.2 The set of 7 semantic relations, where N_1 is the head noun and N_2 is a modifier 155
7.3 Breakdown of the SemEval-2007 data across the two tasks and 7 SRs 155
7.4 Extracting binary NCs from higher-order NCs 156
7.5 Analysis of the performance of constituent substitution with *synonym*s across two iterations 158
7.6 Analysis of the performance of constituent substitution with *hypernym*s across 4 iterations 159
7.7 Analysis of the performance of constituent substitution with *sister word*s across 4 iterations 160
7.8 Accumulated error in Experiment I 161
7.9 Dataset sizes and quality in Experiment II 161
7.10 Results of NC interpretation 162
7.11 Results of interpretation over differing datasets for the methods of Moldovan and Kim 162

7.12 Systems tested over the SemEval-2007 data 163
7.13 Results for the 7-way classification task 164
7.14 A breakdown of results for hybrid over the 7-way classification task, across the component methods (SC = sense collocation; CS = constituent similarity; extSC = SC + SC co-training; extCS = constituent similarity + SC co-training; and CStt = the final step over both the original and expanded training data) 164
7.15 A breakdown of results for $\mathsf{CSim}_{\mathrm{CT}}$ over the 7-way classification task, across the individual iterations 165
7.16 Results for the 2-way classification task 167
7.17 A breakdown of results for hybrid over the 2-way classification task, across the component methods (SC = sense collocation; CS = constituent similarity; extSC = SC + SC co-training; extCS = constituent similarity + SC co-training; and CStt = the final step over both the original and expanded training data) 168
7.18 A breakdown of results for $\mathsf{CSim}_{\mathrm{CT}}$ over the 7-way classification task, across the individual iterations 168
7.19 Sense distribution for *art* and *day* as an NC modifier, head noun and overall in SemCor 172
7.20 Example of the substitution method for each word sense of a polysemous NC element 176
7.21 Target noun set, and the polysemy of each 177
7.22 WSD accuracy over each target noun in the modifier and head noun positions (best-performing method in each row is indicated in **boldface**; R = random baseline; F = first sense baseline; M = majority class baseline; C = supervised classifier using CoreLex; W = supervised classifier using WordNet) 181
7.23 Accuracy of interpreting SRs in NCs 182
7.24 WSD accuracy over each target noun in the modifier and head noun positions (the best-performing method in each row is indicated in **boldface**; R = random baseline; F = first sense baseline; M = majority-class baseline; $C1$ = supervised classifier with CoreLex and grammatical role; $C1$ = supervised classifier with CoreLex, grammatical role and SR; $W1$ = supervised classifier with WordNet and grammatical role; $W2$ = supervised classifier with WordNet, grammatical role and SR; SL = SenseLearner) 186

8.1 False-positive rate (FPR), false-negative rate (FNR), and inter-annotator agreement across the four groups of token instances 195
8.2 The number of VPC and Verb-PP token instances in groups A, B, and C at varying frequency cut-offs 196
8.3 The number of subject and object head nouns of different type 196

8.4 Example proper nouns and their common noun hypernyms 198
8.5 Final number of training instances from the Brown Corpus and Wall Street Journal 199
8.6 Data set sizes at different frequency cut-offs 199
8.7 Results for VPC identification only 201
8.8 Results for VPC and Verb-PP identification 201
8.9 Results for VPC and verb-PP identification without resolving pronouns and proper nouns 203
8.10 Results for VPC and verb-PP identification without resolving pronouns and proper nouns 203
8.11 Results for VPC identification only when partially resolving pronouns and proper nouns using WordNet 204
8.12 Results for VPC and verb-PP identification when partially resolving pronouns and proper nouns using WordNet 205
8.13 Results with hypernym expansion (4WS) vs. only the first sense (1WS) 205
8.14 Performance of different parsers at VPC identification 207

B.1 The set of 7 semantic relations from SemEval-2007, where N_1 is the head noun and N_2 is a modifier 243

C.1 Group A: WordNet = NO & Query = NO 244
C.2 Group B: WordNet = YES & Query = NO 245
C.3 Group C: WordNet = NO & Query = YES 245
C.4 Group D: WordNet = YES & Query = YES 245

D.1 Results of Experiment 1 (E1) with TiMBL 246
D.2 Results of Experiment 1 (E1) with Maximum Entropy 247
D.3 Results of Experiment 2 (E2) with TiMBL 248
D.4 Results of Experiment 2 (E2) with Maximum Entropy 249
D.5 Results of Experiment 3 (E3) with TiMBL 249
D.6 Results of Experiment 3 (E3) with Maximum Entropy 250
D.7 Results of Experiment 4 (E4) with TiMBL 250
D.8 Results of Experiment 4 (E4) with Maximum Entropy 251

Citations to Previously Published Work

Sections of this thesis have appeared in the following papers:

Su Nam Kim and Timothy Baldwin. 2005. Automatic Interpretation of Compound Nouns using WordNet Similarity. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP)*. Jeju island, South Korea. pp. 945–956.

Su Nam Kim and Timothy Baldwin. 2006. Automatic Identification of English Verb Particle Constructions using Linguistic Features. In *Proceedings of the 11th Conference of European Chapter of the Association for Computational Linguistics (EACL): the Third ACL-SIGSEM Workshop on Prepositions*. Trento, Italy. pp. 65–72.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting Semantic Relations in Noun Compounds via Verb Semantics. In *Proceedings of COLING/ACL*. Sydney, Australia. pp. 491–498.

Su Nam Kim and Timothy Baldwin. 2007. MELB-KB: Nominal Classification as Noun Compound Interpretation. In *Proceedings of the 4th International Workshop on Semantic Evaluation*. Prague, Czech Republic. pp. 231–236.

David Martinez, **Su Nam Kim** and Timothy Baldwin. 2007. MELB-MKB: Lexical Substitution system based on Relatives in Context. In *Proceedings of the 4th International Workshop on Semantic Evaluation*. Prague, Czech Republic. pp. 237–240.

Su Nam Kim and Timothy Baldwin. 2007. Disambiguating Noun Compounds. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI)*. Vancouver, Canada. pp. 901–906.

Su Nam Kim and Timothy Baldwin. 2007. Interpreting Noun Compounds via Bootstrapping and Sense Collocation. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*. Melbourne, Australia. pp. 129–136.

Su Nam Kim and Timothy Baldwin. 2007. Detecting the Compositionality of English Verb-Particle Constructions using Semantic Similarity. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*. Melbourne, Australia. pp. 40–48.

Su Nam Kim, Meladel Mistica and Timothy Baldwin. 2007. Extending Sense Collocations in Interpreting Noun Compounds. In *Proceedings of*

Australasian Language Technology Workshop (ALTW). Melbourne, Australia. pp. 49–56.

Su Nam Kim and Timothy Baldwin. 2008. Benchmarking Noun Compound Interpretation. In *Proceedings of the Joint Conference on Natural Language Processing (IJCNLP)*. Hyderabad, India. pp. 569–576.

Acknowledgments

My heartfelt gratitude goes to:

∞ my advisor, Dr. Timothy Baldwin, for giving me the opportunity to study under his supervision and for inspiring me with his research and patience,

∞ Dr. Steven Bird and Dr. David Martinez for valuable input on my work

∞ Dr. Barbara Di Eugenio and Dr. Amanda Stent for their support

∞ Dr. James Bailey for administrative support

∞ everyone (Edward, Jeremy, Karl, Lars, Martina, Mel, Olivia, Ondrej, Patrick, Phil, Rebecca and Trevor) in the Language Technology Group at the University of Melbourne for providing me with a friendly environment, and for great coffee and chocolate breaks,

∞ the anonymous reviewers who provided valuable comments on my work,

∞ the University of Melbourne for financial support (J.H. Mirams and Stewell research scholarships, NICTA scholarship)

∞ Bart and Two-Socks for keeping me company day by day,

∞ my families for their sincere support,

∞ and Bharanee for his endless support, help and encouragement to become a better researcher.

Dedicated to my dearest Bharaneedharan Rathnasabapathy

Chapter 1

Introduction

1.1 Research Motivation

Lexemes are the basic unit of natural (i.e. human) language. In sentences, they combine together and interact to form structures and meaning. Lexemes can occur in single units called *simplex words*, which is the smallest lexical unit that contains meaning, or as multiple simplex words that function as a single lexical unit, called *multiword expressions* (MWEs).[1] (1.1)–(1.4) show a number of MWEs (underlined) in context.[2]

(1.1) The *marketing manager* can learn how to *take advantage of* the growing database...

(1.2) Most of the time it failed to *make it out* of the pit lane...

(1.3) They were *by and large* of the type postulate...

(1.4) You should also *make a note* of the *serial number* of your *television video*...

Both simplex words and MWEs function as structural and conceptual units of language. However, MWEs often require deeper syntactic and semantic reasoning

[1] The definition of MWEs here is deliberately brief. See Section 2.1 for a complete definition of MWEs.

[2] All examples are taken from the British National Corpus.

due to subtle interactions with the syntax and semantics of their component simplex words, or alternatively behavior which is completely at odds with their parts.[3] In the following examples, the relationship between the MWEs and their component simplex words is relatively transparent. Note that the MWEs are indicated by *italic-underlining* and the remaining terms are all simplex words.

(1.5) Kim bought fresh *kidney beans* at the market.

(1.6) He immediately *got on* the bus.

(1.7) Everyone *makes mistakes*.

(1.8) The *bus driver* accidentally hit the *garbage bin*.

(1.9) Everyday, Kim goes to the office *by bus*.

In (1.5)–(1.9), the MWEs are relatively easy to detect as their components occur continuously. The semantics of the MWEs in these examples is also predictable. The meaning of *bus driver* as "one who drives a bus" is easily accessible despite *bus* having meanings including "an electrical conductor that makes a common connection between several circuits" and "a car that is old and unreliable", and *driver* having meanings including "a golf club (a wood) with a near vertical face that is used for hitting long shots from the tee" and "a program that determines how a computer will communicate with a peripheral device".[4] The process for disambiguating the semantics in context here is identical to that for determining the word sense of *hit*, e.g. from among its many senses, based on analysis of the combinatoric interaction between possible word senses of the lexemes in the sentence (Agirre and Edmonds 2006). We can also predict the meaning of *by* in *by bus* as "manner/method", and notice that the MWE is syntactically marked in that *bus* is ordinarily a countable noun and thus requires a determiner when used with singular number.

[3] In this thesis, we refer that components are simplex words making up of a MWE. We use the terms, *components* and *parts* to refer the simplex words in MWEs.

[4] Glosses taken from WordNet 3.0.

However, while at this simplistic level MWEs are similar to simplex words in terms of their function within a sentence, they pose bigger challenges due to their syntactically and semantically unexpected behavior (Sag *et al.* 2002). (1.10)–(1.14) show more complicated MWEs where knowledge of the components alone is insufficient to predict the observed linguistic behavior. Once again, the MWEs are *italic-underlined*.

(1.10) The subject is about *language learning system design*.

(1.11) Kim *took* her pen *out*.

(1.12) She likes to *take a* long *bath* for relaxation after exams.

(1.13) He will inherit when his grandfather *kicks the bucket*.

(1.14) The survey shows that *by and large* people skip breakfast.

(1.10)–(1.12) are MWE examples which are hard to recognize as a single unit due to their length or the fact that they are discontinuous. For example, although *take out* is an MWE, it is not immediately apparent that (1.11) includes a token instance of it since *out* is separated from the verb *take*. Also, due to the internal modification by *long*, *take a bath* is not easily recognizable as a unit, or analogously, it is not immediately apparent that *long* is **not** a component of the MWE. In addition, MWEs are often confused with non-MWEs, e.g. the MWE vs. non-MWE usages of *put on* in *put the coat on* vs. *put the coat on the table*, respectively. As a result of such variations in the context of usage of MWEs, it is sometimes difficult to distinguish MWEs from compositional usages of the individual simplex words that they are formed from. Though often understated, understanding and processing language is overwhelmingly difficult without the means to syntactically recognize MWEs.

MWEs are problematic semantically as well. The meaning(s) of an MWE cannot always be directly predicted from its component words. The contribution of the MWE components to its semantics can vary widely from no single word contributing obviously (e.g. *kick the bucket*), to a single component making the most significant contribution (e.g. *finish up*), to all words in an MWE contributing equally (e.g. *bus driver*). (1.13) and (1.14) illustrate the difficulty in predicting the meaning of MWEs from their components. The meaning of *kick the bucket* as an MWE is “pass from physical life and lose all bodily attributes and functions necessary to sustain life”. However, unfortunately, neither *kick* nor *bucket* contains this meaning. Hence, estimating the exact meaning of *kick the bucket* from its parts is futile.[5] It is also impossible to estimate the meaning of *by and large* as “mostly” from the components *by* and *large*. Hence, semantically, (some) MWEs need a different treatment to simplex words.

As observed above, MWEs pose significant challenges to syntactic and semantic processing. Below, we outline other significant aspects of MWEs that make them significant to natural language processing (i.e. NLP).

The number of MWEs is estimated to be of the same order of magnitude as the number of simplex words in a speaker’s lexicon (Wood 1964; Gates 1988; Jackendoff 1997; Peabody 1981; Side 1990; Tschichold 1998).

To add to this, new (types of) MWE are continuously created as languages evolve (e.g. *shock and awe*, *cell phone*, *ring tone*) (Gates 1988; Tschichold 1998; Dias *et al.* 1999; Stevenson *et al.* 2004).

MWEs are also broadly used to enhance fluency and understandability, or mark the register/genre of language use (Fillmore *et al.* 1988; Liberman and Sproat 1992; Nunberg *et al.* 1994; Dirven 2001). For example, MWEs can make language more or less formal. *Piss off* in *It pisses me off when people do that* is an informal variant of *annoy*.

Regionally, MWEs vary considerably. For example, *take away* and *take out* have an identical meaning in the context of fast food outlets, but the former is the preferred expression in Australian English, while the latter is the preferred expression in American English. Another example is *mail box* and *post box* in the context of a postal service, where the former is the preferred form in American English and the latter is the preferred form in Australian English.

MWEs can also be used to represent information concisely (Levi 1978). For example, *winter school* is a compact way of expressing “a school which is held in the

[5] Except, possibly, via historical linguistics and analysis of antiquated usages of *bucket*.

winter".

MWEs can also lend nuance/emphasis to language (Bolinger 1976b; Brinton 1985; Side 1990; Ramchand and Svenonius 2002; Butt 2003; Fazly *et al.* 2005; Cook and Stevenson 2006). For example, *up* in *finish up the food* adds the meaning of "completion". That is, *finish up* has the meaning of *finish*, but it also contains the entailment that the *food* is completely consumed and emphasises the completeness of the eating action.

There is a modest body of research on modeling MWEs which has been integrated into NLP applications, e.g. for the purposes of fluency, robustness or better understanding of natural language. Understanding MWEs has broad utility in tasks ranging from syntactic disambiguation to conceptual (semantic) comprehension. Explicit lexicalized MWE data helps simplify the syntactic structure of sentences that include MWEs, and conversely, a lack of MWE lexical items in a precision grammar is a significant source of parse errors (Baldwin *et al.* 2004). Additionally, it has been shown that accurate recognition of MWEs influences the accuracy of semantic tagging (Piao *et al.* 2003), and word alignment in machine translation (MT) can be improved through a specific handling of the syntax and semantics of MWEs (Venkatapathy and Joshi 2006).

However, the varied nature of MWEs—i.e. syntactic, semantic and pragmatic idiomaticity, syntactic and semantic flexibility, and productivity—makes it difficult to harness them in NLP applications. Fortunately, there are classes of MWEs which have internal consistency to varying degrees within a given language, such as **noun compounds** (e.g. *apple pie*, *computer science*) and **idioms** (e.g. *in one's shoes*, *spill the beans*). We return to detail the properties and types of MWE in Section 2.

The complexities of MWEs have been recognized from linguistic (Bolinger 1976a; Downing 1977; Levi 1978; Bauer 1983; Nunberg *et al.* 1994; Fillmore *et al.* 1988; Sag *et al.* 2002; Lohse *et al.* 2004), cognitive (Side 1990; Dras and Johnson 1996; Jackendoff 1997; Gries 1999; Lynott and Keane 2004) and statistical (Lin 1999; Calzolari *et al.* 2002; Baldwin *et al.* 2003a; McCarthy *et al.* 2003; Stevenson *et al.* 2004; Van Der Beek 2005; Kim and Baldwin 2006b) perspectives. In this thesis, we draw together linguistic, cognitive and statistical sources in focusing on the syntactic and semantic complexities of MWEs.

Syntactically, one of the major issues with MWEs is recognition, due to idiomatic and syntactically-flexible expressions. MWEs are often found in the form of semi- or non-fixed expressions. The components often inflect for number or tense (e.g. *family cars*, *The plane has taken off.*). The occurrence of the components also vary with context. For example, modifiers can internally modify the components of MWEs (e.g. *make a big mistake*, *Unfair advantage was taken of him*).

Semantically also, MWEs can cause difficulties for comprehension. MWEs can be semantically idiomatic, i.e. the meaning can be explicitly or implicitly derived from the components of MWEs or be completely unrelated to the semantics of the parts (e.g. see (1.13) and (1.14)). It is also relatively common for the components of MWEs to combine compositionally to form competing analyses. For example, *a piece of cake* can be an MWE with meaning "any undertaking that is easy to do", or alternatively it can be a simple compositional expression referring to a portion of cake. Moreover, MWEs are highly productive, and their components are often used to generate novel MWEs. The verb *take*, for example, combines with a number of prepositions to form verb particle constructions including *take away*, *take off* and *take up*, each of which has distinctive semantics.

To add to these difficulties, MWEs occur in a bewildering array of syntactic and semantic types which are interrelated to varying degrees, such that neither is it possible to come up with a genuinely general-purpose analysis of all MWEs, nor is it adequate to try to document each individual MWE type independently. For example, while syntactically identifying instances of noun compounds such as *paper submission* and *chocolate bar* is relatively easy, it is much harder with other types of MWEs such as *in one's shoes* and *break the ice*. Semantically, predicting the meaning of MWEs is relatively easy with some types of MWEs such as *take a walk* and *make a note (of)*, whereas with other MWEs such as *make out* and *kick the bucket* it is considerably more difficult.

In sum, MWEs pose significant challenges for NLP, and developing a framework for modeling MWEs syntactically and semantically is both vital to the furtherance of NLP and a daunting task.

1.2 Research Issues, Related Work and Focus

Two major computational tasks relating to MWEs are: (i) syntactically identifying and extracting MWEs from corpus data, and (ii) semantically measuring the compositionality of MWEs, analyzing the semantics of MWEs, and interpreting MWEs based on their components. However, depending on the type of MWE, the relative import of these syntactic and semantic tasks varies. For example, with noun compounds, the identification and extraction (i.e. syntactic) tasks are relatively trivial, whereas interpretation is considerably more difficult.

First, prior to detailing the computational tasks relating to MWEs, let us briefly define a number of MWE types which will recur in later discussions. For full details of the MWE types, see Section 2. A *noun compound* (NC, e.g. *golf club* or *paper submission*) is an $\bar{\mathrm{N}}$ made up of two or more nouns. A *verb-particle construction* (VPC, e.g. *hand over* or *battle on*) is a verbal MWE made up of a verb and obligatory particle(s). A *light-verb construction* (LVC, e.g. *take a walk* or *make a mistake*) is a verbal MWE made up of a verb and (usually indefinite singular) object NP, where the verb has bleached semantics and the noun complement to a large degree determines the semantics of the MWE. A *determinerless prepositional phrase* (D-PP, e.g. *at school* or *on air*) is an adverbial MWE made up of a preposition and a singular noun without a determiner. Finally, an *idiom* (e.g. *kick the bucket* or *take a turn for the worse*) is an amalgam of words in a construction other than those explicitly identified above, which has different semantics to that of the combination of the individual components (Potter *et al.* 2000).

In the following sections, we discuss the primary research issues relating to MWEs, and prior work done in each area. In doing do, we offer our perspective on why these issues (continue to) pose a challenge for NLP.

1.2.1 Identification

Identification is the task of determining the existence and extent of MWEs in context. That is, the purpose of the task is to determine combinations of multiple simplex words which give rise to an MWE in corpus text. As a result, the task is at the token (instance) level, that is we may identify 50 distinct occurrences of *pick up* in a given corpus. To give an example of an identification task, given the corpus fragment in (1.15) (taken from "The Frog Prince", a children's story), we might identify the MWEs identified in (1.16).

(1.15) One fine evening a young princess put on her bonnet and clogs, and went out to take a walk by herself in a wood; ... she ran to pick it up; ...

(1.16) One fine evening a young princess put on her bonnet and clogs, and went out to take a walk by herself in a wood; ... she ran to pick it up; ...

In (1.16), the indicated occurrences of *put on*, *go out*, *take a walk* and *pick up* are identified.

Note that with LVCs and idioms, there is often ambiguity with simple compositional word combinations:

(1.17) (**LVC**) Kim took a walk.

(1.18) (**non-LVC**) Kim took a dish.

In (1.17), *take a walk* is a LVC whereas in (1.18) *take a dish* is used literally (as in there is a literal *take* event) and thus not a LVC. Due to the interaction of the component words with surrounding words and the potential for ambiguity in most cases, identifying MWEs in context is a crucial part of NLP.

MWE identification has a relatively rich research tradition for English verb-particle constructions, light-verb constructions and idioms (Dras and Johnson 1996; Li *et al.* 2003; Thanopoulos *et al.* 2003; Van Der Beek 2005; Kim and Baldwin 2006a; Villada Moiron 2005; Kan and Cui 2006; Cook and Stevenson 2007). These types of MWEs are often ambiguous with a literal usage. For example, *kick the bucket* in *Last night, grandfather kicked the bucket* is used as an idiom (meaning "pass from physical life and lose all bodily attributes and functions necessary to sustain life"), while that in *Kim accidentally kicked the bucket and fell over* has nothing to do with death (i.e.

it is used literally). Here, idiom identification is clearly a semantic task and the two competing analyses have identical syntax. In cases such as *Kim carried on Tuesday night*, however, we must weigh up both syntactic and semantic considerations in determining whether *on* is a particle (leading to the interpretation of Tuesday night being the occasion that Kim carried on) or a (transitive) preposition (leading to the interpretation of Kim carrying [as distinct from loading, e.g.] on Tuesday night).

In research to date, good results have been achieved for particular MWEs, especially VPCs. However, proposed methods have tended to rely heavily on existing resources such as parsers (Kim and Baldwin 2006a) and handcrafted lexical resources (Li *et al.* 2003). Moreover, the proposed methods are generally tuned to particular MWE types. Hence, there is a need to develop methods that are more generally applicable across a range of MWEs.

1.2.2 Extraction

MWE extraction is a type-level task, wherein the MWE lexical items attested in a predetermined corpus are extracted out into a lexicon or other lexical listing. For example, with a given verb *take* and preposition *off*, we wish to know whether the two words combine to together to form a VPC (i.e. *take off*) in a given corpus. This contrasts with MWE identification, where the focus is on individual token instances of MWEs, although obviously extraction can be seen to be a natural consequence of identification (in compiling out the list of those attested MWEs). The underlying assumption in MWE extraction is that there is evidence in the given corpus for each extracted MWE to form an MWE in some context, without making any claims about whether there also exist simple compositional combinations of those same words. The motivation for MWE extraction is generally lexicon development or expansion, e.g. in recognizing newly-formed MWEs (e.g. *ring tone* or *shock and awe*) or domain-specific MWEs (e.g. *bus speed* or *boot up* in an IT domain).

In general, MWE extraction pulls MWEs out of context as standalone lexical items, although this generally involves analysis of the context of occurrence of a given combination of words. However, as stated above, extraction often takes advantage of the results of MWE identification. For example, Baldwin (2005a) extracted English VPCs based on identifying VPC candidates using resources including a parser and chunker.

Extracting MWEs is relevant to any lexically-driven application, such as grammar development or information extraction. In addition, it is particularly important for productive MWEs or domains that have distinctive MWE content. MWE extraction is difficult for many of the same reasons as MWE identification, namely syntactic flexibility and ambiguity.

The bulk of research on MWE extraction has focused on extracting English verb-particle constructions, light-verb constructions and idioms (Baldwin and Villavicencio 2002; Piao *et al.* 2003; Villavicencio *et al.* 2004; Widdows and Dorow 2005; Baldwin 2005a; Baldwin *et al.* 2006; Van de Curys and Villada Moiron 2007). Despite a healthy body of research on MWE extraction, however, the results have not been as compelling as for MWE identification. Baldwin (2005a) achieved high accuracy on an English VPC extraction task, whereas others such as verb-noun pair extraction (Venkatapathy and Joshi 2005; Fazly and Stevenson 2007) still have considerable room for improvement. Part of the complexity here is that the target lexical resource for the MWE extraction often introduces its own constraints or requirements for extra lexical properties (e.g. valence in the case of Baldwin (2005a)).

1.2.3 Modeling/Measuring Compositionality

In the context of MWEs, compositionality denotes the degree to which the properties of the MWE are inherited directly from those of the components. While there are various definitions of ***compositionality***, for the purposes of this thesis, we focus specifically on semantic compositionally and consider compositional MWEs to be those where the meaning of the MWE is fully or largely derived from the semantics of its components. Conversely, with non-compositional MWEs there is a marked difference in the semantics of the MWE vs. the semantics of the components. For example, with the two VPCs *spill down* and *conk out*, we would claim that *spill down* is compositional whereas *conk out* is non-compositional. That is, *spill* and *down* determine the full semantics of *spill down*, while *conk* has little or no bearing on the semantics of *conk out* and *out* contributes aspectually rather than literally. While we will tend to refer to compositionality as a binary distinction, in practice there is a cline of compositionality from complete compositionality to complete non-compositionality.

Modeling/measuring the compositionality of MWEs is the task of predicting the semantic association between an MWE and its components under the assumption that, to a certain degree, the meanings of MWEs and the components can be semantically resolved using WordNet or other semantic classes. We consider compositionality modeling to be a type-level task and to be invariant across individual senses (i.e. meaning) of the MWE. This is clearly an oversimplification and there are certainly cases of different senses having different degrees of compositionality, however we claim that such cases are rare and that the assumption of monosemy makes the task considerably more tractable. Hence, the task aim is to find whether the components of a given MWE semantically contribute to the semantics of the MWE, and if so, how much. The task of modeling compositionality, i.e. whether the components contribute to the meaning of the MWE, is a binary decision. Measuring compositionality, on the other hand, is a more semantically intensive task, where we not only predict whether a MWE is compositional, but we also estimate the degree of compositionality.

Studying compositionality has its own benefits. It provides information for improving output quality in NLP applications such as machine translation and text generation (Nunberg *et al.* 1994; Sag *et al.* 2002; Venkatapathy and Joshi 2006). It is also a prerequisite task for semantic interpretation over compositional MWEs (see Section 1.2.5 for details). Previous research on modeling/measuring the compositionality of MWEs has primarily focused on English noun compounds and verb-particle constructions (Schone and Jurafsky 2001; Baldwin *et al.* 2003a; Bannard *et al.* 2003; McCarthy *et al.* 2003; Stevenson *et al.* 2004; Venkatapathy and Joshi 2005; Piao *et al.* 2006; Kim and Baldwin 2007a).

Recently, MWE compositionality has been studied not only to detect or measure the degree of the compositionality, but also to utilize this in NLP applications. Venkatapathy and Joshi (2006) successfully showed the utility of MWE compositionality in a word alignment task between English and Hindi. However, since the task

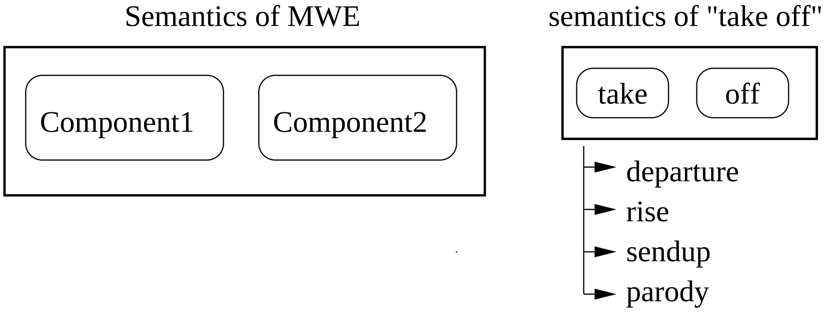

Figure 1.1: The semantic classification task

setup was supervised, large amounts of training data were necessary. There is a gap in the research literature on measuring the degree of MWE compositionality, and also on the utility of compositionality in NLP applications.

1.2.4 Semantic Classification

Semantic classification is the task of specifying the semantics of MWEs based on a generalised semantic inventory (compatible with both simplex words and MWEs). It tends to presuppose the ability to classify the (degree of) compositionality of MWEs and apply only to compositional MWEs. That is, the task focus of MWE semantic classification is to specify the meanings of MWEs according to predefined semantic categories such as WordNet. Figure 1.1 illustrates an example task in the context of VPCs.

In Figure 1.1, the target is to determine the semantics of a given MWE. Often the meaning of the components is employed to specify the semantics of the whole. Hence, compositionality is a very useful clue in estimating the meaning of compositional MWEs. In our example, the target of the task is to determine the different senses of *take off* (i.e. “departure, rise, send up, parody”). This can be performed based on individual analysis of *take* and *off* to some degree. WordNet is commonly used as a sense inventory for semantic classification tasks, although there are instances of user-defined sense inventories (e.g. *particle semantics* in Bannard (2003) and Cook and Stevenson (2006)).

Semantic classification in the context of MWEs is non-trivial due to the varying degrees of opacity in MWEs. The contribution of the individual components can vary (e.g. *eat up* and *start over*, where the verb is the primary determinant of the semantics). Sometimes none of the parts contribute to the semantics of the MWE (i.e. in fully non-compositional VPCs such as *make out*).

Prior work related to the semantic classification of MWEs has been undertaken from both the linguistic and computational perspectives (Fraser 1976; Bame 1999; Gries 1999; Bannard 2003; O'Hara and Wiebe 2003; Patrick and Fletcher 2004; Villavicencio *et al.* 2004; Villavicencio 2005; Cook and Stevenson 2006).

Most of the research on the semantic classification of MWEs has focused on English VPCs. The relatedness between semantic classification and measuring the compositionality of MWEs is not well understood, warranting further study.

1.2.5 Semantic Interpretation

As MWEs are made up of two or more simplex words, syntactic and semantic associations arise between the components. The semantic interpretation of MWEs is the task to determining the semantic relation between the components, in the form of a relation set which is specific to an MWE construction type. Note that the semantic interpretation, once again, relates closely to compositionality, in that compositionality is a claim on whether the semantic association between the components is transparent or not, whereas semantic interpretation seeks to unearth a precise description of the semantic relation between those components. For example, the knowledge that *bus driver* is fully compositional provides us the means to infer the semantics of the components, but semantic interpretation seeks to specify exactly how *bus* and *driver* relate to each other, e.g. in predicting that the *driver* is the agent of control of the *bus*. If we knew that *bus driver* were non-compositional, however, we would know not to attempt to semantically interpret it based on the components. In this sense, modeling/measuring compositionality is a prerequisite for semantic interpretation.

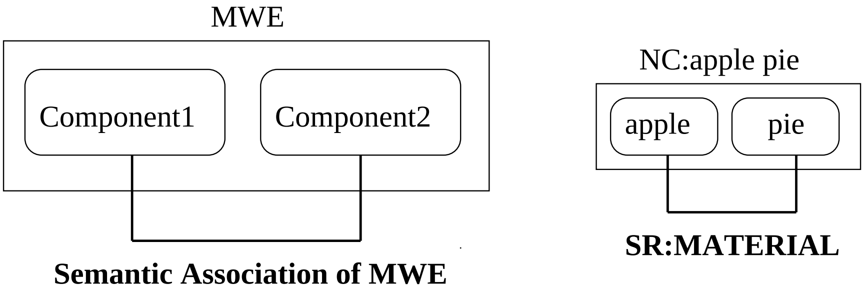

Figure 1.2: The semantic interpretation task

Figure 1.2 depicts the task of semantic interpretation with an example.

In Figure 1.2, the target is to interpret the semantic relation between the components. For example, *apple pie* can be interpreted as "pie made from apple". The semantic relation between *apple* and *pie* is specified as MAKE, where the head noun is made from the modifier.[6]

Semantic relations (or associations) are most commonly used to interpret noun compounds and determinerless prepositional phrases. The semantic relation used to interpret a given MWE varies with the components. For example, the semantic relation in *morning juice* is TIME ("juice in the morning") whereas that in *orange juice* is MAKE ("juice made from orange(s)"). Another example with D-PPs is *by car/bus/plane..*, where a mode of transportation combined with the method/manner preposition *by* leads to the semantic relation MANNER, whereas other nouns such as *day* lead to specific TEMPORAL interpretations.

The majority of past research on semantic interpretation has focused on interpreting noun compounds (Vanderwende 1994; Copestake and Lascarides 1997; Lapata 2002; Moldovan *et al.* 2004; Kim and Baldwin 2005; Kim and Baldwin 2006b; Nastase *et al.* 2006; Girju 2007; Ó Séaghdha and Copestake 2007) and D-PPs (Baldwin *et al.* 2003b; Van Der Beek 2005; Baldwin *et al.* 2006). This research, particularly that on NC interpretation, has been suggested to be relevant for QA and IR (Moldovan

[6] The semantic relations used in this thesis are from Barker and Szpakowicz (1998)

et al. 2004), although there is no definitive empirical evidence to support this claim.

In all prior work, however, a major difficulty in semantic interpretation has been the design of a standard set of semantic relations with which to perform the interpretation. For interpreting noun compounds, the scalability and portability to novel domains/NC types is questionable, as methods make specific assumptions about the domain or range of NC interpretation. The current level of accuracy of NC interpretation over open domain data is not high enough to utilize the acquired data for NLP applications. Also, lack of agreement on the semantic relations used for MWE interpretation makes it hard to incorporate NC interpretation into applications. Another point is that much of the work on semantic interpretation is based on supervised methods, which raises questions about the amount of training data and effective learning algorithms for a particular method or set of semantic relations.

1.2.6 Cross-over and Cross-lingual study

Clearly there is merit in applying MWE research over as broad a range of MWEs as possible in order to avoid having to develop specialized methods for large numbers of discrete MWE types. That is, it is in the interests of the MWE community to apply the results of research over one type of MWE to similar computational tasks involving other types of MWE. Cross-lingual studies of MWEs have similar benefits in terms of rapid prototyping of MWE methods in novel languages, and also refining monolingual MWE analysis through cross-comparison and exposing general mechanisms in MWEs. Data gleaned from monolingual studies on MWEs can easily be adopted to studies on similar types of MWEs in other languages.

There is broad similarity between compositionality modeling and semantic classification methods used across various types of MWE and languages (Baldwin *et al.* 2003a; Katz and Giesbrecht 2006; Uchiyama *et al.* 2005; Kim and Baldwin 2007a). For example, Baldwin *et al.* (2003a) used LSA to measure the degree of compositionality of English MWEs, while Katz and Giesbrecht (2006) used the same technique to detect the compositionality of idioms in German. Uchiyama *et al.* (2005) proposed a classifier for the semantic interpretation of semi-productive Japanese compound verbs based on a matrix-style analysis of the combinatorics of main and auxiliary verbs. Later, Kim and Baldwin (2007a) used the same technique to build a classifier to measure the compositionality of English VPCs. Baldwin *et al.* (2006) and Van Der Beek (2005) performed parallel analysis of D-PPs in English and Dutch and unearthed remarkable consistencies and divergences in the two languages. Venkatapathy and Joshi (2006) used MWE information to align English and Hindi words in order to improve the performance of statistical MT. Girju (2007) applied multiple cross-linguistic features to the interpretation of NCs.

Such cross-over and cross-lingual studies increase the re-usability of existing techniques and models, and also reduce the overhead in dealing with individual types of MWE across different languages. Both cross-over and cross-lingual MWE studies face difficulties in coming up with common or comparable sets of features to use in analysis.

In this thesis, we do not conduct any specific cross-over or cross-lingual research on MWEs. Instead, we attempt to integrate the intuitions and methods used to model a variety of MWE types in English and other languages. For example, the intuition behind our method for identifying English VPCs is based on the finding of Baldwin *et al.* (2003a), and the method for modeling the compositionality of English VPCs is based on the claims of Villavicencio (2005). Also, we use the classification method developed by Uchiyama *et al.* (2005) for the semantic classification of Japanese compound verbs, and apply it to detecting the compositionality of English VPCs.

1.3 Focus of the Thesis

1.3.1 Our Aims and Approaches

The goals of this MWE study are to shed light on underlying linguistic processes giving rise to MWEs across constructions and languages, to generalize techniques for analyzing MWEs, to abstract away from individual MWE types to develop general-purpose interpretation methods, to cross-compare pre-existing MWE classifications, and finally to exemplify the utility of MWE interpretation within general NLP tasks.

To develop a framework for modeling MWEs in this thesis, our principal approach is to employ statistical approaches, and furthermore to integrate symbolic approaches wherever possible to build from richer syntactic and semantic representations.

The main hurdle in modeling MWEs is their high productivity as well as immense variety of linguistic features. Some types of MWE such as noun compounds, verb-particle constructions and light-verb constructions produce (virtually) unlimited numbers of lexical items. Statistical techniques provide the means to scale over such large numbers of MWEs and capture the massive diversity, and generalize much more readily than hand-crafted rules or methods. For example, in the context of determinerless prepositional phrases, the rules are not easily simplified. Some D-PPs such as *by car* and *by foot* can be effectively captured by hand-crafted rules based on the observation that any form of transport or movement can combine with *by*, but this doesn't generalize to examples such as *by cloud* that is not a D-PP, or alternatively overgeneralizes to examples such as *by night* which do not have a transport meaning. Furthermore, in previous research, statistical methods have been shown to repurpose readily across related MWEs (Grefenstette and Teufel 1995; Lauer 1995; Schone and Jurafsky 2001; Baldwin and Villavicencio 2002; Baldwin *et al.* 2003a; McCarthy *et al.* 2003; Kim and Baldwin 2005).

Finally, we believe that by hybridizing across statistical and symbolic approaches we should be able to maximize the applicability of this research. While statistical approaches are good for rapid prototyping, they have the tendency to hit a ceiling because of not having access to rich enough data representations, while symbolic approaches can be brittle but provide access to deeper features which provide the means to generalize across linguistic phenomena. For example, existing lexical resources such as WordNet and CoreLex provide a powerful means to capture the semantics of MWEs and their components, and specify the subtle semantic interactions found in MWEs.

1.3.2 Scope of Research

In this thesis, we exclusively deal with English MWEs. Our motivation in targeting English MWEs relates largely to resource availability.

There is currently a large number of lexical resources (e.g. WordNet and CoreLex) and tools/software (e.g. RASP and WordNet::Similarity) available for English. Resources such as WordNet and RASP have been widely used as a means of syntactic and semantic analysis for various NLP tasks in English. In developing our framework for MWEs, we attempt to leverage such resources wherever possible.

We focus our attention primarily on noun compounds (NCs) and verb-particle constructions due to their high frequency and productivity. We will focus predominantly on binary NCs, i.e. NCs made up of two nouns (e.g. *computer science* and *golf club*), since we estimate that 88.4% of NCs in the Wall Street Journal and 90.6% of NCs in the British National Corpus are binary. We do, however, look at higher-arity NCs briefly in the context of NC bracketing.

1.4 Thesis Outline and Our Contribution

In this section, we provide a synopsis of each chapter and outline our contribution in the context of modeling English MWEs.

Chapter 2: The Linguistics of Multiword Expressions

Chapter 2 provides an overview of MWEs, and details the linguistic properties of various types of English MWE, namely: noun compounds, verb-particle constructions, light-verb constructions, idioms and determinerless-prepositional phrases. Further, we present computational tasks involving each type of MWE, and briefly review prior research in computational linguistics.

Chapter 3: Statistical Frameworks and Related Work

Chapter 3 presents seven general statistical approaches to MWEs: substitutability, distributional similarity, co-occurrence properties, statistical tests, semantic similarity, recovering ellipsed predicates and linguistic properties. We also review prior research on MWEs which has made use of each of these approaches, and point out the advantages and disadvantages of each approach.

Chapter 4: Resources

Chapter 4 describes the resources that we use in this research. It contains three parts: corpora (e.g. Wall Street Journal and Brown Corpus), lexical resources (e.g. Moby Thesaurus and CoreLex) and tools (e.g. WordNet::Similarity and TiMBL).

Chapter 5: MWEs and Semantic Similarity

Chapter 5 presents methods based on semantic similarity for modeling the semantics of noun compounds and verb-particle constructions.

The first task is to interpret the semantic relations in noun compounds (Kim and Baldwin 2005; Kim and Baldwin 2007d; Kim *et al.* 2007; Kim and Baldwin 2008). The method is built around word similarity, based on nearest-neighbor matching over the union of senses of the modifier and head noun computed by WordNet::Similarity. Based on the proposed method, we carry out various sub-experiments. First, we evaluate the relative contribution of the modifier and head noun in noun compounds in predicting the semantic relation. Second, we apply the proposed method over ternary NCs to testify the adaptability of the method. Third, we extend the basic model based on bootstrapping and the k-nearest neighbor algorithm in an attempt to improve performance. Finally, we utilize the acquired semantic relations in a syntactic disambiguation task (bracketing).

The second task is to model the compositionality of English verb-particle constructions in terms of the contribution of the simplex verb and particle (Kim and Baldwin 2007a). We check for overlap in the semantics of the VPC and verb in isolation to estimate the compositionality of the VPC. We also model the contribution of the verb and particle in compositional VPCs using the proposed method. Finally, we check the correlation between the performance of our method and human judgments.

Our contribution in modeling two distinctive MWE tasks is: (1) to showcase general-purpose interpretation methods using semantic similarity and hand-crafted resources; (2) to exemplify the utility of MWE interpretation within NLP tasks; and (3) to port a method developed for another MWE type in another language to English VPCs. We testify that our semantic similarity method provides the means to effectively capture the semantic variation found in NCs and VPCs, in combination with hand-crafted resources (i.e. WordNet and CoreLex). Our proposed method for NC interpretation is shown to enhance the performance of NC task, a highly novel outcome. In addition, we demonstrate the potential for cross-lingual MWE research by repurposing a method developing for Japanese compound verbs to English VPCs.

Chapter 6: MWEs and Ellipsed Predicates

Chapter 6 presents a novel method for automatically interpreting noun compounds based on a predefined set of semantic relations (Kim and Baldwin 2006b). First, we map verb tokens in sentential contexts to a fixed set of seed verbs using WordNet::Similarity and Moby Thesaurus. We then match the sentences with semantic relations based on the semantics of the seed verbs and grammatical roles of the head noun and modifier. Based on the semantics of the matched sentences, we then build a classifier using TiMBL.

In this chapter, our main contribution is to generalize the underling linguistic processes relating to NCs and emphasize the benefits of hand-crafted resources (i.e. WordNet and CoreLex), NLP tools (i.e. WordNet::Similarity and RASP), and linguistic classifications (i.e. grammatical roles of nouns). That is, we evidence the potential for combining the benefits of statistical and symbolic approaches. We develop a framework for interpreting NCs based on the notion of ellipsed predicates, and blend NLP resources to acquire deeper, richer syntax and semantics.

Chapter 7: MWEs and Substitutability

Chapter 7 is concerned with the interaction between word sense disambiguation and the interpretation of noun compounds (NCs) in English. We investigate two computational tasks involving NCs based on substitutability and co-occurrence.

The first method we investigate is a novel method to acquire noun compounds automatically tagged with semantic relations (Kim and Baldwin 2007c; Kim and Baldwin 2007d; Kim *et al.* 2007; Kim and Baldwin 2008). The method is motivated by the semantic collocation of constituents in noun compounds. We automatically generate tagged NCs by replacing one constituent at a time with similar words. As similar words, we use *synonym*s, *hypernym*s or *sister word*s. Also, we attest the usability of the automatically-acquired NCs by incorporating them within existing interpretation methods. That is, we apply the proposed method in expanding the training instances for SemEval-2007. In addition, we hybridize a basic method based on semantic similarity, and demonstrate the potential to improve the performance of NC interpretation.

Secondly, we develop techniques for disambiguating word sense specifically in the context of NCs, and then investigate whether word sense information can aid in the semantic relation interpretation of NCs (Kim and Baldwin 2007b). To disambiguate word sense, we combine the one-sense-per-collocation heuristic with the grammatical role of polysemous nouns and analysis of word sense combinatorics. We built supervised and unsupervised classifiers for the task and demonstrate that the supervised methods are superior to a number of baselines and also a benchmark state-of-the-art WSD system. Finally, we show that WSD can significantly improve the accuracy of NC interpretation.

Our main contribution in these tasks are: (1) to shed light on the underlying linguistic nature of NC interpretation; (2) to exemplify an extension to a basic interpretation method which enhances performance; and (3) to cross-compare existing methods on NC interpretation. We expose some of the underlying linguistic processes relating to the semantics of NCs. We also demonstrate the usability of the proposed method on word sense disambiguation for NC interpretation. Finally, we benchmark existing semantic similarity-based approaches and further the performance of NC interpretation.

Chapter 8: MWEs and Linguistic Properties

Chapter 8 presents a method for automatically identifying token instances of VPCs based on the output of RASP (Kim and Baldwin 2006a). The proposed method pools together instances of VPCs and verb-PPs from the output of RASP and uses the sentential context of each such instance to differentiate VPCs from verb-PPs. We show our technique to perform at an F-score of 97.4% at identifying VPCs in Wall Street Journal and Brown Corpus data. We further investigate the performance of existing resources on the VPC identification task. Finally, we correlate VPC identification with the compositionality of VPCs.

Our contribution in this chapter is to shed light on the linguistic processes giving rise to English VPCs, and harness linguistic properties in a generalized framework. We also testify the effectiveness of existing resources (i.e. a chunker and parsers) on differentiating the VPCs and Verb-PPs, to provide tentative guidelines for further usage. In addition, we compare the relative performance at VPC identification over VPCs of varying compositionality.

Chapter 9: Conclusion

Chapter 9 summarizes our attempts at modeling English MWEs and presents the findings of each chapter. We also describe the weaknesses of our proposed methods, and propose possible applications for this research in NLP applications. Finally, we conclude the thesis by suggesting further improvements to our methods and pointing the way to future work.

Chapter 2

The Linguistics of Multiword Expressions

Multiword expressions can be syntactically and semantically categorized into various types, including noun compounds and idioms. Each type of MWE has distinctive linguistic features which we will describe in Section 2.1.1. Due to these differences, for distinct MWEs, we have disparate objectives for knowledge acquisition and different obstacles to overcome. For example, interpreting semantic relations in noun compounds is a hard task while extracting or identifying them is relatively trivial. On the other hand, extracting or identifying verb-particle constructions is challenging since there is often ambiguity with a verb-PP analysis. Also, measuring compositionality is an important task for VPCs as there is a more uniform distribution of VPCs across the spectrum of compositionality, whereas it is less of an issue for noun compounds as they are mostly compositional.[1]

In this chapter, we will survey the linguistics of the major types of English MWE. Specifically, we will describe the relevant linguistic properties of noun compounds, verb-particle constructions, light-verb constructions, idioms and determinerless prepositional phrases.

[1] That is noun compound *types* are mostly compositional; noun compound *tokens* are arguably not.

2.1 Overview of Multiword Expressions

Multiword expressions are lexical items that can be decomposed into multiple simplex words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag *et al.* 2002; Calzolari *et al.* 2002).

2.1.1 Linguistic Properties of Multiword Expressions

The linguistic properties of MWEs can be used to determine whether multiple simplex words form an MWE or not. Particularly, these properties are effective indicators in distinguishing MWEs from simple word collocations (see below), which are often misidentified as MWEs due to their high frequency.

The main properties of English MWEs are: *idiomaticity*, *non-identifiability*, *figuration*, *situatedness*, *single-word paraphrasability*, *proverbiality* and *prosody* (Cruse 1986; Fillmore *et al.* 1988; Liberman and Sproat 1992; Nunberg *et al.* 1994; Jackendoff 1997; Sag *et al.* 2002). None of these provides a water-tight test of MWEhood, and we are instead on the lookout for at least one of these properties to hold in order to classify a given word combination as an MWE. In this thesis, we review these properties under two different headings: *idiomaticity* and *other properties*. That is, *idiomaticity* contains elements of *lexico-syntactic*, *semantic*, *pragmatic* and *statistical idiomaticity*. We also consider *non-identifiability* and *figuration* as *semantic idiomaticity*, and *situatedness* as *pragmatic idiomaticity*, although some researchers define them separately. *Other properties* includes *single-word paraphrasability*, *proverbiality* and *prosody*.

Idiomaticity

Idiomaticity is defined as lexico-syntactic, semantic, pragmatic, and statistical markedness (Katz and Postal 2004; Wood 1964; Chafe 1968; Cruse 1986; Jackendoff 1997; Sag *et al.* 2002). Lexico-syntactic idiomaticity means that the MWE has surprising syntax given the syntax of the individual simplex words. For example, *apple pie* is the entirely unsurprising combination of the nouns *apple* and *pie* into an $\bar{\text{N}}$, whereas *by and large* is a coordination of a preposition and an adjective to form an adverbial phrase, an effect which is not predicted by standard English grammar rules. As such, *apple pie* is not lexico-syntactically idiomatic while *by and large* is. Semantic irregularity commonly happens in idioms such as *in one's shoes*, where the semantics is not immediately predictable from the simplex semantics of *shoes*. Pragmatic idiomaticity occurs in situated expressions such as *good morning* and *all board*. That is, these MWEs are associate with very particular situations and are anomalous in other contexts (e.g. *good morning* when finishing a meal, or *all aboard* when watching a soccer match). Statistical idiomaticity occurs with MWEs such as *black and white* where they occur with uncommonly high frequency in contrast to alternative forms of the same expression. I.e., it is perfectly acceptable to say *white and black*, but the skew towards the first form is sufficiently great that *white and black photograph*, e.g., is marked in English.

Below, we present a more detailed account of these different forms of idiomaticity.

- *Lexico-syntactic Idiomaticity*

  Lexico-syntactic idiomaticity refers to syntactic markedness in an MWE, in that the syntax of the MWE is not derived directly from that of its components (Katz and Postal 2004; Chafe 1968; Bauer 1983; Sag *et al.* 2002). For example, *by and large* is lexico-syntactically idiomatic in that it is an adverbial phrase made up on the coordination of a <u>preposition</u> (*by*) and an <u>adjective</u> (*large*). Similarly, *ad hoc* is syntactically marked in that it is an adjective phrase but formed from components (*ad* and *hoc*) which do not have syntax as simplex words in a standard English lexicon. On the other hand, *take a walk* is not syntactically marked as it is a simple verb–object combination which is derived transparently from a transitive verb (*walk*) and a noun (*walk*). Figure 2.1 summarizes these three examples.

- *Semantic Idiomaticity*

  Semantic idiomaticity is a reflection of the meaning of a MWE not being explicitly or implicitly derivable from its parts (Katz and Postal 2004; Chafe 1968; Bauer 1983; Sag *et al.* 2002). For example, *birds of a feather* usually indicates

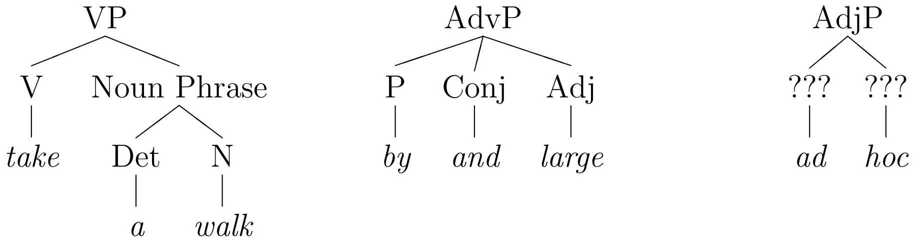

Figure 2.1: Examples of syntactic non-markedness vs. markedness

"people with similar interests", which we could not predict from either *birds* or *feather*. On the other hand, *all aboard* is not semantically marked as its semantics is fully predictable from its parts. Many cases are not as clear cut as these, however. The semantics of *blow hot and cold* ("constantly change opinion"), for example, is partially predictable from *blow* ("move" and hence "change"), but not as immediately from *hot and cold*. There are also cases where the meanings of the parts are transparently inherited in the MWE but there is additional semantic content which has no overt realization. One such example is *bus driver* where, modulo the effects of word sense disambiguation, *bus* and *driver* both have their expected meanings, but there is additionally the default expectation that a *bus driver* is "one who drives a bus" and not "one who drives *like* a bus", e.g. *Semantic idiomaticity* is directly related to compositionality, as the degree of semantic contribution of the components indicates the *semantic idiomaticity* as well as the compositionality.

Related to the issue of *semantic idiomaticity*, there has been discussion of the notions of *non-identifiability* and *figuration* (Fillmore *et al.* 1988; Liberman and Sproat 1992; Nunberg *et al.* 1994). We roughly classify these properties under our definition of *semantic idiomaticity* for the purposes of this thesis.

Non-identifiability (Nunberg *et al.* 1994) is the notion of the meaning of an MWE not being easily predictable from the surface form (components), much like our definition of *semantic idiomaticity*. For example, the meaning of *kick the bucket*

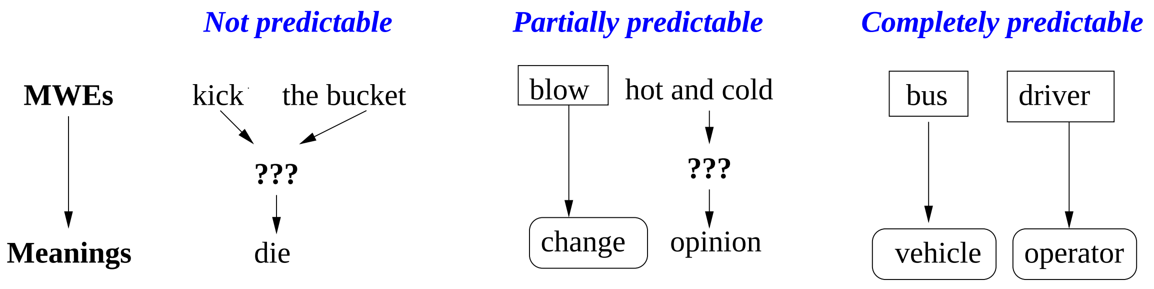

Figure 2.2: Examples of semantic idiomaticity

("die") cannot be derived from either *kick* or *bucket*. Another example is *make out*, where the parts (i.e. *make* and *out*) do not semantically contribute to the meaning of the whole. This property relates closely to compositionality. That is, when MWEs are compositional, the meaning of MWEs can be predicted from the parts. Hence, non-identifiability coincides with non-compositionality (other examples of non-identifiable and non-compositional MWEs are *on ice*, *cock up*, *chicken out* and *by and large*).

Figuration (Fillmore *et al.* 1988; Nunberg *et al.* 1994) is an attribute of encoded expressions such as metaphors (e.g. *take the bull by the horns*), metonymies (e.g. *lend a hand*) and hyperboles (e.g. *not worth the paper it's printed on*). It is defined as the property of the components of an MWE having some metaphoric or hyperbolic meaning in addition to their literal meaning. That is, the semantics of the MWE is derived from the components through a process of metaphor, hyperbole or metonymy, although the precise nature of the figuration may be more or less obvious. Hence, *figuration* involves subtle interactions between idiomatic and literal meaning.

We return to touch on the relationship between *figuration* and *semantic idiomaticity* below.

- *Pragmatic Idiomaticity*

  Pragmatic idiomaticity is the condition of an MWE being associated with a fixed set of situations or a particular context (Kastovsky 1982; Jackendoff 1997; Sag *et al.* 2002). *Good morning* and *all aboard* are examples of pragmatic MWEs, where the first is a greeting associated specifically with mornings[2] and the second is a command associated with the specific situation of train or ship stations and the imminent departure of a train or ship. Although these examples can be interpreted literally, e.g. as "pleasant morning" in the former case (c.f. *Kim had a good morning*), their meaning differs slightly from the pragmatically-marked sense. Examples of MWEs which are not pragmatically marked are *on TV* and *to and fro*.

- *Statistical Idiomaticity*

  Statistical idiomaticity occurs when a combination of words occurs with surprising frequency, relative to the component words or alternative phrasings of the same expression (Pawley and Syder 1983; Cruse 1986; Sag *et al.* 2002). Cruse (1986:p281) provides some nice examples of *statistical idiomaticity* in the matrix of adjectives and nouns presented in Table 2.1. The adjectives are largely synonymous, and yet different nouns have particular preferences for certain subsets of the adjectives as modifiers, as indicated by the cells in the matrix ("+" indicates a strong lexical affinity, "?" indicates a marginal lexical affinity, and "−" indicates a negative lexical affinity).

  Note that the statistical idiomaticity (i.e. the alternative phrasing) can be in terms of alternative orderings of the components. For example, *black and white* in much more common in English than *white and black*, while the reverse holds in the case of other languages such as Japanese and Spanish (see Table 2.1).

  For the purposes of this thesis, we will follow Sag *et al.* (2002) in referring to MWEs which are only statistically idiomatic (i.e. not also lexico-statistically, semantically or pragmatically idiomatic) as *collocations*.

[2] Which is not to say that it can't be used ironically at other times of the day!

Statistical idiomaticity relates to the notion of *institutionalization/conventionalization*, i.e. a particular word combination coming to be used to refer a given object (Fernando and Flavell 1981; Bauer 1983; Nunberg *et al.* 1994; Sag *et al.* 2002). For example, *traffic light* is the conventionalized descriptor for "a visual signal to control the flow of traffic at intersections". There is no reason why it shouldn't instead be called a *traffic director* or *intersection regulator*, but the simple matter of the fact is that it is not referred to using either of those expressions; instead, *traffic light* was settled on as the canonical term for referring to the object. Similarly, it is an arbitrary fact of the English language that we say *many thanks* and not **several thanks*, and *salt and pepper* in preference to *pepper and salt.*[3]

Nunberg *et al.* (1994) consider *collocation* (*conventionality* in their terms) to be a mandatory property of MWEs. We consider conventionality to relate to semantic, pragmatic and statistical idiomaticity, but consider that MWEs do not have to have any one of these three forms of markedness (e.g. MWEs which are strictly lexico-syntactically idiomatic are classified as MWEs in this research). Collocations are most apparent when observed in contrast with *anti-collocations*. *Anti-collocations* are lexico-syntactic variants of collocations which have unexpectedly low frequency (Pearce 2001). For example, *pepper and salt* is an anti-collocation for *salt and pepper*, and *traffic director* is an anti-collocation for *traffic light*.

It is important to acknowledge that our use of the term *collocation* differs from the mainstream usage in computational linguistics, where a collocation is often defined as an arbitrary and recurrent word combination that co-occurs more often than would be expected by chance (Choueka 1988; Lin 1998b; Evert 2004). Often, the recurrence is the result of syntactic or semantic markedness, such that the more standard definition of collocation is much broader than that used in this thesis.

[3] Which is not to say there wasn't grounds for the selection of the canonical form at its genesis, e.g. for historical, crosslingual or phonological reasons.

	unblemished	spotless	flawless	immaculate	impeccable
performance	−	−	+	+	+
argument	−	−	+	−	?
complexion	?	?	+	−	−
behavior	−	−	−	−	+
kitchen	−	+	−	+	−
record	+	+	−	?	+
reputation	?	+	−	?	?
taste	−	−	?	?	+
order	−	−	?	+	+
credentials	−	−	−	−	+

Table 2.1: Examples of statistical idiomaticity ("+" = strong lexical affinity, "?" = marginal lexical affinity, "−" = negative lexical affinity) (Cruse 1986)

Above, we described four different forms of *idiomaticity*. We bring these together in categorizing a selection of MWEs in Table 2.2.

In Table 2.2, some examples such as *kick the bucket*, *make out* and *traffic light* are marked with only one form of idiomaticity, which is sufficient for them to be classified as MWE. On the other hand, others such as *shock and awe* and *to and fro* are idiosyncratic in more ways than one. We analyze *shock and awe* as being pragmatically idiomatic because of its particular association with the bombardment of Baghdad at the commencement of the Iraq War, and *to and fro* as being lexico-syntactically idiomatic because of the relative syntactic opacity of the antiquated *fro*.

Other Properties

Other properties of MWE relevant to this thesis are *single-word paraphrasability*, *proverbiality* and *prosody*. Unlike *idiomaticity*, where some form of idiomaticity is a necessary feature of MWEs, these other properties are neither necessary nor sufficient. *Prosody* relates to *semantical idiomaticity*, while the other properties are independent of idiomaticity as described above.

	Lexico-syntactic	Semantic	Pragmatic	Statistical
all aboard	−	−	+	+
black and white	−	?	−	+
by and large	+	+	−	−
kick the bucket	−	+	−	−
make out	−	+	−	−
shock and awe	−	−	+	+
to and fro	+	−	−	+
bus driver	−	+	−	+
traffic light	−	−	−	+

Table 2.2: Classification of MWEs in terms of different forms of idiomaticity

- Single-word paraphrasability

  Single-word paraphrasability is the observation that significant numbers of MWEs can be paraphrased with a single word (Chafe 1968; Gibbs 1980; Fillmore *et al.* 1988; Liberman and Sproat 1992; Nunberg *et al.* 1994). While some MWEs are single-word paraphrasable (e.g. *leave out* = "omit"), some are not (e.g. *look up* = ?). Also, MWEs with arguments can sometimes be paraphrasable (e.g. *take off clothes* = "undress"), just as multi-word non-MWEs can be single-word paraphrasable (e.g. *not sufficient* = "insufficient").

- Proverbiality

  Proverbiality is the ability of an MWE to "describe and implicitly to explain a recurrent situation of particular social interest in the virtue of its resemblance or relation to a scenario involving homely, concrete things and relations" (Nunberg *et al.* 1994). For example, VPCs and idioms are often indicators of more informal situations (e.g. *piss off* is an informal form of *annoy*, and *kick the bucket* is an informal form of *die, demise*). Nunberg *et al.* (1994) treat *informality* as a separate category, where we combine it with *proverbiality*.

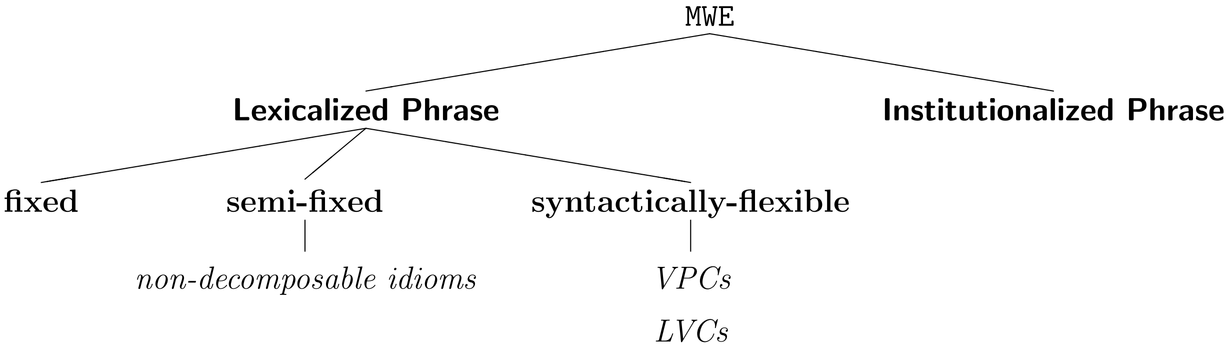

Figure 2.3: MWE types (Sag et al. 2002)

- Prosody

  MWEs can have distinct *prosody*, i.e. stress patterns, from compositional language (Fillmore *et al.* 1988; Liberman and Sproat 1992; Nunberg *et al.* 1994). For example, when the components do not make an equal contribution to the semantics of the whole, MWEs can be prosodically marked, e.g. *soft spot* is prosodically marked (due to the stress on *soft* rather than *spot*), although *first aid* and *red herring* are not. Note that *prosodic* marking can equally occur with non-MWEs, such as *dental operation.*

2.1.2 Types of Multiword Expression

English MWEs can be syntactically and semantically categorized in various ways. In this thesis, we adopt the classification and terminology of Bauer (1983) and Sag *et al.* (2002), as outlined in Figure 2.3. The classification of MWEs into *lexicalized phrases* and *institutionalized phrases* hinges on whether the MWE is lexicalized (i.e. explicitly encoded in the lexicon) on the grounds of lexico-syntactic or semantic idiomaticity, or a simple collocation (i.e. only statistically idiosyncratic). Note that we will largely ignore pragmatic idiomaticity for the remainder of this thesis.

Lexicalized phrases are MWEs in which the components have idiosyncratic syntax or semantics in part or in combination. Lexicalized phrases can be further split into:

fixed expressions (e.g. *by train*, *at first*), *semi-fixed expressions* (e.g. *spill the beans*, *car dealer*, *Chicago White Socks*) and *syntactically-flexible expressions* (e.g. *add up*, *give a demo*). Following Sag *et al.* (2002), we assign individual MWE construction types to each of these sub-categories of lexicalized phrases, namely: adjective-noun combinations, non-decomposable idioms, verb-particle constructions and light-verb constructions. Determinerless PPs, on the other hand, cut across all three sub-categories. Note that we slightly differ from the original categories presented in Sag *et al.* (2002) that compound nouns occur across all categories.

- *fixed expressions* are fixed strings that undergo neither morphosyntactic variation nor internal modification. For example, *by and large* is not morphosyntactically modifiable (e.g. $*$*by and larger*) or internally modifiable (e.g. $*$*by and very large*). Non-modifiable determinerless prepositional phrases such as *on air* and *by car* are also fixed expressions.

- *semi-fixed expressions* are lexically-variable MWEs that have hard restrictions on word order and composition but undergo some degree of lexical variation such as inflection (e.g. *kick/kicks/kicked/kicking the bucket* vs. $*$*the bucket was kicked*), variation in reflexive pronouns (e.g. *in her/his/their shoes*) and determiner selection (e.g. *The Beatles vs. a Beatles album* that show the first is a lexically-variable MWE while the second is not.). Note that the determiner *the* in *The Beatles* is obligatory in the case that *The Beatles* forms a noun phrase (i.e. *Beatles* can only be quantified by *the*), but in cases where *Beatles* forms an $\bar{\text{N}}$, e.g. in [$_{NP}$ *a* [$_{N'}$[$_{N'}$ *Beatles'*] *album*]], the lexical item is realized without a determiner. Non-decomposable idioms (e.g. *kick the bucket*, *shoot the breeze*) and compound nominals (e.g. *attorney general, part of speech*) are also classified as semi-fixed expressions (Sag *et al.* 2002).

- *syntactically flexible expressions* are MWEs which undergo syntactic variation, such as verb-particle constructions, light-verb constructions and decomposable idioms. The nature of the flexibility varies significantly across construction types. Verb-particle constructions, for example, are syntactically flexible with respect to the word order of the particle and NP in transitive usages: *hand in the paper* vs. *hand the paper in*. They are also usually compatible with internal modification, even for intransitive VPCs: *the plane took right off*. Light-verb constructions (e.g. *give a demo*) undergo full syntactic variation, including passivization (e.g. *a demo was given*), extraction (e.g. *how many demos did he give?*) and internal modification (e.g. *give a clear demo*). Decomposable idioms are also syntactically flexible to some degree, although the exact form of syntactic variation is hard to predict (Riehemann 2001).

As described in Section 2.1.1, *collocations* (or *institutionalized phrases*) are MWEs that occur with surprising frequency, relative to the component words or alternative phrasings of the same expression (i.e. they are strictly statistically idiosyncratic), but which are otherwise unmarked. Examples include *peanut butter and jam*, *salt and pepper*, *telephone booth*, *many thanks* and *traffic light*.

In the remainder of this thesis we will focus on lexicalized phrases and make little mention of collocations.

2.2 Noun Compound

Noun compounds (i.e. *NCs*) form an $\bar{\mathrm{N}}$ and are made up of two or more nouns, such as *golf club* or *computer science department* with and/or without a space or with a hyphen (Lauer 1995; Sag *et al.* 2002; Huddleston and Pullum 2002). Note that in this thesis, we deal with NCs with a space only. We will refer to the rightmost noun in the NC as the *head noun* (i.e. *club* and *department*, respectively) and the remainder of the component(s) as *modifier(s)* (i.e. *golf* and *computer science*, respectively). There is some disagreement in the precise scope of the term noun compound. Lauer (1995), for example, required that the head be a noun but allowed the modifiers to be nouns, verbs, adjectives or adverbs.

The main syntactic challenge with NCs is to disambiguate the phrase structure of NCs with 3 or more terms. For example, *glass window cleaner* can be syntactically analyzed as either *(glass (window cleaner))* (i.e. a window cleaner made of glass, or similar) or *((glass window) cleaner)* (i.e. a cleaner of glass windows). Syntactic ambiguity impacts on both the semantic interpretation and prosody of the NC. Various methods have been proposed for disambiguating the syntax of NCs with 3 terms or more, based on frequency and context (Marcus 1980; Lauer 1995; Buckeridge and Sutcliffe 2002; Nakov and Hearst 2005). Disambiguating syntactic ambiguity is called ***bracketing*** and will be a component task of Section 5.1.

The main semantic challenge with NCs is to determine their semantic interpretation by way of ***semantic relations*** (SRs) (Downing 1977; Warren 1978; Levi 1978; Finin 1980; Sparck Jones 1983). SRs provide the means to identify the underspecified two-place directed relation that holds between the head noun and modifier(s). For example, the NC *orange juice* can be interpreted by way of the SR MAKE (i.e. it is juice made from oranges), while *morning juice* can be interpreted by way of the SR TIME (i.e. it is juice associated with the time of morning). Interpreting NCs by way of SRs is one of the major tasks of this thesis, and will be described in detail in Section 5.1.

2.3 Verb-Particle Constructions

Verb-particle constructions (i.e. *VPCs*) are made up of a verb and obligatory particle(s) such as *hand in* and *take off* (Bolinger 1976b; Jackendoff 1997; Huddleston and Pullum 2002; Sag *et al.* 2002). The obligatory particles are usually intransitive prepositions, adjectives or verbs, as shown in (2.1)–(2.3).

(2.1) verb + intransitive prepositions : *battle on*, *take off*

(2.2) verb + adjectives : *cut short*, *band together*

(2.3) verb + verbs : *let go*, *let fly*

Generally, VPCs are both idiosyncratic or semi-idiosyncratic combinations although some are adverbial and/or non-lexical particle cases (Dehe *et al.* 2001). In this thesis, we will focus exclusively on VPCs where the particle is prepositional.

VPCs often involve subtle interactions between the verb and particle (Bolinger 1976b; Jackendoff 1973; Fraser 1976; Lidner 1983; Kayne 1985; Svenonius 1994; Dehe *et al.* 2001; Dehe 2002). For example, the particle can impact on various properties of the verb, including: aspect (e.g. *eat* vs. *eat up*), valency (e.g. *make* vs. *make out*), conation (e.g. *hit* vs. *hit at*), reciprocity (e.g. *ring* vs. *ring back*) and repetition (e.g. *start* vs. *start over*).

Note that VPCs are termed ***phrasal verbs*** by some researchers (Bolinger 1976b; Side 1990; Dirven 2001; McCarthy *et al.* 2003) and verb-particle constructions by others (Dehe *et al.* 2001; Bannard *et al.* 2003; Bannard 2003; Baldwin *et al.* 2003a; Cook and Stevenson 2006; Kim and Baldwin 2007a). In this thesis, we will refer to them exclusively as VPCs.

One MWE type which relates closely to VPCs is ***prepositional verbs*** (Jackendoff 1973; O'Dowd 1998; Huddleston and Pullum 2002; Baldwin 2005b), which are similarly made up of a verb and preposition, but the preposition is transitive and selected for by the verb (e.g. *refer to, look for*). It is possible to differentiate transitive VPCs[4] from prepositional verbs via their respective linguistic properties (Bolinger 1976b; Jackendoff 1973; Fraser 1976; Lidner 1983; O'Dowd 1998; Dehe *et al.* 2001; Jackendoff 2002; Huddleston and Pullum 2002; Baldwin 2005b):

- in the case that the object NP is not pronominal, transitive VPCs can occur in either the joined or split word order (c.f. (2.4)), while prepositional verbs must always occur in the joined form (c.f. (2.7));
- in the case that the object NP is pronominal, transitive VPCs must occur in the split word order (c.f. (2.5)), while prepositional verbs must occur in the joined form (c.f. (2.8));
- manner adverbs cannot occur between the verb and particle in VPCs (c.f. (2.6)), while they can with prepositional verbs (c.f. (2.9)).

[4] Prepositional verbs are obligatorily transitive, so there is no ambiguity with intransitive VPCs.

Verb-particle constructions

(2.4) **Non-pronominal object: optional joined/split word order**

- *Put on the sweater.*
- *Put the sweater on.*

(2.5) **Pronominal object: obligatory split word order**

- *Finish it up.*
- ∗*Finish up it.*

(2.6) **With manner adverb**

- *Quickly eat up the food.*
- ∗*Eat quickly up the food.*

Prepositional verbs

(2.7) **Non-pronominal object**

- *Look for a word.*
- ∗*Look a word for.*

(2.8) **Pronominal object**

- *Look for it.*
- ∗*Look it for.*

(2.9) **With manner adverb**

- *Come with me quickly.*
- *Come quickly with me.*

Semantic Contribution	Compositional	Examples
verb & particle	Yes	*get down, take off*
verb	Yes	*lie down, eat up*
particle	Yes	*close off, be away*
none	No	*chicken out, make out*

Table 2.3: Classification of the compositionality of VPCs (Bannard et al. 2003 vs. McCarthy et al. 2003)

VPCs undergo morphological, syntactic and semantic variation. Morphologically, VPCs inflect for tense and number (e.g. *take/takes/took/have taken/is taken/... off*).

Syntactically, VPCs undergo word order variation, and are internally modifiable by a small set of adverbs (e.g. *right*, *back*, *way* and *all the way*).

Semantically, VPCs populate the spectrum of compositionality relative to their components (Lidner 1983; Brinton 1985; Ishikawa 1999; Olsen 2000; Jackendoff 2002; Bannard *et al.* 2003; Cook and Stevenson 2006). According to the view of Bannard *et al.* (2003), VPCs can be subclassified into four compositionality classes based on the independent semantic contribution of the verb and particle: (1) both the verb and particle contribute semantically, (2) only the verb contributes semantically, (3) only the particle contributes semantically, and (4) neither the verb nor the particle contributes semantically. Other researchers such as McCarthy *et al.* (2003) employ a one-dimensional classification of VPC compositionality (over a cline or a number of discrete sub-classes): compositional vs. non-compositional. Table 2.3 details the two classification systems, with examples.

2.4 Light-Verb Constructions

Light-verb constructions (i.e. LVCs) are made up of a verb and a noun complement, often in the indefinite singular form (Jespersen 1965; Abeillé 1988; Miyagawa 1989; Grefenstette and Teufel 1994; Hoshi 1994; Sag *et al.* 2002; Huddleston and Pullum 2002; Butt 2003; Stevenson *et al.* 2004). The name of the construction comes from the verb being semantically bleached or "light", in the sense that their contribution to the meaning of the LVC is relatively small in comparison with that of the noun complement. LVCs are also sometimes termed ***verb-complement pairs*** (Kan and Cui 2006) or support verb constructions (Calzolari *et al.* 2002). Our definition of ***light-verb constructions*** is in line with that of Huddleston and Pullum (2002).

The principal light verbs are *do, give, have, make, put* and *take*, for each of which we provide a selection of LVCs in (2.10)–(2.15). English LVCs generally take the form verb+*a*+object, although there is some variation here.

(2.10) **do:** *do a demo, do a drawing, do a report*

(2.11) **give:** *give a wave, give a sigh, give a kiss*

(2.12) **have:** *have a rest, have a drink, have (a) pity (on)*

(2.13) **make:** *make an offer, make an attempt, make a call*

(2.14) **put:** *put the blame (on), put an end (to), put stop (to)*

(2.15) **take:** *take a walk, take a bath, take a photograph (of)*

There is some disagreement in the scope of the term LVC, most notably in the membership of verbs which can be considered "light". Calzolari *et al.* (2002), e.g., argued that the definition of LVCs (or support verb constructions in their terms) should be extended as follows: (1) when the verbs combine with an event noun (deverbal or otherwise) and the subject is a participant in the event most closely identified with the noun (e.g. *take an exam, ask a question, make a promise*); and (2) when the subject of these verbs belongs to some scenario associated with the full understanding of the event type designated by the object noun (e.g. *pass an exam, survive an operation, answer a question, keep a promise*).[5]

Morphologically, LVCs inflect but the noun complement tends to have fixed number and a preference for determiner type (Wierzbicka 1982; Alba-Salas 2002; Kearns 2002; Butt 2003; Folli *et al.* 2003; Stevenson *et al.* 2004). For example, *put an end*

[5] All examples are from Calzolari *et al.* (2002).

_(to)_ undergoes full verbal inflection (_put/puts/putting an end (to)_), but the noun complement cannot be pluralized or modified derivationally (e.g. *_put an ending (to)_, *_put ends to_).[6]

As described above, there is little constraint on the syntax of LVCs. Semantically, although the meaning of the verb in LVCs is bleached, a given noun will usually have strong constraints on which light verb(s) it combines with to form an LVC (e.g. _put blame (on)_ vs. *_do/give/have/make blame_), and different light verbs can lead to VPCs with different semantics (Butt 2003). For example, _put blame (on)_ and _take blame_ are both LVCs but have very different semantics: the subject of _put blame (on)_ is the Agent of the blaming and the object of the PP headed by _on_ is the Patient, while the subject of _take blame_ is the Theme. Also, what light verb a given noun will combine with to form an LVC is often consistent across semantically-related noun clusters (e.g. _give a cry/moan/howl_ vs. *_take a cry/moan/howl_[7]).

2.5 Idioms

An _idiom_ is an MWE whose meaning is fully or partially unpredictable from the meanings of its components (e.g. _kick the bucket_, _blow hot and cold_) (Nunberg _et al._ 1994; Potter _et al._ 2000; Sag _et al._ 2002; Huddleston and Pullum 2002). Huddleston and Pullum (2002) identified subtypes of idioms such as _verbal idioms_ (e.g. _jump off, get out, run ahead_) and _prepositional idioms_ (e.g. _for example, in person, under the weather_) which we classify as VPCs/prepositional verbs and determinerless PPs, respectively. In our terms, therefore, idioms are those non-compositional MWEs not included in the named construction types of VPCs, prepositional verbs, noun compounds and determinerless PPs.

While all idioms are non-compositional (to varying degrees), we further categorize them into two groups: decomposable and non-decomposable (Nunberg _et al._ 1994). With _decomposable idioms_, given the interpretation of the idiom, it is possible to

[6] But also note other examples where the noun complement can be pluralized, e.g. _take a bath_ vs. _take baths._

[7] Examples are from Stevenson _et al._ (2004).

associate components of the idiom with distinct elements of the idiom interpretation based on semantics not immediately accessible from the components in isolation. Assuming an interpretation of *spill the beans* such as `reveal'(X,secret')`, e.g. we could analyze *spill* as having the semantics of `reveal'` and *beans* having the semantics of `secret'`, and hence arrive at a post hoc explanation for the interpretation of the idiom via the reverse-engineered semantics of the components (through figuration of some description). Note that the interpretations of the components (*spill* as `reveal'` and *beans* as `secret'`) are removed from those for the simplex words, and it is on this basis that we consider the idiom non-compositional. Other examples of decomposable idioms are *pull one's leg* and *pull strings*. Examples of non-decomposable idioms where a post hoc semantic decomposition is not accessible are *break a leg* and *kick the bucket*.

Decomposable idioms tend to be syntactically flexible, as defined by the nature of the semantic decomposition, whereas non-decomposable idioms tend not to be syntactically flexible (Katz and Postal 2004; Wood 1964; Chafe 1968; Kastovsky 1982; Pawley and Syder 1983; Cruse 1986; Jackendoff 1997; Sag *et al.* 2002). For example, *spill the beans* can be passivized (*It's a shame the beans were spilled*) and internally modified (*AT&T spilled the Starbucks beans*).

2.6 Determinerless-Prepositional Phrases

Determinerless prepositional phrases (i.e. D-PPs) are MWEs that are made up of a preposition and a singular noun without a determiner (Quirk *et al.* 1985; Huddleston and Pullum 2002; Sag *et al.* 2002; Baldwin *et al.* 2006).

Syntactically, D-PPs are highly diverse, and display differing levels of syntactic markedness, productivity and modifiability (Chander 1998; Ross 1995). That is, some D-PPs are non-productive (e.g. *on top* vs. ∗*on edge*) and non-modifiable (e.g. *on top* vs. ∗*on table top*), whereas others are fully-productive (e.g. *by car/foot/bus/...*) and highly modifiable (e.g. *at high expense*, *on summer vacation*). In fact, while some D-PPs are optionally modifiable (e.g. *on vacation* vs. *on summer vacation*), others require modification (e.g. ∗*at level* vs. *at eye level*, and *at expense* vs. *at company*

Class	Examples
institutional	*at school, in church, on campus, in gaol*
media	*on TV, on record, off screen, in radio*
metaphor	*on ice, at large, at hand, at liberty*
temporal	*at breakfast, on holiday, on break, by day*
means/manner	*by car, by hammer, by computer, via radio*

Table 2.4: A semantic classification of D-PPs

expense) (Baldwin *et al.* 2006).

Syntactically-marked D-PPs can be highly productive (Ross 1995; Grishman *et al.* 1998). For example, *by* combines with a virtually unrestricted array of countable nouns (e.g. *by bus/car/taxi/...*) but not with uncountable nouns (e.g. ∗*by information/linguistics/...*).

Semantically, D-PPs have a certain degree of semantic markedness on the noun (Haspelmath 1997; Mimmelmann 1998; Stvan 1998; Bond 2001; Borthen 2003). For example, *in* combines with uncountable nouns which refer to a social institution (e.g. *school, church, prison* but not *information*) to form syntactically-*un*marked D-PPs with marked semantics, in the sense that only the social institution sense of the noun is evoked (e.g. *in school/church/prison/...* vs. ∗*in information*) (Baldwin *et al.* 2006). Note that some D-PPs with *in* combine with countable nouns such as *pub* and *hospital* but they do not refer to social institution. In general, D-PPs have been categorized into five semantic groups by Stvan (1998). These classes often correlate with a particular compositionality, e.g. metaphorical D-PPs are non-compositional while the other classes are compositional.

2.7 Human annotation experiments on Multiword Expressions

Human annotation experiments were carried out to acquire the upper boundary over different tasks.

Annotation was done for the following data sets: (a) assigning semantic relations to 2- and 3-term NCs in Section 5.1.4; (b) assigning word senses to polysemous noun in 2-term NCs in Section 7.2.4; and (c) assigning semantics to particles in VPCs in Section 5.2.4. Although we will describe the details of the individual annotation tasks in each designated section, we briefly overview the annotation experiments here, including the annotator qualifications and how the annotation was done.

Two PhD students performed all of the annotations. One is an English native speaker with some experience in computational linguistics, and the other is a non-native English speaker with wide experience in computational linguistics and various annotation tasks.

To annotate the 2-term and 3-term NCs, the annotators were instructed to tag with a unique SR where possible, but also that multiple SRs were allowed in instances of genuine ambiguity. For example, *cable operator* can be interpreted as corresponding to the SR TOPIC (as in *operator is concerned with cable(s)*) or alternatively OBJECT (as in *cable is acted on by operator*). On completion of the annotation, the two annotators were instructed to come together and resolve any disputes in annotation. See Section 5.1.4 for more details.

For the word sense distribution analysis, the same annotators were provided with 45 NCs (5 NCs for each target noun) not occurring in our data set, and asked to annotate the component nouns for word sense with reference to sentences incorporating those NCs. In the main experiment, they were provided with 900 NCs including a target polysemous noun as either modifier or head noun, along with 50 sentences for each NC. The 50 sentences were meant to help the annotators identify the majority sense of the target noun in the given NC. See Section 7.2.4 for more details.

Particle semantics was annotated by the same two annotators. As we used the same particle semantic classification as Bannard (2003), we followed the procedure of Bannard in assigning the particle semantics. Before the actual experiment, the annotators were given samples taken from Bannard (2003) to familiarise themselves with the annotation task. During the actual experiment, they were given particles and their VPCs only.

After the different annotation tasks, we analyzed the results in terms of initial agreement. Despite training the annotators in advance, we achieved lower agreement than expected over some tasks. We hypothesise this was caused by: (a) the task itself being harder than expected (e.g. extra terms in NCs providing distraction due to syntactic and semantic ambiguity), (b) the provided evidences not being very helpful to the annotators (e.g. VPCs themselves could introduce ambiguity into the particle semantics annotation task). The results suggest that the annotation methodology needs to be carefully designed to achieve high agreement. The detail of the tasks, including the annotator agreement and difficulties in the annotation task, are described in each section.

2.8 Chapter Summary

In this chapter, we have described the linguistic properties of MWEs, provided a classification of English MWEs, and provided details of a number of key English MWE types.

First, we defined the following linguistic properties in the context of MWEs: *idiomaticity*, *non-identifiability*, *situatedness*, *figuration*, *single-word paraphrasability*, *proverbiality* and *prosody*. We identified *idiomaticity* as a primary defining property of MWEs, and described the relevance of the various properties to it. In particular, we subclassified idiomaticity according to the four areas of: lexico-syntactic, semantic, pragmatic and statistical idiomaticity. *Lexico-syntactic idiomaticity* was defined to be a mismatch in the syntax of the MWE relative to the properties of the simplex components. *Semantic idiomaticity* was defined to be (semantic) non-compositionality, i.e. a mismatch in the semantics of the MWE and that of its components. *Pragmatic idiomaticity* was defined to be the situatedness of an MWE, or association of the MWE with a particular situation. Finally, *statistical idiomaticity* was defined to occur when the frequency of the MWE is unusually high compared to that of its components or alternative phrasings of the same expression. From this, we then defined a *collocation* to be an MWE which was strictly *statistical idiomatic*.

Other properties we identified were *single-word paraphrasability*, *proverbiality* and *prosody*. *Single-word paraphrasability* is the ability of an MWE to be paraphrased with a single simplex word; *proverbiality* is the property of MWEs to represent a recurrent situation of particular social interest; and *prosody* relates to the observation that certain MWEs occur with abnormal stress patterns.

We also provided a detailed description of the syntax and semantics of the following MWE types: *noun compounds*, *verb-particle constructions*, *light-verb constructions*, *idioms* and *determinerless prepositional phrases*. *Noun compounds* are comprised solely of nouns and make up an $\bar{\text{N}}$. *Verb-particle constructions* are combinations of a verb and one or more particle. *Light-verb constructions* are made up of one of a small subset of verbs with bleached semantics, and a noun complement. *Idioms* are non-compositional MWEs which do not fall into any of the identified MWE types. We further sub-classified idioms into *decomposable* and *non-decomposable* idioms. Finally, *determinerless PPs* are made up of a preposition and a singular noun without a determiner.

Chapter 3

Statistical Frameworks and Related Work

3.1 Introduction

In this chapter, we will look at the underlying methods commonly used in statistical approaches to MWE extraction: co-occurrence properties, substitutability, distributional similarity, semantic similarity, recovering ellipsed predicates and linguistic properties. We will take a look at how these methods are used for computational tasks relating to MWEs, and weigh up the advantages and disadvantages of each approach. We will also look at prior approaches, and provide an overview and comparison of the methods used in this thesis.

3.2 Co-occurrence Properties

3.2.1 Overview of Co-occurrence Properties

The use of *co-occurrence properties* in modeling MWE involves analyzing the co-occurrence of the components of an MWE under the assumption that two or more words occur together with markedly high frequency iff they form an MWE. This basic approach forms the basis of a plethora of association measures and has been used extensively for collocation extraction (in the standard use of the term) (Choueka *et al.* 1983; Smadja 1993; Lin 1998b; Pearce 2001; Evert 2004; Pecina 2005). This property has been found to be highly effective for extracting statistically-marked MWEs such as *shock and awe* as their co-occurrence tends to have abnormally high frequency relative to the alternative ordering. (3.1) is a sample of high-frequency such binomials (relative to their alternative ordering) while (3.2) is a sample of binomials where both orderings have approximately the same frequency. This method can also be paired with analysis of alternative wordings for a given phrase in the form of substitutability (see Section 3.3). Note that when we say co-occurrence here we refer to the co-occurrence of the parts rather than co-occurrence with any specific context, which is the basis of distributional similarity in Section 3.4.

(3.1) MWEs: *black and white, by and large, salt and pepper, shock and awe*

(3.2) Non-MWEs: *blue and red, small and large, salt and sugar*

Note that the underlying mechanism driving co-occurrence is statistical idiomaticity, as most MWEs are statistically idiomatic to some degree. In (3.1), for example, the method can be seen to have extracted statistically-marked MWEs (*by and large*) as well as semantically- (*black and white*) and pragmatically-marked MWEs (*shock and awe*).

Co-occurrence properties are often measured by association measures such as pointwise/specific mutual information (Church and Hanks 1989), the Dice coefficient (Church and Hanks 1989), the student's t-test, Pearson's chi-square (Dunning 1993) and log likelihood (Dunning 1993). For a detailed overview of these association measures, see Pecina (2005).

This method is useful when the components combine together with markedly high frequency relative to the components, or alternatively relative to an alternative form of the same MWE. However, quantitatively measuring co-occurrence properties via a given association measure has its limitations. As most of the measures rely on lexicalized corpus frequencies, they are vulnerable to the effects of data sparseness. Additionally, highly-productive MWEs such as VPCs and NCs tend to have Zipfian distributions, with large numbers of low-frequency instances. Furthermore, it is often difficult to predict which association method will perform best over a given MWE type and corpus (Pecina 2005).

Co-occurrence properties have been used widely in tasks such as extracting collocation and MWEs (Smadja 1993; Grefenstette and Teufel 1995; Villavicencio *et al.* 2004; Baldwin 2005b; Fazly *et al.* 2005; Villada Moiron 2005; Pecina 2005; Widdows and Dorow 2005; Kan and Cui 2006), modeling the compositionality of MWEs (Bannard 2003; McCarthy *et al.* 2003; Venkatapathy and Joshi 2005; Fazly and Stevenson 2007; Kim and Baldwin 2007a), and classifying MWE semantics (Fraser 1976; Lapata and Keller 2004).

Below, we outline a representative selection of papers on the co-occurrence properties of MWEs, in the context of extraction, compositionality modeling and semantic classification tasks, respectively. Note that in some instances, the original research uses the term *collocation* in the broader sense of the term to mean MWE. In our description of the research, we will use the terms MWE and collocation as outlined in Chapter 2.

3.2.2 Co-occurrence Properties for Extraction

Smadja (1993) proposed the XTRACT system for extracting MWEs from raw text, building on a number of ideas from previous work (Choueka *et al.* 1983; Church and Hanks 1989). The basis for XTRACT is that the components of MWEs co-occur with unexpectedly high frequency, and also that they tend to occur in fixed word positions relative to each other (an assumption which clearly falls down with VPCs in the split configuration, e.g.).

The method is made up of three steps. The first (similar to Church and Hanks (1989)) is to extract binary MWE candidates within a 5-word window based on strength (frequency of collocation), spread (burstiness) and peakiness (identification of a particular word order/positioning which is notably more frequent than others). For example, for a given target word *takeover*, occurring with words $pill_{+2}$, $make_{+2}$ and $attempt_{+2}$, $attempt_{-1}$ (where $word_N$ is an occurrence of *word* N words to the left of the target word, considering each combination of word and position as a distinct data point), the method may filter out all but $attempt_{-1}$, corresponding to *takeover attempt* (i.e. *attempt* -1 words to the left = one word to the right of the target word).

The second step (similar to Choueka *et al.* (1983)) is to combine binary MWE candidates into multiple-word combinations and complex expressions. That is, from binary collocations extracted in the first stage, the method generates n-gram collocations from individual occurrences of the two words and analyzes the distribution of words and POS in surrounding context, and identifies any extra components which commonly co-occur with the elements of the bigram. For example, *chip stocks* may be expanded into *blue chip stocks*, and *price index* into *the consumer price index.*

The third step involves syntactic analysis of the binary or larger MWEs from the second step to ensure they follow constituent boundaries and correspond to common syntactic configurations, e.g. modifier-modifiee, subject–verb or verb–object.

In more recent work, Pecina (2005) tested a large number of co-occurrence-based extraction methods proposed in previous work in an MWE extraction task. The aim of the work was to empirically evaluate a comprehensive list of automatic MWE extraction methods using precision–recall curves, and to propose a new approach for combining individual extraction methods using supervised learning methods. Pecina used a total of 84 association measures based on occurrence frequencies (i.e. co-occurrence properties) over binary MWEs. As association measures, he used simple probabilities, mutual information and derived measures, statistical tests of independence, likelihood measures, and various heuristic association measures and coefficients. He also used context association measures based on syntactic and semantic units, with a more sound linguistic foundation. The final conclusion of this work was that the combination of multiple independent measures is superior to any one individual extraction method at MWE extraction.

Grefenstette and Teufel (1995) developed a method for extracting light verbs and their complements (i.e. *LVCs*) using co-occurrence properties. The basic idea behind this work is that the noun complements in LVCs are often deverbal (e.g. *proposal*), and that the distribution of nouns in PPs post-modifying noun complements in genuine LVCs (e.g. *(make a) proposal of marriage*) will be similar to that of the object of the underlying verb (e.g. *propose marriage*). Grefenstette and Teufel collected verbs and their nominalized forms, along with verb–object relations for the verbs and verb–noun–PP relations for the nouns, based on a low-level parser and heuristics.[1] From this, they selected the most common verb supporting the structure NP PP where the given nominalization heads the NP and the prepositional head of the PP is most similar to that of the underlying verb of the nominalization. In this case, therefore, multiple co-occurrences are considered (verb–noun and noun–preposition) to predict the light verb associated with a given nominalization.

Baldwin (2005b) employed several statistical tests to extract prepositional verbs (see Section 2.3). The main idea in this work is that the verb and preposition in prepositional verbs co-occur more frequently than for simple verb–preposition combinations. Baldwin proposed a number of unsupervised methods to extract prepositional verbs based on statistical tests such as chi-square and Dice's coefficient, as well as substitutability with highly frequent verbs and transitive prepositions (see Section 3.3). The method also adopted linguistic features of prepositional verbs, and demonstrated that co-occurrence properties were effective in the extraction task, but that the combination of all extraction method strategies was superior overall.

3.2.3 Co-occurrence Properties for Compositionality

McCarthy *et al.* (2003) proposed a method to measure the compositionality of English VPCs based on the intuition that compositional MWEs are more likely to occur in similar contexts to their component words, than is the case for non-compositional MWEs. In detail, McCarthy *et al.* used distributional similarity (see Section 3.4) and statistical tests to model the compositionality of English VPCs. First, the authors

[1] Note that a parser has been employed in several MWE extraction methods, including Baldwin (2005a) in the context of English VPC extraction. However, in Baldwin (2005a), the parser(s) are used extensively not only to extract VPC candidates but also to analyze the argument structure of the VPC, as described in Section 3.7.2.

identified VPC, verb and preposition instances from the output of the RASP parser, and from these calculated context vectors. They then calculated the distributional similarity between different combinations of VPCs and verbs, and used six different methods to estimate VPC compositionality. One such method was *overlap*, which is overlap in the top X neighbors of the VPC (not including the simple verb itself) and the same number of neighbors of the simplex verb. Another is *same particle − simplex* which is the number of neighbors in the top X which share the same particle as the VPC, minus the number of neighbors in top X for the simplex verb which share the same particle as the VPC.

In addition to distributional similarity, the authors employed several statistical tests to measure compositionality, both based on corpus statistics and dictionary occurrence. In evaluation, they found high correlation between the best of the distributional methods (**same particle − simplex**) and the human-annotated compositionality values, and that simple co-occurrence in the form of statistical tests performed very badly over the target task.

Venkatapathy and Joshi (2005) proposed a method for measuring the relative compositionality of a verb–noun pair such as *take place* or *feel safe*. Verb–noun pairs often occur with high frequency, making them suited to co-occurrence-based analysis.

The proposed methods are based on various types of collocation and context. The authors used five different co-occurrence tests, namely frequency, point-wise mutual information, least mutual information difference with similar collocations, distributed frequency of object and distributed frequency of object using the verb information. They also used distributional similarity based on the approach of Baldwin *et al.* (2003a) to model the compositionality of English MWEs. They evaluated the proposed methods using correlation, following the methodology of McCarthy *et al.* (2003).

The authors concluded that collocation features are better for measuring the relative compositionality of verb–noun pairs than distributional similarity, and that the correlation between the combined features and the human ranking was much better than that using individual features.

3.2.4 Co-occurrence Properties for Semantics

Lapata and Keller (2004) used co-occurrence properties in a variety of NLP tasks, including bracketing of compound nouns and interpreting compound nouns. The main motivation for this research was to evidence the usefulness of web data by employing it for probabilistic modeling. The probabilistic models they used for NC bracketing and interpretation were very simplistic, and based on simple co-occurrence of parts of the NC (in the first instance) and parts with different prepositions (in the second). That is, for the bracketing task, they tested 10 different probabilistic models integrating the frequencies of bracketed candidates (e.g. *((back up compiler) disk)* vs. *(backup (compiler disk)))*. For NC interpretation, they tested the method proposed in Lauer (1995) based on the co-occurrence of nouns with different prepositions (e.g. *night flight* paraphrased as *flight at night*). Their research demonstrated that simple web frequencies were highly successful when applied to these two (and other) tasks.

3.3 Substitutability

3.3.1 Overview of Substitutability

Substitutability is the ability to replace parts of MWEs with alternative lexical items, and involves comparison of the target MWE with anti-collocations (defined in Section 2.1). Also, this method is directly related to *single-word paraphrasability* described in Section 2.1.1. This approach is effective when parts of an MWE occur with unusually high frequency relative to lexical alternatives, i.e. their collocational association is high. In this thesis, we consider substitutability to be a subset of co-occurrence properties.

Substitutability can be applied to either compositional or non-compositional MWEs. Substitutability is closely related to anti-collocation, as when parts of the MWE are replaced, the new lexical items are generally no longer MWEs. Note that in substitutability, we always consider the whole MWE (in the form of the original or the anti-collocation), while in co-occurrence properties, we sometimes compare the whole to a variant word order, and sometimes compare the whole to its parts. Analysis

MWE	→	*Non-MWE*
frying fan	→	*frying* ***pot***
salt and pepper	→	*salt and* ***sugar***
many thanks	→	***several*** *thanks*
red tape	→	***yellow*** *tape*

Table 3.1: MWEs and Non-MWEs based on substitution

of substitutability tends to be based on the same inventory of statistical tests as for co-occurrence, as outlined in Section 3.2.1.

In generating substitution candidates, we often replace components of the original MWE with synonyms, sister words or antonyms, depending on the task and approach. This is based on the assumption of institutionalization, i.e. that a particular word combination has been established as an MWE to the exclusion of other plausible possibilities based on related words. Table 3.1 details examples where substitution leads to syntactically and/or semantically anomalous word combinations.

In Table 3.1, when parts such as *fan* and *many* are replaced with related words, the newly-formed word combinations (i.e. *frying pot* and *several thanks*, respectively) are no longer MWEs. Similarly, *yellow tape*, formed by substituting *red* with *yellow* in *red tape*, does not preserve the original meaning of "bureaucracy".

Substitutability can also be used to investigate the limits of productivity of MWEs such as VPCs and NCs. Despite various semantic restrictions, certain MWEs are highly productive. Hence, substitutability can be employed in order to construct new MWEs while maintaining the original "semantic collocation" (e.g. the same verb synset combined with the same particle). It is interesting to note it is possible to construct many NCs while preserving the same basic sense collocation, a point we return to in Section 7.1.

(3.3) *call up* → *phone/ring up* vs. **telephone up*

(3.4) *lemon juice* → *orange/fruit/lime juice*

In (3.3), *call up* is the basis for generating the VPCs *phone up* and *ring up*, but anomalously not *telephone up*, despite *telephone* being a lexical variant of *phone*. Starting with *lemon juice* in (3.4), we form the three NCs *orange juice*, *lime juice* and *fruit juice*, based on substituting *lemon* with a synonym, hypernym and sister word, respectively.

In a computational context, substitutability is broadly used to classify word combinations as MWEs or non-MWEs (Lin 1998b; Lin 1998d; Lin 1999; Pearce 2001). Substitutability is also applicable to the modeling of MWE compositionality (Bannard *et al.* 2003; Bannard 2003; McCarthy *et al.* 2003; Kim and Baldwin 2007a), the generation of MWEs with related semantics or compositionality (Stevenson *et al.* 2004; Baldwin 2005b; Turney 2005; Kim and Baldwin 2007b; Kim and Baldwin 2007c), and semantic classification (Villavicencio *et al.* 2004; Villavicencio 2005; Uchiyama *et al.* 2005).

3.3.2 Substitutability for Extraction

Lin (1999) proposed a method for classifying word combinations as non-compositional and compositional using the substitution method. The idea behind this work is that when a phrase is non-compositional, substitution candidates will tend to have markedly different frequencies of occurrence. For example, *red tape* occurs with much higher frequency than *yellow tape* or *orange tape*, indicating that it is non-compositional. On the other hand, *economic impact* has similar frequency to alternative wordings such as *financial impact* and *economic effect* and is hence predicted to be compositional.

As the source of the substitution candidates, Lin used a distributional thesaurus (Lin 1998a), which was pre-computed from the output of the Minipar dependency parser (Lin 1993). He also used the output of Minipar over a large-scale corpus to compute frequencies of different word combinations in particular syntactic configurations, from which he calculated the degree of association via a variant of point-wise mutual information. He compared the degree of association of the target word combination with substitution candidates via a Z-score, which provides an indication of

MWEs	*Anti-collocations*
*emotional **baggage***	*emotional **luggage***
***many** thanks*	***several** thanks*
***strong** coffee*	***powerful** coffee*

Table 3.2: Examples of MWEs and anti-collocations (Pearce 2001)

the relative differential in the association values. Only if the differential is high over all substitution candidates is the target word combination considered to be an MWE.

The study found that substitutability was a successful means of predicting the non-compositionality of word combinations.

Pearce (2001) proposed a method for extracting MWEs using substitution over WordNet. The motivation for the substitution method is that parts of compositional MWEs can be substituted with related words such as synonyms and hypernyms while maintaining the same basic semantics. Similar to Lin (1999), if the substitution candidates occur markedly less frequently than the original, it is an interpreted to be an indication that the original was an MWE. Table 3.2 illustrates examples of MWEs and corresponding ***anti-collocations*** generated by this method.

In Table 3.2, *emotional baggage* is an MWE whereas *emotional luggage* is not, despite *baggage* and *luggage* being synonyms. That is, in terms of the MWE properties described in Section 2.1.1, the MWEs in Table 3.2 are institutionalized, as indicated by their unusually high frequency relative to their anti-collocations.

In evaluation, Pearce classified the test instances into three classes: MWE, potential and unknown. The experimental results were promising, and demonstrated the power of the rich hierarchical structure of WordNet.

3.3.3 Substitutability for Semantic Classification

Turney (2005) proposed a method for measuring the relational similarity between a pair of nominal phrases, for use in analogical reasoning. For example, the noun pair *cat:meow* is analogous to the pair *dog:bark*, because both represent an animal and its sound. Likewise, *milk:drink* and *pie:eat* form a relational pair in which the relation would be of the type food and how to consume it.

The particular task Turney is interested in is the SAT test, where given a target noun pair such as *quart:volume* and a set of 5 candidate noun pairs, such as in (3.5), the task is to select the candidate noun pair that is most relationally similar to the target pair.

(3.5) *day:night*, *mile:distance*, *decade:century*, *friction:heat*, *part:whole*

In this case, the answer would be *mile:distance* on the basis that the first noun is a specific measurement of the second noun.

While the noun pairs are not in fact MWEs, they are closely related to NCs, and the methodology proposed by the author is closely related to methods used for interpreting NCs.

To measure the similarity between a giving combination of two noun pairs, the author employs substitution relative to the target noun pair A:B, replacing a word at a time based on the top 10 related words using synonymy, hypernymy and sister words. He then filters generated word pairs based on frequency, and measures the similarity of phrases based on clustering to confirm that they preserve the same relational semantics.

Two notable aspects of this research are that: (1) it is based on substitutability; and (2) it makes use of clustering and not classification, and as such does not attempt to resolve the exact relation between the nouns in a given pair.

Uchiyama *et al.* (2005) used the co-occurrence properties of Japanese compound verbs to predict their semantics. Japanese compound verbs are made up of a verb in the continuative form ($V1$) and an auxiliary verb ($V2$), as in *tabe-sugiru/eat too much*. Japanese compound verbs are highly productive and semantically ambiguous, and are subject to semantic constraints between the first verb and the second verb.

(3.6)–(3.8) show examples of Japanese compound verbs and a classification according to the semantics of the $V2$ (i.e. spatial, aspectual and adverbial), which also correspond to distinct translation strategies into English (as indicated). Note that the translation between Japanese and English has been carried out base on the fact that they have a semi-similarity due to their loose connection.

(3.6) **Spatial compound verbs:** $V2$ is translated as a verb in English.
nage-ageru ≡ throw (a ball) up
keri-ageru ≡ kick (a ball) up

(3.7) **Aspectual compound verbs:** $V2$ is translated as a particle in English.
yude-ageru ≡ finish boiling (vegetables)
musi-ageru ≡ finish steaming (vegetables)

(3.8) **Adverbial compound verbs:** $V2$ is translated as a adverb in English.
donari-ageru ≡ shout
odosi-ageru ≡ threaten

Uchiyama *et al.* (2005) proposed a novel machine learning method to disambiguate the semantics of $V2$, based on the co-occurrence of $V1$ and $V2$. The method is based on a matrix analysis of $V1$–$V2$ combinatorics. That is, the features used to classify a given combination of $V1$ and $V2$ are based on the semantic classes of each $V2'$ which co-occurs with $V1$, and each $V1'$ which co-occurs with $V2$, based on the row containing $V1$ and column containing $V2$. Figure 3.1 shows how to construct this feature vector.

3.4 Distributional Similarity

3.4.1 Overview of Distributional Similarity

Distributional similarity is a method for estimating semantic similarity based on the analysis of the contexts in which two lexical items are used. The basic idea behind this method was popularized by Firth (1957), and states that when two words are similar,

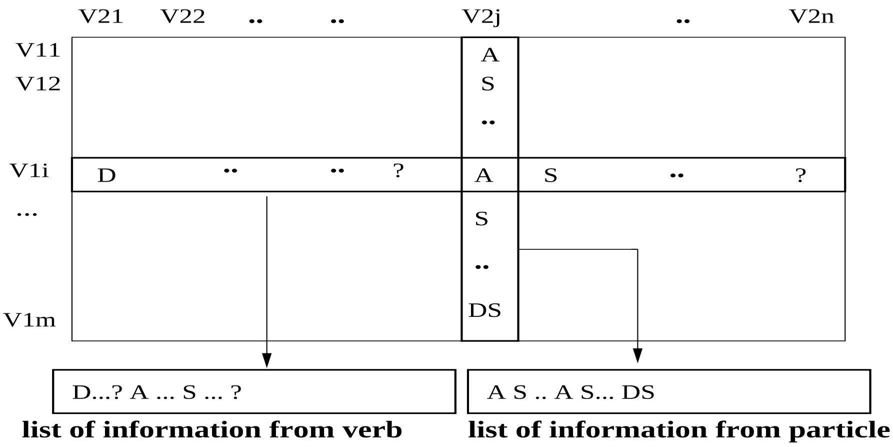

Figure 3.1: Co-occurrence-based feature representation (Uchiyama et al. 2005)

they will occur in similar contexts (i.e. their neighboring words within a word window will be similar). In the context of MWEs, distributional similarity is frequently used to compare the token occurrences of an MWE with the token occurrences of its components outside of the MWE. For example, when *kick the bucket* is used as an idiom, it may occur commonly with words such as *mourn*, *sad* and *bury*, while *kick* and *bucket* may occur commonly with very different words such as *water*, *accident* and *container*. This suggests that the semantics of *kick the bucket* differs from that of its parts, and that it is therefore a non-compositional MWE.

A common window size used to model contextual similarity is 25 words to either side of a given lexical item token. The similarity between the context vectors associated with two lexical items is commonly measured with *cosine similarity* (Salton *et al.* 1975).

(3.9) is an example of the idiom *kick the bucket*, where a context window of 5 words has been indicated via underlining; (3.10) is a literal usage of *kick the bucket* with its corresponding 5-word context window.

(3.9) The old man requested, “When I ***kick the bucket***, bury me on top of that mountain.”

(3.10) When we were about to enter the room, Kim accidentally ***kicked the bucket*** next to the door.

Comparing distributional similarity with the previous two methods, it is similar to co-occurrence properties in that it compares word combinations, with the big distinction that distributional similarity analyses the context of token occurrences of a given lexical item, whereas co-occurrence properties analyses the frequencies of components. Distributional similarity is a more powerful method in that there is greater scope for parameterization/reformulating in terms of: how the context window is defined, how token counts are translated into feature vectors, and how context vectors are compared. In the context of translating token counts into feature vectors, e.g. a considerable amount of work has been done on dimensionality reduction, such as with *latent semantic analysis* (LSA) (Landauer *et al.* 1998) to overcome data sparseness. Co-occurrence properties, on the other hand, are based fundamentally on token counts of components/re-orderings of the original lexical item, with the only place for innovation in the numeric interpretation of those numbers.

One way in which researchers have extended the basic distributional similarity method is by redefining the context window to look at the second-order co-occurrence of words. Here, rather than using the neighboring words of the target lexical item's neighboring words across multiple contexts as a direct representation of the target expression, the neighboring words of a specific token occurrence of the target expression are in turn modelled via their neighboring words. For example, let's assume that the target word *bank* has neighboring words *money*, *stock* and *savings* in a given context window. Rather than represent these directly as a 3-term (sparse) vector, we look to see what words each of them co-occurs with across the sum total of their usages. For example, *money* might co-occur with terms such as *banking* and *market* across all of its token occurrences, giving us a rich vector with which to present that one context term. We similarly generate individual vectors for the other two context terms and use the combination of the three to represent the original context. If we were then to compare the original token instance of *bank* with a single token instance of *financial institution*, say, although the immediate context words may not overlap, there is a good chance that the context vectors for each of the context words will. Second-order co-occurrence therefore provides a powerful mechanism for performing token-level analysis of context, e.g. in disambiguating individual occurrences of word sequences (such as *kick the bucket*) as either MWEs or simple compositional combinations.

The main weakness of distributional similarity is that it relies on large amounts of corpus data to operate effectively.

Distributional similarity has been employed to model the compositionality of MWEs (Schone and Jurafsky 2001; Bannard 2003; Baldwin *et al.* 2003a; Venkatapathy and Joshi 2005; McCarthy *et al.* 2007), to identify MWEs (Katz and Giesbrecht 2006), and to classify the semantics of MWEs (Stevenson *et al.* 2004).

3.4.2 Distributional Similarity for Compositionality

Bannard *et al.* (2003) used the distributional semantics of English VPCs to measure their compositionality and to model the contribution of the verb and particle in the overall semantics of the VPC. The basic idea behind this work is that if an MWE is compositional, then it will occur in the same lexical context as its components. The authors assumed that VPCs populate a continuum between fully compositional and fully non-compositional structures, which can be discretized according to the contribution of each of the verb and particle (as described above).

Bannard *et al.* used four different classification methods: the method of Lin (1999), the context space model of Schutze (1998), a substitution method, and distributional similarity between each of the components and the overall VPC. The authors found that the mixed methods performed best, and the third and fourth methods outperformed the first and second methods. Significantly, this paper showed that distributional semantics can be applied to the analysis of particles and MWEs, where previous work had tended to focus exclusively on simplex content words.

Baldwin *et al.* (2003a) used distributional similarity to compare MWEs with their components, focusing on NCs and VPCs. The proposed method was based on the context space model of Schutze (1998), which incorporates LSA.[2] (3.11) illustrates the outputs of the method for the VPCs *cut out* and *cut off* with the component verb *cut*. Based on the similarity values, the model is predicting that *cut out* is more compositional than *cut off*.

(3.11) similarity(*cut*, *cut out*) = .433 vs. similarity(*cut*, *cut off*) = .183

To evaluate their method, the authors compared the predicted similarity between VPCs and their component verbs, and NCs and their component nouns, with similarities generated from WordNet. They found a weak correlation between the two, and once again demonstrated the potential for distributional semantics to model the compositionality of MWEs.

3.4.3 Distributional Similarity for Identification

Katz and Giesbrecht (2006) used second-order distributional similarity to identify non-compositional MWEs (i.e. idioms) in German. As outlined above, the intuition behind the method is that non-compositional MWEs will co-occur with significantly different words to their components, as can be captured in their second-order co-occurrence. For example, when *kick the bucket* is used as an idiom (meaning "die"), then the context words around it will be very different to those for both *kick* and *bucket* in isolation, whereas when it is used compositionally, it will be more similar in usage to the component words. To measure the similarity between German MWEs and their components, they once again employed the context space model of Schutze (1998).

Figure 3.2 shows the context vector associated with an idiomatic usage of *den loffel Abgeben* (corresponding to *kick the bucket* in German, and literally meaning "to eat the spoon"), compared to each of its component words vs. a paraphrase for the MWE (*sterben*, meaning *die*). Here, therefore, the prediction would be that the usage is idiomatic rather than literal.

The authors concluded that it is possible to identify MWEs in context using distributional similarity.

[2] `http://infomap.stanford.edu/`

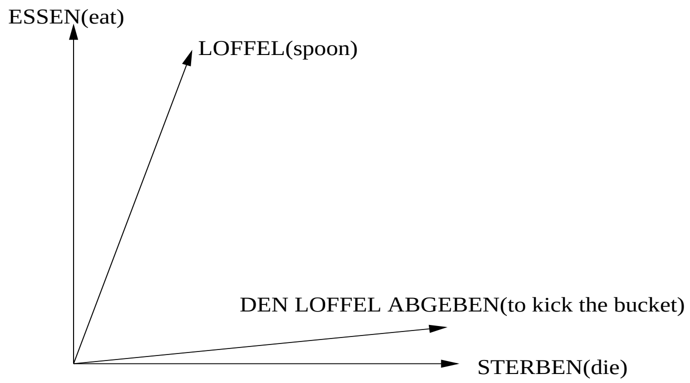

Figure 3.2: Distributional semantics of the German idiom *den loffel Abgeben* (Katz and Giesbrecht 2006)

3.5 Semantic Similarity

3.5.1 Overview of Semantic Similarity

Semantic similarity uses a direct model of the semantics of the parts (and possibly the whole) of an MWE to measure compositionality. The underlying assumption is that with compositional MWEs, the semantics of the whole MWE can be decomposed into the semantics of the parts. For example, we would expect the semantics of *add up* to be closely related to that of *add*, and to a lesser degree *up*. Similarly, we would expect *sum up* to have similar properties to *add up* based on them both incorporating the same particle and *sum* and *add* being similar (Villavicencio 2005; Kim and Baldwin 2007a). Compared with *distributional similarity*, the main difference is that *semantic similarity* employs the semantics from the MWE parts whereas *distributional similarity* uses the information from the target word's neighboring words. See *distributional similarity* in Section 3.4 for comparison.

One application of semantic similarity is in the interpretation of MWEs. That is, when the corresponding components of a pair of MWEs are similar (such as with *sum up* vs. *add up* above), it is generally the case that they have a similar interpretation, e.g. via a semantic relation. This gives rise to a method for interpreting MWE semantics (Rosario and Marti 2001; Moldovan *et al.* 2004; Kim and Baldwin 2005; Nastase *et al.* 2006; Girju 2007; Kim and Baldwin 2007c). (3.12) and (3.13) show how to interpret the semantic relations in NCs using semantic similarity.

(3.12) modifier = FRUIT, head noun = LIQUID → SR = MAKE
e.g. *apple juice, orange juice, grapes nectar, chocolate milk*

(3.13) modifier = LOCATION, head noun = LIQUID → SR = LOCATION
e.g. *Fuji apple, California orange, Bordeaux wine*

In (3.12) and (3.13), despite different combinations of lexical items, NCs such as *apple juice* and *chocolate milk* are predicted to have the same SR of MAKE, as the modifier and head noun, respectively, have similar semantics.

The advantage of this method comes from the ability to use existing similarity measures for simplex words (e.g. based on lexical resources such as WordNet or CoreLex) to accurately interpret MWEs, although such methods are limited by the coverage of the underlying similarity measures (and hence the coverage of any base lexical resources).

This method is employed in computational tasks such as interpreting NCs (Rosario and Marti 2001; Moldovan *et al.* 2004; Kim and Baldwin 2005; Girju 2007), and modeling the compositionality of MWEs (Piao *et al.* 2006; Kim and Baldwin 2007a).

3.5.2 Semantic Similarity for Compositionality

Piao *et al.* (2006) proposed the use of semantic similarity to test the compositionality of MWEs. The basic idea was that there is a correlation between compositionality and the relative similarity between the semantics of an MWE and its parts. To model semantics, they used a field taxonomy based on the Lancaster English Semantic Lexicon,[3] which is derived from the McArthur (1981) Longman Lexicon of Contemporary English. The lexicon has 21 major semantic fields, further divided into 232 subcategories, and contains nearly 55,000 single-word entries and over 18,800 MWEs entries. (3.14) shows the semantic hierarchy for *food & farming*, e.g.

[3] www.comp.lancs.ac.uk/ucrel/usas/

(3.14) **F: FOOD & FARMING**

Food $\supset$ Drinks $\supset$ Cigarettes & Drugs $\supset$ Farming & Horticulture

The paper proposes a novel method for measuring the semantic distance between an MWE and its component words based on hand-tagged hierarchical semantic information. Piao *et al.* evaluated the proposed method over 89 MWEs, scoring each from 0 (least compositional) to 10 (completely compositional). They used Spearman's correlation coefficient to measure the correlation between the automatic and manual rankings, and claimed results comparable to human performance.

3.5.3 Semantic Similarity for Semantic Relations

Rosario and Marti (2001) used semantic similarity to interpret NCs in the medical domain based on the CUI and MeSH medical ontologies. CUI is part of UMLS (Humphreys *et al.* 1998), and is comprised of three resources: a meta-thesaurus, semantic network and specialist lexicon. MeSH is one of the source vocabularies of UMLS, where concepts are identified by unique concept identifiers in hierarchical structures. (3.15) shows how the hierarchical classes for the modifier and head noun in *flu vaccination*.

(3.15) *flu vaccination* → SR = PURPOSE

- CUIs: C0016366 | C0042196
- MeSH: D4.808.54.79.429.154.349 | G3.770.670.310.890

To classify NCs which are manually tagged with medical classes, Rosario and Marti used a neural network. They found that the domain-specific lexical hierarchy successfully captured the semantic similarity of NCs to interpret SRs, but also that coverage is a significant bottleneck for the medical domain.

Moldovan *et al.* (2004) used word sense collocations in NCs to interpret SRs. The basic idea is that when NCs have the same sense collocation, i.e. corresponding components are semantically similar, then they most likely have the same SR. For example, *car factory* and *automobile factory* have identical sense collocation, and as such, have the same SR MAKE.

Moldovan *et al.* proposed a probabilistic model called SEMANTIC SCATTERING to implement their sense collocation-based interpretation method. Semantic scattering is based on Equation 3.16 and Equation 3.17.

$$P(r|f_{ij}) = \frac{n(r, f_{ij})}{n(f_{ij})} \tag{3.16}$$

where f_{ij} is a simplified feature pair $f_i f_j$ (i.e. the word senses of the modifier and head noun in an NC) and r is the semantic relation. The preferred SR r^* for the given word sense combination is that which maximizes the probability:

$$\begin{aligned} r^* &= \text{argmax}_{r \in R} P(r|f_{ij}) \\ &= \text{argmax}_{r \in R} P(f_{ij}|r) P(r) \end{aligned} \tag{3.17}$$

In evaluation, the authors found that their method performed at about 43% accuracy over open domain NCs.

3.6 Recovering Ellipsed Predicates

3.6.1 Overview of Ellipsed Predicates

Ellipsed predicates, taken from Pustejovsky (1995), are a means of interpreting MWEs via an underlying predicate without surface realization. For example, *GM car* means “car made by GM”, which we can represent as the directed predicate `make` (i.e. the formal interpretation is `make’(GM’,car’)`, where GM is the maker and car is the “makee” or item of manufacture).

The assumption with *ellipsed predicates* is that the implicit predicate can be recovered from analysis of the semantics of the components. Normally, the predicate is verbal (such as *make* above), and the analysis takes place via the semantics of the associated verb(s).[4] Depending on the set of predicates, it is possible for a single MWE to be associated with multiple ellipsed predicates. For example, *family car* can be interpreted as `possess'(family',car')` (meaning "(the) family possesses a car") or alternatively `belong'(car',family')` (meaning "(the) car belongs to a family"), noting the different ordering of the components relative to the two predicates.

Ellipsed predicate recovery is an effective way of interpreting MWEs, as it allows for the analysis of word combinations via overt templates or paraphrases. That is, it allows us observable with unobservable MWEs. For example, Lauer (1995) showcased ellipsed predicate to recover the hidden verb semantics between nouns in noun compounds. The major barrier, however, is in finding a suitable set of verbs for each semantic relation.

In the context of computational tasks, ellipsed predicates are used primarily as a means of semantic interpretation, e.g. of NCs (Levi 1978; Vanderwende 1994; Lapata 2002; Kim and Baldwin 2006b; Nakov and Hearst 2006), but they can also be used to measure the compositionality of MWEs (Cook and Stevenson 2006).

3.6.2 Ellipsed Predicates for Semantic Relation Interpretation

Levi (1978) studied SRs in NCs and argued that there exists a small set of discrete semantic relations between the components of NCs, unlike Finin (1980), e.g. who argued that the number of semantic relations is unbounded. Levi proposed an interpretation method via verb semantics where NCs are said to be derived via one of two syntactic processes: (1) predicate nominalization, or (2) predicate deletion. In the first case, she defined 9 predicates: CAUSE, HAVE, MAKE, USE, BE, IN, FOR, FROM and ABOUT. She also argued that NCs can be derived by predicate deletion via various transformation, as shown in (3.18).

(3.18) *virus infection* → CAUSE

[4] Exceptions are *coach player*, where the predicate is non-verbal and symmetric (`equative'(coach',player') = equative'(player',coach')`), and *night flight*, where the predicate is once again non-verbal (`time'(flight,night')`).

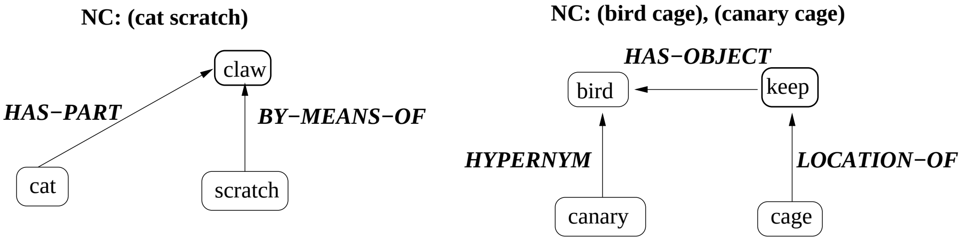

Figure 3.3: Interpretation via verb semantics (Vanderwende 1994)

1. *infection* (virus causes infection)
2. *infection* (infection is caused by virus) → Passive
3. *infection* (infection is virus-caused) → Compound adjective formation
4. *infection* (which is virus-caused) → Relative clause formation
5. *infection virus-caused* → WH-be deletion
6. *virus-caused infection* → Predicate proposing
7. *virus infection* → RDP deletion: CAUSE$_2$
8. *viral infection* → Morphological adjectivalization

Levi's research provided the foundation for much of the subsequent work on SRs in NCs.

Vanderwende (1994) interpreted SRs in NCs via verb semantics, and the relation between the components and different verbs. The intuition behind this work is much the same as that of Levi (1978), that NCs can be interpreted using SRs and ellipsed predicates (i.e. verb semantics). The verb semantics associated with different SRs were extracted from definition sentences in a machine-readable dictionary. Also, Vanderwende used hypernymy over the components to select the most plausible SR. Figure 3.3 outlines her approach to NC interpretation with an example.

The examples *cat scratch* and *bird cage* are interpreted via the verbs *claw* and *keep*, respectively. Also, since *canary* is a hyponym of *bird*, *canary cage* is also interpretable from *bird cage*.

This study was one of the first successful instances of an automated method based on verb semantics being used to interpret NCs.

Lapata (2002) proposed a probabilistic framework for interpreting the SRs in *compound nominalizations*, a proper subset of NCs. In compound nominalizations, the head noun is derived from a verb and the modifier is interpretable as an argument of this underlying verb. For example, in the case of *child behavior*, *behavior* is derived from *behave*, and *child* can be interpreted as the subject of this predicate. Note that with general-purpose SRs, there is generally a fixed set of predicates associated with each SR, whereas with compound nominalizations, the predicate is defined by the head noun and thus a more open set.

Lapata assumed that in compound normalizations, the modifier can be interpreted as either the subject (e.g. (3.19)) or the object (e.g. (3.20)) of the underlying verb form of the head noun.

(3.19) **SUBJECT**

- *child behaves* → *child behavior*
- *government refuses* → *government refusal*

(3.20) **OBJECT**

- *love car* → *car lover*
- *synthesize sound* → *sound synthesizer*

The basic model proposed by Lapata is:

$$\begin{aligned} RA(rel, n_1, n_2) &= \log_2 \frac{P(OBJ|n_1, n_2)}{P(SUB|n_1, n_2)} \\ P(rel|n_1, n_2) &\approx \frac{f(v_{n2}, rel, n_1)}{\sum_i f(v_{n2}, rel_i, n_1)} \end{aligned} \tag{3.21}$$

where n_1 is the modifier, n_2 is the head noun, v_{n2} is the underlying verb form of the head noun, rel is a semantic relation (either subject or object), and $f(v, rel, n)$ is the frequency of occurrence of n as the rel (subject or object) of v in corpus data.

In this paper, the author conducted a number experiments, focusing particularly on smoothing based on distributional evidence, statistical smoothing and extra contextual information. The author achieved a final accuracy of 86.1% over the British National Corpus.

In addition to this work, Grover *et al.* (2004) applied the same technique in the biomedical domain over an expanded set of relations (to allow for prepositional objects), integrating the paraphrase representation of Lauer (1995). Similar to the original research, they examined the influence of smoothing methods to overcome data sparseness. Later, Nicholson and Baldwin (2005) proposed an algorithm for both identifying compound nominalizations based on a chunker, and interpreting them via corpus evidence. Their trial is a replication of the 3-way classification of Grover *et al.* (2004), but they additionally attempted to predict the preposition and evaluated their method over open domain data.

3.7 Linguistic Properties

3.7.1 Overview of Linguistic Properties

A final method is the use of *linguistic properties* to analyze MWEs. The assumption is that certain linguistic properties correlate with MWE compositionality, as well as particular syntactic and semantic types. Linguistic properties are generally considered in combination with other computational methods, rather than forming a standalone computational method, as they tend to suffer from data sparseness, i.e. have high precision but low recall over a given set of MWE types.

(3.22)–(3.24) show an example of using linguistic properties to extract VPCs from corpus data.

(3.22) **Particle position**

- *lead (the donkey/them) on*
- ∗*draw (inner strength/it) on*

(3.23) **Particle modifiability**

- *pick (back/right back) up the pencil*
- ∗*draw (back/right back) on (inner strength/it)*

(3.24) **Nominalization**

- *feedback, backup*
- ∗*drawon*

In (3.22), we are able to rule out *draw on* as a VPC based on the fact that the preposition is not compatible with the split word order. In (3.23), we are once again able to rule out *draw on* as a VPC on the grounds that it is not possible to modify the preposition *on* with *back* or *right back*. Finally, in (3.24), the fact that *draw on* does not nominalize to ∗*drawon* is suggestive of the fact that it is not a VPC (although in this last case, it is a sufficient but not necessary condition on VPCs).

Linguistic properties often take the form of highly-specific syntactic features of MWEs, either in context or at the type-level. While linguistic properties can rely on context, they differ from distributional similarity in that they are very selective and fine-tuned to particular construction types.

As shown in the examples above, linguistic properties can provide very reliable features for identifying or otherwise classifying MWEs. Their main drawbacks are that they do not easily generalize, and rely on the occurrence of very particular usages/contexts.

Example tasks where linguistic properties have been employed are MWE extraction (Baldwin and Villavicencio 2002; Baldwin 2005a; Nakov and Hearst 2005), identification (Patrick and Fletcher 2004; Van Der Beek 2005; Kim and Baldwin 2006a) and semantic classification (O'Hara and Wiebe 2003; Stevenson *et al.* 2004; Cook and Stevenson 2006).

3.7.2 Linguistic Properties for Extraction

Baldwin (2005a) used linguistic properties (among many other features) to extract fully-specified English VPCs from raw text corpora. The basic approach in this work was to boost the precision of more general-purpose features with linguistic properties, based on the output of various preprocessors (e.g. parsers and chunkers). Specific linguistic properties used by the author were analysis of the word order of the object NP and preposition in transitive VPC candidates, particularly when the object NP is pronominalized. That is, transitive English VPCs undergo the particle alternation, producing the joined and split word orders. Also, pronominal objects must be expressed in the split configuration, and manner adverbs cannot occur between the verb and particle in either transitive or intransitive VPCs. These properties are illustrated in (3.25)–(3.27).

(3.25) **Particle alternation**

- `joined`: *Kim handed in the paper.*
- `split`: *Kim handed the paper in.*

(3.26) **Pronominalized object word order**

- *hand it in.*
- ∗*hand in it.*

(3.27) **Manner adverb word order**

- *Hand it in promptly.*
- ?∗*Hand it promptly in.*

The task in Baldwin (2005a) was undertaken with no assumptions about corpus annotation, using only information from pre-processors such as a part-of-speech tagger, chunker and RASP. It also evaluated VPC extraction as both shallow and deep lexical acquisition tasks, that is either as the simple task of determining what combinations of verb and preposition can form a VPC, or as the harder task of determining what combinations of verb and preposition can form an intransitive and transitive VPC (e.g. for the purposes of a deep grammar lexicon).

The proposed method was tested over three corpora (Brown Corpus, Wall Street Journal and British National Corpus), and linguistic properties were shown to provide valuable evidence in the extraction task, especially over low-frequency VPCs.

Nakov and Hearst (2005) employed linguistic properties in a probabilistic model for bracketing NCs with 3 or more terms in the medical domain, building on the work of Marcus (1980) and Lauer (1995).[5] Nakov and Hearst extended this earlier work by integrating linguistic features into their model, based on analysis of surface features in web data as illustrated in (3.28)–(3.30).

(3.28) **Dash or hyphen**

- *left bracketing*: *cell-cycle analysis* → *((cell-cycle) analysis)*
- *right bracketing*: *donor T-cell* → *(donor (T-cell))*

(3.29) **Genitive ending or possessive marker**

- *left bracketing*: *brain stem's cells* → *((brain stem) cells)*
- *right bracketing*: *brain's stem cells* → *(brain (stem cells))*

(3.30) **Capitalization**

- *left bracketing*: *Plasmodium vivax Malaria* → *((Plasmodium vivax) Malaria)*
- *right bracketing*: *brain Stem cells* → *(brain (Stem cells))*

Based on these and other features, Nakov and Hearst (2005) developed an unsupervised method for NC bracketing using chi-square, and achieved 89.34% bracketing accuracy.

[5] We return to survey related work on bracketing in Section 5.1.1.

3.7.3 Linguistic Properties for Semantics

Cook and Stevenson (2006) classified particle semantics in English VPCs using linguistic properties of VPCs. The authors observed the following facts: (1) semantically-similar verbs combine with a similar range of target particles; and (2) what verbs can combine with a given particle is an indicator of the semantics of the target particle. Based on these observations, the authors used slot features to encode the relative frequencies of the syntactic slots (i.e. subject, direct and indirect object, and object of a preposition), and particle features to encode the relative frequency of the verb co-occurring with high frequency particles.

Cook and Stevenson classified the particle *up* into four different semantic classes, as illustrated in (3.31)–(3.34).

(3.31) **Vertical:** *The price of gas jumped up.*

(3.32) **Goal-oriented:** *The deadline is coming up quickly.*

(3.33) **Completive:** *Finish up your dinner quickly.*

(3.34) **Reflexive:** *Roll up the curtain.*

The paper also used co-occurrence features to classify particle semantics. In their experiments, the authors found that the method based on linguistic properties outperformed that using word co-occurrence features in the task of classifying particle semantics.

3.8 Chapter Summary

In this chapter, we have provided an overview of the statistical approaches most commonly used in modeling MWEs. In detail, we presented six different statistical approaches: co-occurrence properties, substitutability, distributional similarity, semantic similarity, ellipsed predicate recovery and linguistic properties. For each approach, we described the basic ideas and reviewed a small sample of related work in the context of MWE tasks such as *MWE identification/extraction*, *semantic classification* and *interpreting semantic relations* (as defined in Section 1.2).

Co-occurrence properties is the use of the frequencies of the components or an alternative word ordering of a given MWE to analyze whether the MWE is statistically marked. These frequencies are then plugged into a variety of statistical tests to measure the cohesion among the components. In addition to being effective at detecting statistical idiomaticity, it has been employed in modeling compositionality.

Substitutability is analysis of the effect on the MWE of substituting components with related terms. The frequency or other features of the target MWE are then compared to the *anti-collocations* generated through substitution. Substitutability is considered a special case of co-occurrence properties and is particularly suited to the modeling of compositionality, and allows for more fine-grained analysis than co-occurrence properties. This method is often employed to extract MWEs and to classify the semantics of MWEs.

Distributional similarity involves analysis of the context of use of different lexical items. Based on the assumption that similar words will occur in similar contexts, the more similar the context vectors of a given pairing of lexical items, the more similar they are predicted to be. Distributional similarity can be calculated based on the contexts of use of a lexical item across multiple usages, or alternatively based on the context vectors of each of the context words surrounding a lexical item in a given usage (second-order co-occurrence). Unlike co-occurrence properties, distributional similarity is based on co-occurring words rather than component words of the MWE. It has been used to model compositionality, classify semantics and to identify MWEs.

Semantic similarity is the process of modeling the semantics of the whole via the semantics of the parts, notably in comparing corresponding components of a pairing of MWEs and inferring that the MWEs are similar in the instance that the components are similar. This approach is effective for interpreting the semantics of MWEs, and has been applied to the tasks of semantic interpretation and compositionality modeling.

Ellipsed predicate recovery attempts to model the semantics of MWEs by recovering an implicit predicate. The underlying assumption is that certain MWEs can be interpreted by way of a predicate over its components. By mapping semantic relations onto verbal predicates, it becomes possible to interpret MWEs by looking for attested instances of particular predicate-argument structures. This approach has been employed for the semantic interpretation (particularly, noun compounds) of MWEs.

Linguistic properties of MWEs can be used to model MWEs, e.g. based on the output of a parser. They tend to be highly construction-specific, and are high-precision but low-recall. As a result, they tend to be combined with other approaches rather than form standalone methodologies. This approach has been applied to MWE extraction and semantic classification.

Finally, from analysis of statistical approaches over various computational tasks, we have attempted to gauge their utility for the purposes of our research, as we shall see in ensuing chapters.

Chapter 4

Resources

This chapter describes the resources used in our research, including corpora, lexical resources, dictionaries and software. The resources vary in coverage and usage, and not only provide fundamental knowledge to understand context, but are also in some cases used to evaluate the proposed models.

4.1 Corpus

4.1.1 British National Corpus

The British National Corpus (Burnard 1995)[1] is a 100 million word balanced corpus of written and spoken English developed between 1991 and 1994. It contains principally modern British English over a variety of subject areas, genres and registers, as well as styles.

90% of the British National Corpus is written language (or articles) extracted from various sources such as local and national newspapers, special periodicals and journals, academic books, fiction works, memorandums and school essays. The remaining 10% is spoken language (or dialogues), including a large amount of unscripted informal conversations recorded by volunteers selected from different age, regional and social backgrounds. The topics also vary from business or government meetings to

[1] `http://www.hcu.ox.ac.uk/BNC`

radio shows and phone conversations. In this thesis, we used the written portion of the British National Corpus, POS-tagged with fnTBL and pre-parsed with RASP (see below), in the tasks of noun compound interpretation and VPC identification.

4.1.2 Wall Street Journal

The Wall Street Journal section of the Penn Treebank was first published in 1989 and contains around one million words of written American English from the Wall Street Journal, annotated in Treebank II style (Marcus *et al.* 1993). The Treebank II style is designed to allow for the extraction of simple predicate-argument structure, and includes part-of-speech tags as well as phrase structure information. Compared with the British National Corpus, its scale is relatively small, and its domain coverage is limited to financial texts.

In this thesis, we used the Wall Street Journal as POS-tagged by fnTBL and parsed by RASP in the tasks of noun compound interpretation and VPC identification.

4.1.3 Brown Corpus

The Brown Corpus of Standard American English was the first computer-readable, balanced corpus of English, and published in 1961. It was compiled by W.N. Francis and H. Kucera of Brown University. The corpus consists of 500 texts, each made up of just over 2,000 words. The texts were sampled from 15 different text categories such as general fiction, the press and humour, to provide a standard reference for corpus linguistic research. The samples represent a wide range of styles and varieties of prose. The version of the Brown Corpus we use in this research is the subset contained in the Penn Treebank, which is annotated in Treebank II style (similarly to the Wall Street Journal above).

Nowadays, the Brown Corpus is considered relatively small and slightly outdated compared to other corpora such as British National Corpus and Wall Street Journal. However, it is still used alongside other corpora and provided the inspiration for other corpora such as British National Corpus, American National Corpus and Kolhaur

Corpus. In the thesis, we used the section of the Brown Corpus contained in the Penn Treebank (Marcus *et al.* 1993), pre-parsed by RASP, in the tasks of noun compound interpretation and VPC identification.

4.2 Lexical Resources

4.2.1 WordNet

WordNet (Fellbaum 1998)[2] is a large-scale lexical database of English developed at Princeton University under the direction of George A. Miller. It groups English words (nouns, verbs, adjectives and adverbs) into sets of synonyms called synsets. WordNet provides short, general definitions for each synsets and records various conceptual-semantic and lexical relations between pairings of synsets. Initially, it was developed to produce a combination of dictionary and thesaurus to support automatic text analysis and NLP applications. It contains both simplex words and multiword expressions. As described in Table 4.1, the total of all unique noun, verb, adjective, and adverb lexical items is 155,327, contained in 117,597 unique synsets (based on version 2.1). Many lexical items have a unique synset classification within a given syntactic category, but are described under more than one syntactic category.

WordNet has been used in various natural language processing tasks such as lexical semantics (McCarthy *et al.* 2004; Moldovan *et al.* 2004; Nastase *et al.* 2006), PP-attachment (Kim and Baldwin 2006a) and question answering (Prager and Chu-Carroll 2001; Hermjakob *et al.* 2002), and has become a mainstream language resource in NLP. The current version of WordNet is 3.0, although most of our experiments were carried out using WordNet 2.1 as it was the current version at the time.

Table 4.1 summarizes the total number of words and multiword expressions contained in WordNet 2.1.

[2] `http://wordnet.princeton.edu/`

POS	# of lexical entries	# of MWEs
noun	117,097	59,876
verb	11,488	2,777
adjective	22,141	571
adverb	4,601	117
all	155,327	63,341

Table 4.1: Composition of WordNet 2.1

4.2.2 CoreLex

CoreLex (Buitelaar 1989)[3] is a noun classification based on a unified approach to systematic polysemy and the semantic underspecification of nouns, and is derived from WordNet 1.5. While WordNet 1.5 provides around 60,000 different noun synsets, CoreLex collapses these into a concise set of 126 coarse-grained classes, by taking into account systematic polysemy and underspecification.

It contains 45 basic CoreLex types, systematic polysemous classes and 39,937 tagged nouns. The semantic types are underspecified representations based on the generative lexicon theory (Pustejovsky 1995). From a seed collection of hand-tagged nouns, a probabilistic tagger was built to classify unknown nouns (not in CoreLex) and to identify context-specific and new interpretations. The classification algorithm is centered around the computation of a similarity based on the Jaccard coefficient, that compares lexical items in terms of their shared attributes (linguistic patterns acquired from domain-specific corpora). In this thesis, we used this lexical resource to interpret semantic relations in NCs.

The following example shows the `fod` class and its subclasses.

(4.1) **fod**

- **atr fod(attribute)** *chocolate, vintage, wine*
- **fod(food)** *ale, beefsteak, chili*

[3] `http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html`

- **fod frm(form)** *doughnut, suds*
- **fod nat(natural_object)** *berry, java, milk*
- **fod qui(quantify_indefinite)** *cocktail, syllabub, toast*
- **fod sta(state)** *blackheart, pickle, stew*
- **fod sub(substance)** *nectar, paste, wafer*

4.2.3 The Moby Thesaurus

The Moby Thesaurus[4] is the largest and most comprehensive English thesaurus, originally based on Roget's thesaurus. The second edition contains 30,000 root words, and 2.5 million synonyms and related words. It is provided in a simple ASCII format suitable for viewing, editing, and automatic parsing. We used this dictionary to interpret semantic relations in NCs.

4.3 Tools

4.3.1 WordNet::Similarity

We used the open-source WordNet::Similarity package (Patwardhan *et al.* 2003)[5] to compute word similarities. WordNet::Similarity is developed at the University of Minnesota, and provides various methods to measure the similarity or relatedness between a pair of concepts or word senses. It contains implementations of a variety of comparison methods, of three basic types: similarity, relatedness and random.

The similarity methods are categorized into two groups: path-based (LCH (Leacock and Chodorow 1998) and WUP (Wu and Palmer 1994)) and information-content based (RES (Resnik 1995), JCN (Jiang and Conrath 1997), and LIN (Lin 1998c)), as summarized in Figure 4.1.

[4] `http://www.dcs.shef.ac.uk/research/ilash/Moby/mthes.html`

[5] `www.d.umn.edu/~tpederse/similarity.html`

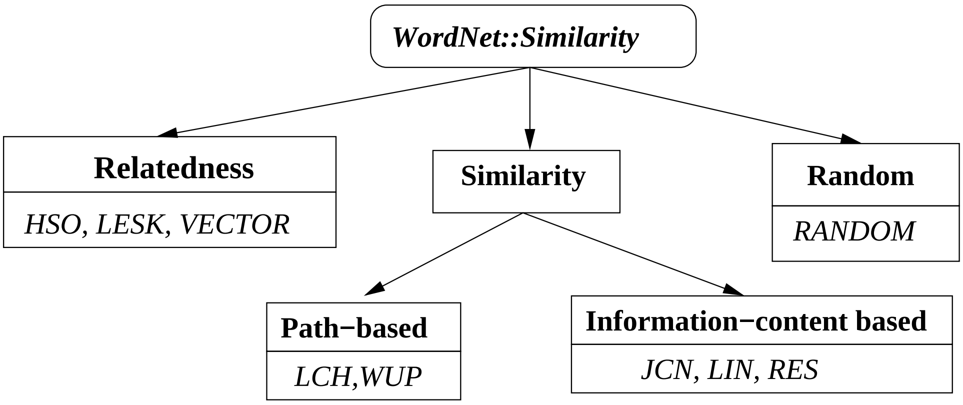

Figure 4.1: Classification of methods in WordNet::Similarity

Path-based methods compute lexical similarity based on the shortest path between two target synsets based on the WordNet *is-a* hierarchy. The difference between LCH and WUP is in the calculation of path length. LCH calculates the path length between two target concepts (c_1 and c_2) based on Equation 4.2:

$$Similarity_{lch}(c_1, c_2) = -\log\left(\frac{p}{2 \times depth}\right) \tag{4.2}$$

where p is the number of nodes in the shortest path connecting c_1 and c_2, and *depth* is the maximum depth of WordNet hierarchy.

WUP, on the other hand, is based on the path length to the root node from the least common subsumer (LCS) of the two target concepts (c_1 and c_2). The LCS is defined as that concept at greatest depth in the WordNet hierarchy that subsumes both c_1 and c_2. The calculation of similarity is based on Equation 4.3

$$Similarity_{wup}(c_1, c_2) = \frac{2 \times p3}{p1 + p2 + 2 \times p3} \tag{4.3}$$

where $p1$ and $p2$ are the number of nodes on the path from c_1 to c_2 respectively and $p3$ is the number of nodes on the path between LCS and root.

RES, JCN and LIN augment the calculation of path length with the information content (IC) of the LCS, calculated as follows:

$$IC(c) = -\log \frac{freq(c)}{freq(root)} \tag{4.4}$$

where $freq(c)$ is the frequency of a given concept c, and $freq(root)$ is the frequency of the root of the hierarchy.

RES calculates the similarity of two concepts by the information of their LCS:

$$similarity_{res} = IC(lcs(c_1, c_2)) \tag{4.5}$$

JCN is an extension of RES, where the path length between the two concepts is included in the calculation, based on:

$$similarity_{jcn} = IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)) \tag{4.6}$$

LIN is a further variant of RES, based on the Dice coefficient:

$$Similarity_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{4.7}$$

The relatedness measures use additional relations such as *has-part, is-made-of* and *is-an-attribute-of* in addition to the *is-a* relation. There are three relatedness measures: HSO (Hirst and St-Onge 1998), LESK (Banerjee and Pedersen 2003) and VECTOR (Patwardhan 2003). HSO is based on path similarity, and takes into consideration sequences of lexical relations connecting synsets in the WordNet hierarchy that are likely to be indicative of word-level (rather than sense) relatedness. LESK is based on the weighted word overlap of different pairings of synset glosses, over a variety of relation types.

VECTOR is a corpus-based measure. Each word is represented as a multi-dimensional vector of co-occurring words. The similarity of a word pair is measured by the cosine similarity of the two vectors. In Equation 4.8, $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ are the vectors of the two target words:

$$Relation_{vector}(c_1, c_2) = \frac{\overrightarrow{v_1} \times \overrightarrow{v_2}}{||\overrightarrow{v_1}|| \times ||\overrightarrow{v_2}||} \tag{4.8}$$

Finally, RANDOM measures similarity by random assignment.

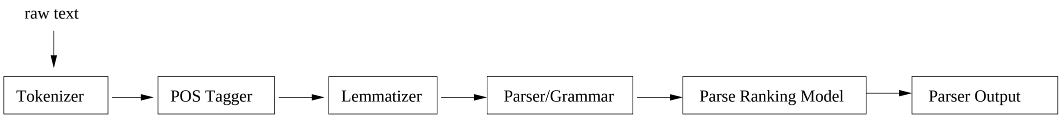

Figure 4.2: RASP pipeline

4.3.2 TiMBL

TiMBL (Daelemans *et al.* 2004)[6] is a memory-based learner. Memory-based learning is based on the classical k-nearest neighbor (i.e. k-NN) approach to classification. To deal with discrete data, large numbers of examples and large numbers of attributes of varying relevance, TiMBL makes extensive use of indexes rather than a typical flat file found in traditional k-NN systems. TiMBL includes a variety of similarity metrics such as overlap and dot product, feature weighting metrics such as information gain and chi square, and distance weighting metrics such as inverse distance and inverse linear distance. Also, it handles user-defined example weighting. For all our experiments, we use version 5.1 of TiMBL.

4.3.3 Robust Accurate Statistical Parsing (RASP) parser

The RASP parser (Briscoe and Carroll 2002)[7] is a parsing toolkit jointly developed at the University of Sussex and University of Cambridge. It integrates and extends several strands of research on robust statistical parsing and automated grammar and lexicon induction, and is based on a tag sequence grammar. The RASP output contains dependency tuples derived from the most probable parse, each of which includes a label identifying the nature of the dependency (e.g. subject, direct object), the head word of the modifying constituent, and the head of the modified constituent. In addition, each word is tagged with a part-of-speech tag from which it is possible to determine the valence of any prepositions. Figure 4.2 depicts the pipeline architecture of RASP.

Features of RASP which make it particularly attractive in this research are its robustness and dependency output format. We use RASP to automatically parse the British National Corpus, Wall Street Journal and Brown Corpus, from which we extract features for use in noun compound interpretation and VPC identification. The relevant parameters we used to run RASP are: "1" as the maximum number of parses that should be produced for each sentence, "grammatical relations" as the output form, "no use" of verb subcategorization frame probabilities, and "turning off" the use of a list of phrasal verbs that normally allows more accurate identification of verb-particle constructions.

4.4 Chapter Summary

In this chapter, we have described various resources employed in modeling MWEs. We use three corpora—the British National Corpus, Brown Corpus, and Wall Street Journal—which we have POS-tagged, chunked and dependency-parsed with RASP. We also employ lexical resources to model the lexical semantics of lexical items, namely WordNet, CoreLex and Moby Thesaurus. Finally, we use WordNet::Similarity to measure word similarity, and the TiMBL memory-based learner to build classifiers.

[6] `http://ilk.uvt.nl/software.html`

[7] `http://www.informatics.susx.ac.uk/research/nlp/rasp/project.html`

Chapter 5

MWEs and Semantic Similarity

In this chapter we will take a look at methods that can be used to interpret NCs and model the compositionality of VPCs. MWEs are often formed from parts which have the same or similar semantics as previously-observed data. Hence, *semantic similarity* has often been used for modeling tasks relating to MWEs, such as NC interpretation (Rosario and Marti 2001; Moldovan *et al.* 2004; Girju *et al.* 2005; Nastase *et al.* 2006), modeling compositionality (Bannard *et al.* 2003; McCarthy *et al.* 2003) and the identification of MWEs (Lin 1998b; Lin 1999; Baldwin *et al.* 2003a).

In this chapter, we will present the motivation behind the use of semantic similarity, and describe and evaluate our approach to interpreting NCs and modeling the compositionality of VPCs based on semantic similarity.

5.1 Noun Compound Interpretation via Constituent Similarity

Noun compound (NC) interpretation has a long history in both theoretical and computational linguistic research (Downing 1977; Levi 1978; Finin 1980; Vanderwende 1994; Barker and Szpakowicz 1998; Rosario and Marti 2001; Lapata 2002; Moldovan *et al.* 2004; Kim and Baldwin 2005; Girju 2007). Conventionally, *semantic relations* (SRs) are used to describe how the components of a given NC interact with each other. Semantic relations specify the underlying relation between a head noun and its modifier(s) in the form of a (directed) binary predicate. For example, *family car* involves the semantic relation POSSESSOR, which describes the ownership of *car* (head noun) by a *family* (modifier). On the other hand, *GM car* contains the relation MAKE, which represents the fact that *GM* (modifier) produces the *car* (head noun). Various sets of semantic relations have been proposed in linguistics for interpreting NCs. In this thesis, our focus is not on deriving a new set of semantic relations, but to compare different interpretation techniques relative to a fixed set of semantic relations. As such, we take a set of semantic relations which is relatively well-established in NLP research (Barker and Szpakowicz 1998) and accept it as suitable for our purposes without modification. Table 5.1 details the 20 semantic relations defined by Barker and Szpakowicz (1998). Note that some examples, such as *solar system* and *fast computer*, are not noun compounds but are included from the description in the original paper.

SRs have been proposed as a means of solving or simplifying many problems faced by NLP applications, including question-answering, machine translation, information extraction and summarization (Cao and Li 2002; Baldwin and Tanaka 2004; Lauer 1995; Venkatapathy and Joshi 2006).

However, despite the huge effort invested in interpreting the semantic relations in NCs, there is still a relative lack of automated approaches. Also, much of the prior work has been built on specific assumptions. Prior to describing our methods, we first briefly present previous research on NC interpretation.

5.1.1 Past Research on Noun Compound Interpretation

The computational tasks related to NCs are: (1) disambiguating the syntax of NCs with 3 or more terms; (2) defining the semantic relations found in NCs; (3) automatically interpreting NCs relative to a predefined set of SRs; and (4) disambiguating the word senses of component nouns to use in automatic interpretation. In this section, we review the first three tasks, and leave the last (i.e. word sense disambiguation in NCs) for Section 7.2.

Relation	***Definition***	***Example***
AGENT	n_2 is performed by n_1	*student protest, band concert, military assault*
BENEFICIARY	n_1 benefits from n_2	*student price, charitable compound*
CAUSE	n_1 causes n_2	*printer tray, flood water, film music, story idea*
CONTAINER	n_1 contains n_2	*exam anxiety, overdue fine*
CONTENT	n_1 is contained in n_2	*paper tray, eviction notice, oil pan*
DESTINATION	n_1 is destination of n_2	*game bus, exit route, entrance stairs*
EQUATIVE	n_1 is also head	*composer arranger, player coach*
INSTRUMENT	n_1 is used in n_2	*electron microscope, diesel engine, laser printer*
LOCATED	n_1 is located at n_2	*building site, home town, solar system*
LOCATION	n_1 is the location of n_2	*lab printer, desert storm, internal combustion*
MATERIAL	n_2 is made of n_1	*carbon deposit, gingerbread man, water vapour*
OBJECT	n_1 is acted on by n_2	*engine repair, horse doctor*
POSSESSOR	n_1 has n_2	*student loan, company car*
PRODUCT	n_1 is a product of n_2	*automobile factory, light bulb, color printer*
PROPERTY	n_2 is n_1	*elephant seal, blue car, big house, fast computer*
PURPOSE	n_2 is meant for n_1	*concert hall, soup pot, grinding abrasive*
RESULT	n_1 is a result of n_2	*storm cloud, cold virus, death penalty*
SOURCE	n_1 is the source of n_2	*chest pain, north wind*
TIME	n_1 is the time of n_2	*winter semester, morning class, late supper*
TOPIC	n_2 is concerned with n_1	*computer expert, safety standard, horror novel*

Table 5.1: The semantic relations used in this research to interpret noun compounds (n_1 = modifier, n_2 = head noun)

Disambiguating Syntactic Ambiguity: Bracketing

The task of disambiguating the syntax of NCs with 3 or more terms is called **NC bracketing**.

(5.1) ((computer science) department) [**left bracketing**]

(5.2) (linguistic (graduate program)) [**right bracketing**]

(5.1) and (5.2) show the two bracketing options for a ternary (i.e. 3-ary) NC (`left bracketing` and `right bracketing`, respectively). Generally, only one of the two bracketing options has a plausible interpretation, but there are examples, such as (5.3) and (5.4), where both bracketing options have plausible interpretations and context is needed to disambiguate.

(5.3) ((glass window) cleaner) [**left bracketing**]

(5.4) (glass (window cleaner)) [**right bracketing**]

In (5.3), *glass window cleaner* is interpreted as a cleaner of glass windows, whereas in (5.4) it is interpreted as a window cleaner made of glass, e.g.. Note that the SRs for the two interpretations differ: *glass window* would conventionally be interpreted as MATERIAL and *(glass window) cleaner* as PURPOSE, while *window cleaner* would conventionally be interpreted as OBJECT and *glass (window cleaner)* as MATERIAL.

To disambiguate the syntactic structure of ternary or higher order NCs, Marcus (1980) proposed the ***adjacency model***. That is, with a given ternary NC (N1 N2 N3), the model compares the relative frequency of (N1 N2) and (N2 N3) in binary NC data. In the case that (N1 N2) is the most frequent, the method selects a left bracketing analysis, and in the case that (N2 N3) is the most frequent, it selects a right bracketing analysis.

Lauer (1995) argued that the ***dependency model*** is a more accurate representation of the underlying syntax, and also that it is empirically superior to the adjacency model. For a given NC (N1 N2 N3), the dependency model compares the frequency of (N1 N2) with that of (N1 N3), based on the observation that the two dependency tuples in the left bracketing case are (N1 N2) and (N2 N3), while those in the right bracketing case are (N2 N3) and (N1 N3); as the dependency (N2 N3) is common to both, we can ignore its relative plausibility. Thus, in the instance that (N1 N2) is more frequent than (N1 N3), the model prefers a left bracketing analysis, and in the converse case, the model prefers a right bracketing analysis. Nakov and Hearst (2005) took the adjacency and dependency models and added various features to achieve an accuracy or 89.34%, which is the highest performance achieved to date. Both models are clearly susceptible to the effects of data sparseness. That is, without sufficient data, the models do not work well.

Identifying the Semantic Relations in Noun Compounds

Prior research has sought to identify the SRs in NCs from either a linguistic perspective (Levi 1978; Finin 1980; Sparck Jones 1983; Downing 1977) or a computational perspective (Vanderwende 1994; Barker and Szpakowicz 1998; Rosario and Marti 2001; Moldovan *et al.* 2004).

In pioneering work, Levi (1978) defined complex nominals as expressions that have a head noun preceded by one or more modifying nouns or denominal adjectives, and proposed nine SRs for non-opaque (i.e. compositional) compounds. Finin (1980) countered Levi's earlier work in claiming that the number of discrete SRs needed to interpret NCs was infinite, citing the impact of various pragmatic factors on the semantics of NCs. Sparck Jones (1983) built on earlier work by Downing (1977) in claiming that the SRs in NCs can be described only in terms of tendencies or preferences, and not absolutes. She further demonstrated that context can affect the interpretation. For example, (5.5) and (5.6) (adopted from Copestake and Lascarides (1997)) show that there is scope to arrive at distinct interpretations for *cotton bag.*

(5.5) (a) Mary sorted her clothes into various bags made from plastic.

(b) She put her skirt into the cotton bag.

(5.6) (a) Mary sorted her clothes into bags made of various materials.

(b) She put her skirt into the cotton bag.

Despite *cotton bag* occurring in identical (b) sentences, (5.5) evokes the interpretation of the *bag* being used for *cotton(s)*, representing the PURPOSE SR, while (5.6) evokes the interpretation of a *bag* made of *cotton*, representing the MATERIAL SR.

Different researchers have proposed varying sets of SRs for interpreting NCs, based on a variety of approaches. Vanderwende (1994) defined SRs based on *WH*-questions. Barker and Szpakowicz (1998) developed 20 SRs in a bottom-up fashion over task-oriented data, and Moldovan *et al.* (2004) proposed 32 SRs for use in open-domain paraphrase-based interpretation. Rosario and Marti (2001) identified 36 domain-specific SRs for the biomedical domain. Nastase *et al.* (2006) defined 30 SRs, along with 5 superclasses, in an attempt to overcome the problems of fine-granularity and unbalanced distribution. Ó Séaghdha (2007) designed a set of SRs with careful annotation guidelines, with the intention of achieving greater class balance and higher inter-annotator agreement, based on the set of SRs proposed by Levi (1978).

The quest to define a commonly agreed-upon set of SRs remains unsolved due to a number of problems: (1) the granularity of SRs, (2) the coverage of SRs over data from different domains, and (3) the class distribution of SRs. Smaller sets of SRs tend to be hard to work with due to their coarse-granularity (Levi 1978; Vanderwende 1994). Larger sets of SRs tend to fit the data better but are associated with ambiguity and have skewed class distribution (Finin 1980; Rosario and Marti 2001; Moldovan *et al.* 2004). Also, pragmatic effects lead to disagreements in SR labelling in- and out-of-context (Downing 1977; Sparck Jones 1983; Copestake and Lascarides 1997).

Interpreting Noun Compounds

Vanderwende (1994) was an early attempt to interpret the semantics of NCs automatically based on a discrete set of SRs. She used semantic information automatically extracted from analyzing definitions in an online dictionary, and interpreted the NCs through the semantics of verbs corresponding to each relation. One drawback was that the system employed hand-written rules, which meant that it wasn't fully automatic and incurred high setup costs in terms of manual labor and time.

Barker and Szpakowicz (1998) used a semi-automatic method for NC interpretation in a technical repair domain. More recent work on this task has led to several methods for automatic NC interpretation with minimal human effort (Fan *et al.* 2003), using a noun taxonomy.

Some research on NC interpretation has focused particularly on compound nominalizations (Isabelle 1984; Hull and Gomez 1996; Lapata 2002; Grover *et al.* 2004), that is NCs where the head noun is a deverbal noun such as *animation* (derived from *animate*). Lapata (2002) proposed a fully automatic method for interpreting compound nominalizations, based on the assumption that the modifier is interpretable as either the subject (e.g. *child behavior*) or object (e.g. *car lover*) of the base verb expressed by the head noun. Grover *et al.* (2004) and Nicholson and Baldwin (2005) extended this work in considering a wider selection of interpretation types, based around preposition selection.

Other previous work has focused on specific domains. Rosario and Marti (2001) used hierarchical tagged nouns from biomedical texts, and classified them according to a domain-specific set of SRs using neural networks. Grover *et al.* (2004) performed compound nominalization interpretation over biomedical data. Nakov and Hearst (2006) also focused on the biomedical domain, using verb semantics based on a web corpus.

Moldovan *et al.* (2004) and Girju *et al.* (2005) proposed methods for open-domain NC interpretation using the pairing of word senses of the component words. Later, Girju (2007) integrated cross-lingual features from five Romance languages into the same method and showed the effectiveness of cross-lingual information. Kim and Baldwin (2005) used a method based on nearest-neighbor classification over the union of senses of the modifier and head noun (see Section 5.1.3). Kim and Baldwin (2006b) proposed a method for interpreting noun compounds via verb semantics. Ó Séaghdha and Copestake (2007) used context, word and relation similarity to classify NCs. Nulty (2007) investigated the effectiveness of three different learning algorithms for NC interpretation at the token level.

There has also been research on NC interpretation for languages other than English. Johnston and Busa (1996) used qualia structure from the generative lexicon theory to interpret NCs in Italian. Utsuro *et al.* (2007) learned the dependency relations of Japanese compound functional expressions by projecting and analyzing their dependency relations through a machine learning technique. Zhao *et al.* (2007) used paraphrase patterns and web statistics to perform Chinese nominalization interpretation, similarly to Lapata and Keller (2004).

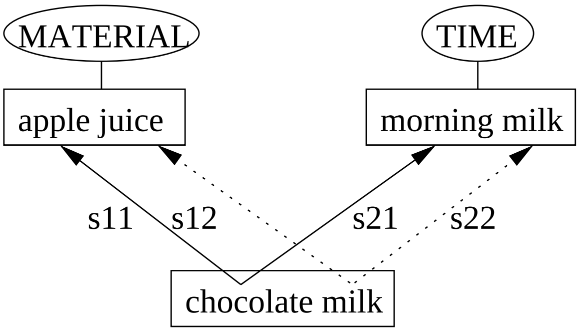

Figure 5.1: Example of constituent similarity

	Training noun	Test noun	S_{ij}	CombSim
n_1	apple	chocolate	0.71	**0.59**
n_2	juice	milk	0.83	–
n_1	morning	chocolate	0.27	0.27
n_2	milk	milk	1.00	–

Table 5.2: WordNet-based similarities for the component nouns in the training and test instances

5.1.2 Motivation for the Constituent Similarity Method

The constituent similarity method for NC interpretation is based on the observation that NCs which contain similar words in corresponding positions tend to have the same semantics. To illustrate how the method works, let us consider *chocolate milk* as a test instance, which we will attempt to interpret based on the training instances *apple juice* and *morning milk*. The SR of *apple juice* is MATERIAL while that of *morning milk* is TIME. As shown in Figure 5.1, we compare the test instance to each of the training instances by calculating the word similarity between the nouns in corresponding positions in the test and each of the training instances, i.e. by comparing the modifier noun to each of the modifier nouns in the training instances, and the head noun to each of the head nouns in the training instances.

Table 5.2 shows the similarity of each constituent (modifier, or n_1, and head

	Training noun	Test noun	S_{ij}	multipledSim
n_1	personal	loan	0.32	0.27
n_2	interest	rate	0.84	–
n_1	bank	loan	0.75	**0.63**
n_2	interest	rate	0.84	–

Table 5.3: The effects of polysemy on the similarities between nouns in the training

noun, or n_2) between the test and each of the training instances. Note that *CombSim* is the product of the component word similarities. The word similarities, S_{ij}, are computed based on the `WUP` measure, as implemented in WordNet::Similarity (see Section 4.3.1). Based on this, the combined similarity of *chocolate milk* with *apple juice* is 0.59, and that for *morning milk* is 0.27 (see below for details). Since the similarity with *apple juice* is higher, the SR for *chocolate milk* is resolved to material, the correct prediction in this case.

Unlike methods which use explicit sense information (e.g. Moldovan *et al.* (2004)), and which are hence susceptible to the unavoidable noise associated with word sense disambiguation, word similarity makes no direct use of sense information. Instead, it models word similarity by the union of the senses of each word, and averages across the sense pairings. As a result, our approach is not directly exposed to any sense ambiguity of the constituent words, and as our constituent representation remains at the word level, we are unexposed to the effects of data sparseness associated with sense-level representations.

The ability of our method to deal with the effects of word sense ambiguity can be seen in Table 5.3, where the training NCs are *personal interest* and *bank interest* (corresponding to SRs possessor and cause/topic, respectively), and the test NC is *loan rate*. Note that both training instances contain the head noun *interest*, but with different semantics: "a diversion that occupies one's time and thoughts" in the case of *personal interest*, and "a fixed charge for borrowing money" in the case of *bank interest*. Despite this, our approach resolves the SR for the test instance *loan*

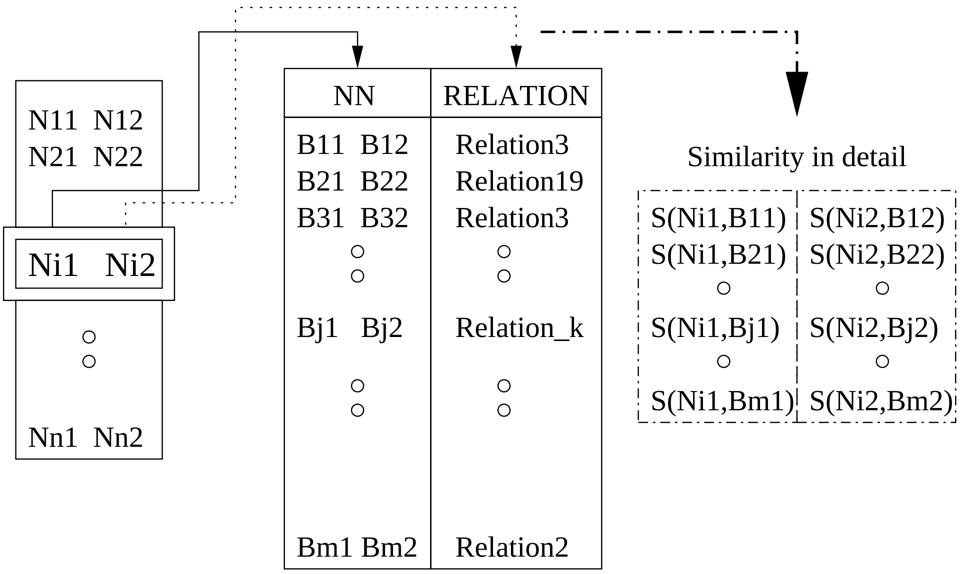

Figure 5.2: Similarity between the i_{th} NC in the test data and j_{th} NC in the training data

rate correctly as CAUSE/TOPIC based on the semantic similarity between the different modifier pairings, i.e. the fact that *loan* is more similar to *bank* than *personal.*

5.1.3 Details of the Constituent Similarity Method

Figure 5.2 shows the architecture of the constituent similarity method for interpreting NCs, where *(N_{i1} N_{i2})* is a test instance and each *(B_{j1} B_{j2})* is a training instance. $S(N_{i1}, B_{j1})$ and $S(N_{i2}, B_{j2})$ are the similarity between modifiers and head nouns, respectively, for *(N_{i1} N_{i2})* and *(B_{j1} B_{j2})*

The first step is to compute the similarities between the pair of modifiers and pair of head nouns for a given test and training instance ($S(N_{i1}, B_{j1})$ and $S(N_{i2}, B_{j2})$). The second step is to compute the combined similarity of modifiers and head nouns, as either a simple product (Equation 5.7) or harmonic mean (Equation 5.8).

Formally, S_A is the similarity between NCs $(N_{i,1}, N_{i,2})$ and $(B_{j,1}, B_{j,2})$:

$$S_A((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) = \frac{((\alpha S1 + S1) \times ((1-\alpha)S2 + S2))}{2} \tag{5.7}$$

where $S1$ is the modifier similarity $(S(N_{i,1}, B_{j1}))$ and $S2$ is head noun similarity $(S(N_{i,2}, B_{j2}))$; $\alpha \in [0, 1]$ is a weighting factor.

S_B is an analogous similarity function, based on the harmonic mean of the two component similarities:

$$S_B((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) \quad = \quad 2 \times \frac{(S1 + \alpha S1) \times (S2 + (1-\alpha)S2)}{(S1 + \alpha S1) + (S2 + (1-\alpha)S2)} \tag{5.8}$$

The final SR is determined by rel:

$$\begin{aligned} rel(N_{i,1}, N_{i,2}) \quad &= \quad rel(B_{m,1}, B_{m,2}) \qquad (5.9)\\ &\text{where } m = \text{argmax}_j S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) \end{aligned}$$

5.1.4 Data Collection for Constituent Similarity Method

We evaluate our proposed method under various experimental settings. The first experiment was run over 2-term NCs, and the second over 3-term NCs. In our third experiment, we checked the relative contribution of the head noun and modifier in different SRs, and in the fourth we used bootstrapping to expand the set of training instances. In the final experiment, we tested an extension of the basic method to incorporate the k-nearest neighbor algorithm (MacQueen 1967).

In order to perform the experiments, we first collected 2- and 3-term NCs from the Wall Street Journal, based on simple POS tag sequence data. Given the large number of NCs, we excluded proper nouns since the coverage of proper nouns in WordNet is poor, meaning that we cannot compute the word similarity for those nouns.[1] Based on this methodology, we collected a total of 2,169 (2,347 w/ multiple SRs) 2-term NC types and 1,571 (1,644 w/ multiple SRs) 3-term NC types. Two trained human annotators then individually bracketed the 3-term NCs, and also tagged all the NCs for SRs. Note that the human annotators considered the 3-term NCs as a single 2-term

[1] We additionally have the problem of determining the extent of multiword named entities, e.g. *North Carlton house*, which we would want to treat as a 2-term rather than 3-term NC.

NC based on the bracketing information. For example, ((N1 N2) N3) is considered as (N2 N3) while (N1 (N2 N3)) is considered as (N1 N3). Annotators were allowed to annotate a given NC with multiple SRs in instances of genuine ambiguity. Any disagreements in the original annotations were resolved based on discussion between the annotators. SR pairs which were particularly contentious and had lower agreement were SOURCE and CAUSE, PURPOSE and TOPIC, and OBJECT and TOPIC. The initial agreement over the SR annotation task was 52.31% for the 2-term NCs and 49.28% for the 3-term NCs. When instances were tagged with multiple SRs at least of which matched the other annotator's tag, it was treated as a point of agreement. Note that the inter-annotator agreement over 3-term NCs is much lower than that over 2-term NCs. We hypothesize that the reason for this was the syntactic ambiguity of 3-term NCs: despite the NCs being provided with manual bracketing, the annotators seemed to find it harder to assign SRs to the 3-terms NCs.

Finally, for evaluation, we randomly selected 50% of the data as our test instances and the remaining 50% as our training instances. A detailed breakdown of the test and training instances is presented in Table 5.4.

Note that in Table 5.4, A are instances which are annotated with both a unique SR (by one annotator) and multiple SRs (by the second annotator), and M are instances which are annotated with multiple SRs by both annotators. The number of unique test and training instances was 1,081 and 1,088 for 2-term NCs, and 785 and 786 for 3-term NCs, respectively.

5.1.5 Experiment (I): 2-term NCs

Our first experiment is to test the reliability of our proposed NC interpretation method over 2-term NCs. The baseline for this experiment is a majority class method, i.e. tagging all test instances as TOPIC.

Table 5.5 shows that the WUP method, using the S_A multiplicative method of combination, provides the highest NC interpretation accuracy. Using the harmonic mean to calculate the combined similarity (S_B), the LCH method has nearly the same performance as WUP. Among the four measures of similarity used in this first

Relation	*2-term NCs*				*3-term NCs*			
	Test		Training		Test		Training	
	A	M	A	M	A	M	A	M
AGENT	10	1	5	0	9	0	7	1
BENEFICIARY	10	1	7	1	2	0	3	0
CAUSE	54	5	74	3	21	0	18	0
CONTAINER	13	4	19	3	13	1	7	2
CONTENT	40	2	34	2	23	0	18	0
DESTINATION	1	0	2	0	0	0	1	0
EQUATIVE	9	0	17	1	1	0	2	1
INSTRUMENT	6	0	11	0	2	0	3	0
LOCATED	12	1	16	2	3	0	5	0
LOCATION	29	9	24	4	19	0	27	0
MATERIAL	12	0	14	1	10	0	11	0
OBJECT	88	6	88	5	22	6	26	3
POSSESSOR	33	1	22	1	25	4	21	6
PRODUCT	27	0	32	6	27	1	26	1
PROPERTY	76	3	85	3	33	0	43	0
PURPOSE	159	13	161	9	89	7	95	6
RESULT	7	0	8	0	3	0	4	0
SOURCE	86	11	99	15	61	0	44	1
TIME	26	1	19	0	19	0	24	0
TOPIC	465	24	447	39	438	16	437	15
TOTAL	1163	82	1184	96	820	35	822	36

Table 5.4: The distribution of semantic relations in 2-term and 3-term noun compounds

experiment, the path-based similarity measures have higher performance than the information content-based methods over both similarity combination methods.

Compared to prior work on the automatic interpretation of NCs, our method achieves relatively good results. Rosario and Marti (2001) achieved about 60% performance over the medical domain. Moldovan *et al.* (2004) used a word sense disambiguation system to achieve around 43% accuracy interpreting NCs in the open domain. Our accuracy of 53% compares favourably to both of these sets of results, given that we are operating over open domain data.

Basis	Method	S_A	S_B
Human annotation	Inter-annotator agreement	52.30%	
Majority class	Baseline	43.00%	
Path-based	WUP	**53.30%**	51.50%
	LCH	52.90%	**52.30%**
Information content-based	JCN	46.70%	43.50%
	LIN	47.40%	42.10%
Relatedness	LESK	42.44%	41.63%
	VECTOR	39.22%	38.85%
Random	RANDOM	21.83%	22.18%

Table 5.5: Accuracy of NC interpretation for the different WordNet-based similarity measures over 2-term NCs

We provide a comparative evaluation between Moldovan *et al.* (2004) and the constituent similarity method over the SemEval-2007 data in Section 7.1.

5.1.6 Experiment (II): 3-term NCs

In this experiment, we test our approach over 3-term NCs. To the best of our knowledge, these are the first reported results for NC interpretation over 3-term NCs. Note that the relative weight of the modifier and head noun is equal, i.e. 0.5.

The experiment is divided into two parts. In the first sub-experiment, we test our method using all three components of the NCs, and multiply all three component similarities together using S_A. Note that despite using all three components, our goal is to interpret only the top-level dependency tuple, not both 2-term NCs that make up the 3-term NC.

In the second sub-experiment, we use 2-term NCs extracted from the 3-term NCs based on gold-standard bracketing information. That is, (N2 N3) is extracted from ((N1 N2) N3), and (N1 N3) is extracted from (N1 (N2 N3)), and it is only these two components that are used by our method.

Tables 5.6 and 5.7 show the performance of our method over 3-term NCs, using

Basis	Method	S_A
Human annotation	Inter-annotator agreement	49.28%
Majority class	Baseline	55.80%
Path-based	WUP	48.28%
	LCH	54.65%
Information content-based	JCN	55.80%
	LIN	**57.71%**
Relatedness	LESK	51.97%
Random	RANDOM	31.72%

Table 5.6: Accuracy of NC interpretation for the different WordNet-based similarity measures over 3-term NCs

all three components in the first case and only a single 2-term NC extracted from the 3-term NC in the second case. Comparing the two NC representations, we find that our method performs better when it has access to all three components of the NC. This suggests that all the NC components contribute to determine the SR to some degree. Interestingly, whereas the path-based methods were the strongest performers in the 2-term NC case, with 3-term NCs, information content-based methods perform the best. While error analysis did not unearth a clear reason for this reverse trend, we suspects that the differing level of information contained in 2- and 3-term NCs is captured differently by the two families of semantic similarity measure. Overall, the results for the 3-term NCs were below those for the 2-term NCs.

5.1.7 Experiment (III): The Relative Contribution of the Components

We also studied the relative contribution of the head noun and modifier in different SRs. We started with the expectation that the contribution of the constituents would vary across different SRs. Table 5.8 shows the predicted contribution of the head noun and the modifier in different SRs.

In Table 5.8, *elephant seal* is interpreted as PROPERTY, where the NC is a hyponym

Basis	Method	S_A	S_B
Human annotation	Inter-annotator agreement	49.28%	
Majority class	Baseline	55.80%	
Path-based	WUP	43.18%	43.95%
	LCH	**44.84%**	**44.33%**
Information content-based	JCN	40.89%	32.82%
	LIN	41.15%	38.60%
Relatedness	LESK	42.73%	42.17%
Random	RANDOM	26.62%	26.37%

Table 5.7: Accuracy of NC interpretation for the different WordNet-based similarity measures over 3-term NCs

Relative contribution of modifier/head noun	Relation	Example
modifier < head noun	PROPERTY	*elephant* ***seal***
modifier = head noun	EQUATIVE	*composer arranger*
modifier > head noun	TIME	***morning*** *class*

Table 5.8: Relative contribution of head noun and modifier for different semantic relations

of the head noun. With *composer arranger* the SR is EQUATIVE, suggesting that the two parts should contribute equally to the overall interpretation. Finally, with *morning class*, the interpretation is TIME, as signified by the temporal modifier.

Comparing these three SRs, we predict that the head noun will contribute more to the semantics than the modifier in PROPERTY, but conversely that the modifier will contribute more in TIME NCs. With EQUATIVE NCs, on the other hand, we expect that the head noun and modifier will make an equal contribution to the overall semantics.

In the third experiment, we therefore explore the relative contribution of NC constituents to the SRs, by testing the relative impact of the α weight in Equations 5.7 and 5.8 to overall performance. Figure 5.3 shows the overall accuracy for SR interpretation as we modify the value of α. The best overall performance is achieved for

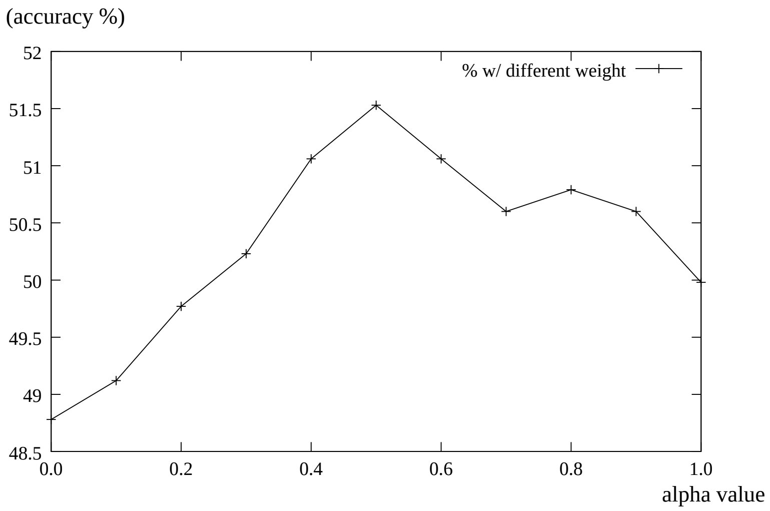

Figure 5.3: Classifier accuracy at different α values

$\alpha = 0.5$, i.e. when the head noun and modifier contribute equally to the determination of the SR.

In Figure 5.4, we explore the impact of different weight values on the accuracy for individual SRs.

Figure 5.4 shows the various performance levels when the modifier and head noun are assigned different weights. Some SRs such as CAUSE and POSSESSOR are influenced more by the modifier, while SRs such as CONTENT and PROPERTY are affected more by the head noun, as predicted (highlighted with the boxes). Contrary to our expectations, however, the head noun seems to be a stronger determinant of the SR than the modifier for EQUATIVE NCs, and both nouns seem to play an equal role for TIME NCs. The results show that despite localised biases for individual SRs, the overall performance is the best when both components have equal say in the prediction of the SR. Note that by varying the weight for the modifier and head noun, we could potentially vary the performance of our method in the other experiments described in this chapter, but we did not carry out a thorough analysis due to the mixed results in this early experiment.

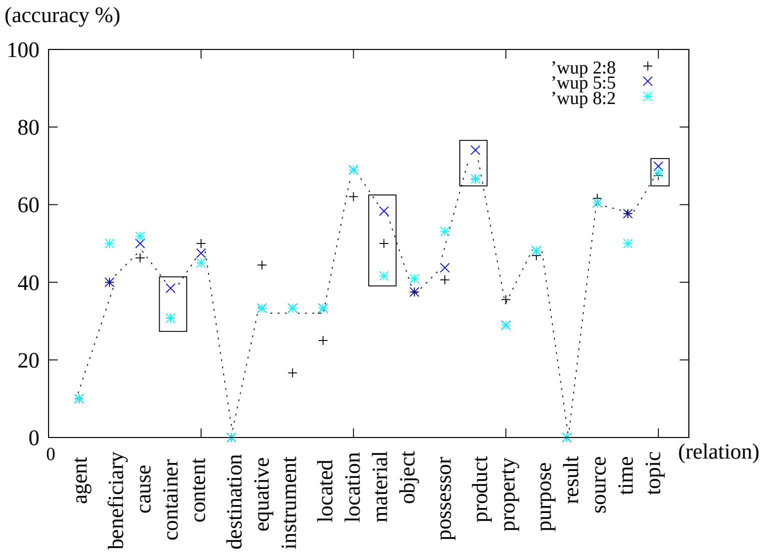

Figure 5.4: Classification accuracy for each semantic relation at different α values

5.1.8 Experiment (IV): Combining the Basic Method with Bootstrapping

The major bottleneck for our method is the availability of training instances. In order to increase the number of training instances with extra annotation, we experiment with combining our method with a bootstrapping technique. While we expect our method to perform better as we increase the number of training instances, there is also the danger of infecting the training data with incorrectly-labelled instances. At each iteration, we select those test NCs which have equal or higher similarity than a given threshold θ, and include them in the training data for next and subsequent iterations. Four different thresholds for inclusion of examples are tested: 0.6, 0.7, 0.8 and 0.9.

To evaluate this approach, we selected the `WUP` measure to calculate similarity since it performed best over 2-term NCs; we target only 2-term NCs in this experiment. Starting with the same seed set of 2,169 (2,347 w/ multiple SRs) NCs as before, we selected 1,559 (1,644 w/ multiple SRs) (70%) NCs as test instances and 610 (30%)

Iteration	Test	Train	Tagged	Correct	Incorrect	Accuracy
1	1,559	610	1,465	503	962	53.64%
2	89	2,075	52	20	32	53.27%
3	37	2,127	1	0	1	53.24%
overall	–	–	1,518	523	995	34.45%
w/org.	1,559	–	–	523	1,036	33.74%

Table 5.9: Bootstrapping with $\theta = 0.6$

Iteration	Test	Train	Tagged	Correct	Incorrect	Accuracy
1	1,559	610	1,244	446	798	56.96%
2	310	1,854	159	62	97	55.54%
3	151	2,013	13	4	9	55.38%
4	138	2.026	1	0	1	55.35%
overall	–	–	1,417	512	905	36.13%
w/org.	1,559	–	–	512	1,047	32.84%

Table 5.10: Bootstrapping with $\theta = 0.7$

NCs as training instances, i.e. we use 40% less seed training data than we did in our original experiment. For each threshold setting, we iterate until no more classifications are made. Note that any predictions made in the second and subsequent iterations are necessarily based on automatically-acquired training instances.

In Tables 5.9 to 5.12, *Test* is the number of test instances, *Train* is the number of training instances, *Tagged* is the number of test instances where the similarity is greater than or equal to the threshold for the given iteration, *Correct* is the number of correctly tagged test instances on that iteration, *Incorrect* is the number of incorrectly tagged test instances on that iteration, and *Accuracy* is the cumulative accuracy for the test instances. Figures 5.5 to 5.8 show the number of interpreted NCs, the number of correctly interpreted NCs and the number of incorrectly interpreted NCs.

The results show that when the threshold is low ($\theta = 0.6$), the method saturates very early, and only very few instances are classified in the second and third iterations.

Iteration	Test	Train	Tagged	Correct	Incorrect	Accuracy
1	1,559	610	727	301	426	68.14%
2	827	1,337	256	113	143	64.28%
3	575	1,593	42	14	28	63.49%
4	533	1,635	1	0	1	63.45%
overall	–	–	1,025	428	905	41.76%
w/org.	1,559	–	–	428	1131	27.45%

Table 5.11: Bootstrapping with $\theta = 0.8$

Iteration	Test	Train	Tagged	Correct	Incorrect	Accuracy
1	1,559	610	196	106	90	88.83%
2	1,358	806	62	34	28	86.41%
3	1,296	868	11	3	8	85.67%
4	1,284	879	3	2	1	85.60%
overall	–	–	272	145	127	46.69%
w/org.	1,559	–	–	145	1414	09.30%

Table 5.12: Bootstrapping with $\theta = 0.9$

Also, the overall accuracy is almost identical to that for the original experiment using considerably more training data, suggesting that the bootstrapping method is an effective way of reducing the need for annotated data. Predictably, as θ gets smaller, the number of acquired instances increases but the amount of infection of the training data increases (i.e. more incorrectly-tagged instances are allowed into the training data), and conversely, as θ gets larger, the number of acquired instances drops but the quality of the predictions increases. All this suggests that combining our approach with bootstrapping can allow us to reliably tag data with less manual labor and time.

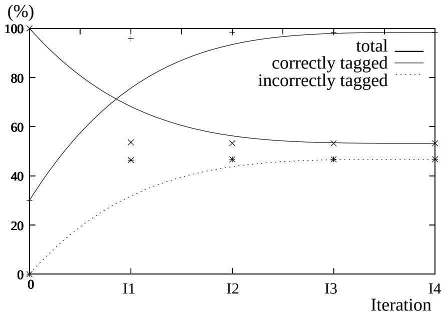

Figure 5.5: Performance with $\theta = 0.6$

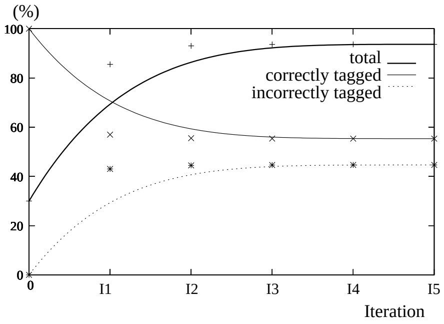

Figure 5.6: Performance with $\theta = 0.7$

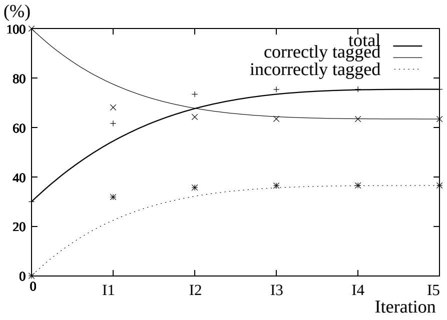

Figure 5.7: Performance with $\theta = 0.8$

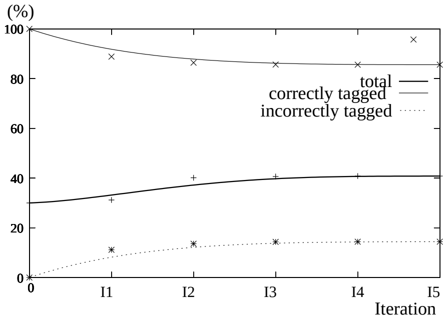

Figure 5.8: Performance with $\theta = 0.9$

5.1.9 Experiment (V): Combining the Basic Method with the k-nearest Neighbor Algorithm

This experiment is targeted at combining the k-nearest neighbor (k-NN) algorithm with our method. Instead of selecting the SR based on the first-ranked training instance, we select the top-k most similar training instances for each test instance, and predict the SR by either simple majority vote (where each training instance's vote has equal weight) or weighted majority vote (where each training instance has a vote proportional to its similarity with the test instance).

In Table 5.13, *first* is the performance when we predict the SR based on the first-ranked training instances (i.e. the basic method), *k-nearest* is the k-nearest neighbor algorithm (with simple voting), and k indicates the value of k at which the indicated accuracy was achieved. Figures 5.9 to 5.14 show the accuracy at different values of k,

Relation	Method	First	k-NN	k
Path	WUP	53.30%	47.73%	7
	LCH	52.90%	48.47%	5
Content Information	JCN	46.70%	46.35%	10,11
	LIN	47.40%	46.44%	19
Relatedness	VECTOR	42.44%	45.33%	18
	LESK	39.22%	46.25%	13

Table 5.13: Accuracy with the k-nearest neighbor algorithm

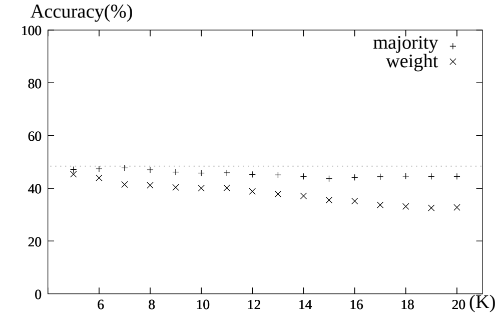

Figure 5.9: Performance of k-NN with WUP, for simple and weighted voting

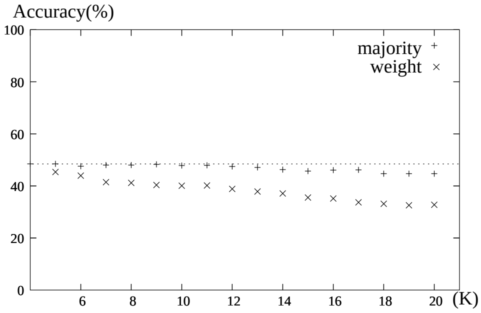

Figure 5.10: Performance of k-NN with LCH, for simple and weighted voting

for both simple voting and weighted voting. In each case, the accuracy of the basic method is indicated as a dotted line.

Although some similarity measures such as VECTOR and LESK perform better in combination with the k-nearest neighbor algorithm, overall, the best results were achieved for the original method. That is, combination with the k-nearest neighbor algorithm lowers performance, or strictly speaking, $k = 1$ is the optimal setting for the k-nearest neighbor algorithm in most cases.

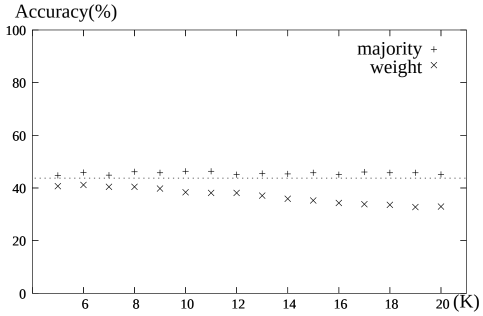

Figure 5.11: Performance of k-NN with JCN, for simple and weighted voting

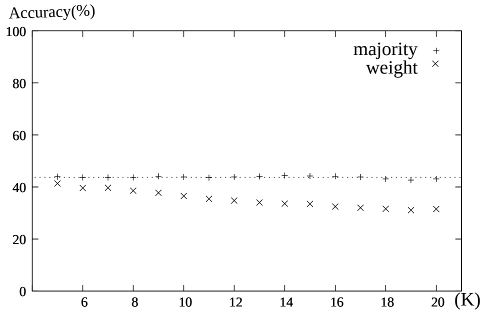

Figure 5.12: Performance of k-NN with LIN, for simple and weighted voting

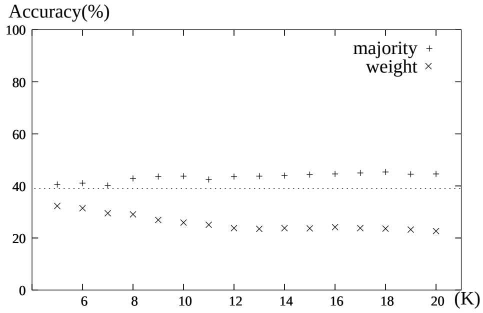

Figure 5.13: Performance of k-NN with VECTOR, for simple and weighted voting

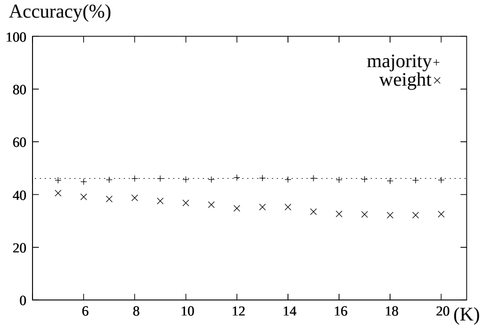

Figure 5.14: Performance of k-NN with LESK, for simple and weighted voting

5.1.10 Experiment (VI): Bracketing using Semantic Relations

Existing methods for bracketing (Marcus 1980; Lauer 1995) rely on a large amount of data to compute the probabilities. Unfortunately, however, a large amount of data is not always available. Here, we test the utility of SRs in bracketing 3-term (or larger) NCs. The basic intuition behind the method is that the SR prediction for the outermost 2-term NC from the correct bracketing, should agree with the SR prediction for the 3-term NC (without bracketing). For example, given the NC *automobile production target*, we would expect the SR for *production target* to be the

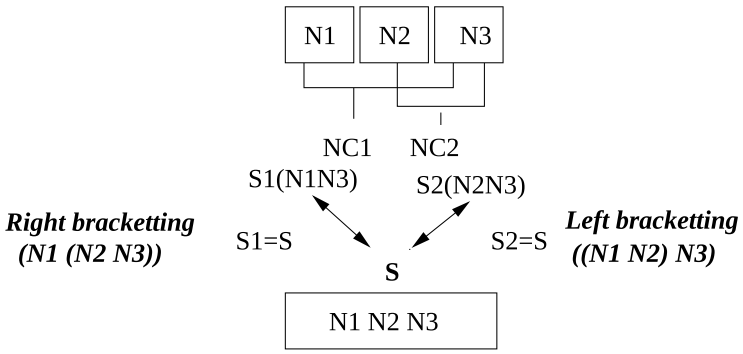

Figure 5.15: Overview of bracketing

same as for the overall 3-term NC, whereas we would not have the same expectation for *automobile target* (derived from the alternative bracketing hypothesis).

We use the WUP similarity measure to interpret NCs, and test bracketing using various similarity thresholds. For example, if the SR for the outermost 2-term NC (based on the dependency model) agrees with that for the overall 3-term NC, both calculated at a similarity value greater than or equal to 0.5, then we use the SR predictions for bracketing. Otherwise, we ignore the SR predictions and fall back on a simple lexical model.

To disambiguate the syntactic structure of 3-term NCs, we use lexical probabilities derived from the combination of our three corpora (British National Corpus, Wall Street Journal and Brown Corpus), and our 3-term NC dataset from Section 5.1.4.

Figure 5.15 shows how to identify the syntactic structure of a 3-term NC using SRs and the 2-term NCs extracted from the 3-term NC. In Figure 5.15, we extract the two outermost 2-term NCs *(N1 N3)* and *(N2 N3)* from the 3-term NC *(N1 N2 N3)*, corresponding to a right dependency bracketing and left dependency bracketing analysis, respectively. In detail, *left dependency* means that the leftmost word (i.e. first word) is in a dependency relation with the head noun, whereas *right dependency* means that there is a dependency relation between the second and final nouns. We then classify the SR for each of *(N1 N3)*, *(N2 N3)* and *(N1 N2 N3)* (i.e. *S1, S2,*

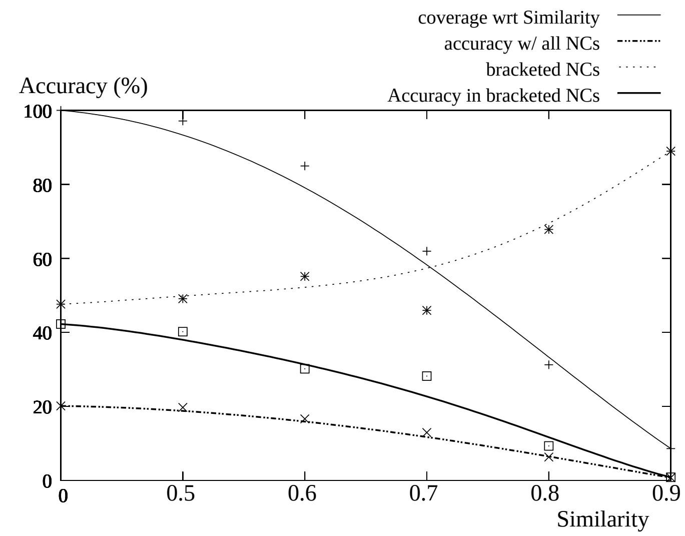

Figure 5.16: Performance and coverage of bracketing by semantic relations

S, respectively), in the latter case, based on the three component model without bracketing information. If the SR for only one of the two 2-term NCs agrees with that for the overall 3-term NC, that gives us our bracketing prediction. That is, if *S1* is the same as S, then we bracket *(N1 N2 N3)* as *(N1 (N2 N3))*, and vice versa. For example, *physics winter school*, with SR TOPIC, is associated with the two candidate 2-term NCs *physics school* (i.e. (N1 N3)) and *winter school* (i.e. (N2 N3)). Since the SR of the first 2-term NC (TOPIC) is the same as the 3-term NC, while that for the second NC (TIME), is different, we can disambiguate the 3-term NC as *(physics (winter school))* (right bracketing).

Figure 5.16 shows the accuracy of the proposed bracketing method. *Coverage wrt Similarity* is the number of NCs which are selected at a given similarity threshold (0.5, 0.6, 0.7, 0.8 or 0.9). i.e. this is the number of test NCs with a given similarity threshold. *Accuracy w/ all NCs* is the accuracy of bracketing, including NCs which are not bracketed among the NCs selected by the given similarity threshold. *Bracketed NCs* is the number of NCs which are bracketed using *selected NCs*, which are selected by the threshold similarity. *Accuracy in bracketed NCs* is the accuracy of bracketing

Training data		Prob only		Combined	
	Models	Coverage	Accuracy	Coverage	Accuracy
Corpus	Adjacency	87.1%	60.2%	93.1%	64.2%
	Left Dependency	64.9%	56.0%	81.3%	**65.8%**
	Right Dependency	63.1%	23.5%	80.1%	34.2%
Web	Adjacency	99.7%	71.2%	100.0%	71.2%
	Left Dependency	99.8%	74.0%	99.9%	**74.1%**
	Right Dependency	99.8%	40.9%	99.9%	40.9%

Table 5.14: Coverage and accuracy of the strictly probabilistic vs. combined methods

computed using only the bracketed NCs, i.e. this accuracy is equivalent to precision.

Figure 5.16 shows that bracketing using SR-tagged NCs suffers from data sparseness. However, it also shows that the method works reasonably well for those instances that are bracketed, although the coverage of the method is low.

Given the limited coverage of the method, we decided to combine our proposed bracketing method with a lexical probabilistic model. That is, we first attempt to determine the bracketing using the probabilistic model, and if the model does not provide an answer due to data sparseness (i.e. if we do not have any instances of both word pairs), then we back off to our SR-based model. As our probabilistic model, we used adjacency and dependency from Lauer (1995) (see Section 5.1.1), using the combination of our three corpora, Brown Corpus, Wall Street Journal and British National Corpus, to compute the lexical probabilities. Also, to compare the method with the current state-of-the-art, i.e. Nakov and Hearst (2005), we used Google web data (i.e. Google page hits) to compute the probabilities of substrings generated based on (left or right) dependencies. Table 5.14 shows the coverage (i.e. proportion of instances where we are able to make a prediction) and accuracy of the probabilistic model in isolation, and also the combined model.

Bracketing using only the probabilistic models suffers from data sparseness, and does not achieve 100% coverage even with web data. When we combine the probabilistic model with our semantic relation method, both the coverage and the accuracy

Method	Models	Untagged	Acc.	SR bracketed	Acc.
Lauer	Adjacency	202	30.1%	94	64.9%
	Left Dependency	552	27.9%	154	59.7%
	Right Dependency	580	29.0%	168	62.9%
Nakov and Hearst	Adjacency	1	0.0%	0	0.0%
	Left Dependency	4	25.0%	3	33.3%
	Right Dependency	2	50.0%	1	100.0%

Table 5.15: Analysis of bracketing using SRs and the probabilistic models

increase in all cases. The best performance for each of the corpus- and web-based methods were achieved with the left-dependency model. Over web data, the increment in coverage and accuracy is tiny, due to the majority of instances being disambiguated by the probabilistic model. However, even here, by combining it with the semantic relation method, we were able to increase accuracy marginally.

We also analyzed the impact of SR-based bracketing over the combined method. Table 5.15 shows the performance of SR-based bracketing over the untagged instances for the probability models (i.e. Lauer and Nakov). *Untagged* is the number of instances which are not bracketed by the probability models, *SR bracketed* is the number of instances bracketed only by SR-based bracketing among the instances untagged by the probability models, and *Acc* is overall accuracy. We present the accuracy for each of these subsets of data.

As the Nakov and Hearst web-based method bracketed the vast majority of the test instances, the impact of SR-based bracketing is negligible. This does not alter the fact, however, that the overall accuracy of both the Lauer and Nakov and Hearst methods increases when we combine it with SR-based bracketing. In conclusion, therefore, we reason that the SR-based bracketing model is a reliable form of smoothing to overcome data sparseness in the task of bracketing with probabilistic models, despite our similarity method performing at an interpretation accuracy of 53% at best.

5.1.11 Summary of the Constituent Similarity Method

In this section, we have proposed a novel method based on component similarity for interpreting NCs. Our intuition is that NCs made up of similar words have the same SR. To test this intuition, we used various word similarity measures to model the word similarity of corresponding NC components, and combined these similarities into an overall NC-level similarity.

In our implementation, we took each training instance and computed the word similarity of first the modifiers then the head nouns in the training and test instances, which we combined together in two different ways to arrive at an overall similarity. We then determined the SR of the test NC by simply taking the SR of the training instance with the highest similarity. To extend the method, we tested the method over 2-term and 3-term NCs. We also tested the contribution of the modifier and head noun on SR prediction over each SR, and experimented with combining the basic method with bootstrapping and the k-nearest neighbor algorithm.

In evaluation, our method achieved an accuracy of 53.3% over 2-term NCs and 57.7% over 3-term NCs (using all three components). Our experiment in combining our method with bootstrapping resulted in equivalent accuracy to the original method with a reduced set of training data, and also established that there is a strong correlation between the similarity value of the best-matching training instance and the quality of the prediction. Combining our method with k-nearest neighbor algorithm improved the performance for a subset of the word similarity measures, but overall the original 1-nearest neighbor method achieved the best performance. Finally, we demonstrated that SRs can be used as a means of NC bracketing for 3-term NCs, and that this method enhances overall bracketing performance when combined with probabilistic models.

In summary, we proposed a method for statistically modeling NC interpretation, and achieved promising results. Also, we attested the utility of the method in the context of an NC bracketing task.

5.2 Modeling the Compositionality of English Verb-Particle Constructions using Semantic Similarity

In this section, we target the task of computationally modeling the compositionality of English verb-particle constructions (VPCs). VPCs are composed of a verb and particle, which have variable impact on the determination of the overall semantics of the VPC. To determine the relative compositionality of the VPC, we focus on modeling the compositionality and semantic contribution using the semantic similarity of the verb and particle relative to other VPC instances.

In the following sections, we briefly describe the previous attempts made in this area, and present our motivation and approach to VPC compositionality. We then take a look at the performance of our approach and provide a discussion of the results, in comparison with prior research.

5.2.1 Past Research on Modeling the Compositionality of MWEs

Most prior work on the compositionality of MWEs has used distributional similarity (Lin 1999; Schone and Jurafsky 2001; Baldwin *et al.* 2003a; Bannard *et al.* 2003; McCarthy *et al.* 2003).

To measure or model the compositionality of MWEs, Lin (1999) primarily focused on modeling non-compositional expressions including noun-noun and verb-noun pairs. The basic hypothesis was that non-compositional phrases have significantly different mutual information from that of anti-collocations generated by replacing one word at a time with a related word (see Section 3.3.2).

Schone and Jurafsky (2001) primarily targeted noun-noun pairs. The underlying intuition behind their method is that the degree of semantic similarity between a MWE and its component indicates the relative compositionality of the MWE. The authors first employed LSA to extract non-compositional MWEs. Then they calculated the cosine similarity between each extracted MWE and the weighted vector sum of its components to demonstrate that lower similarities indicated lower MWE compositionality.

Baldwin *et al.* (2003a) dealt with noun-noun compounds and VPCs only, also using LSA. The basic assumption was that higher similarity between an MWE and its constituent words indicates greater compositionality (see Section 3.4.3).

Venkatapathy and Joshi (2005) measured the relative compositionality of verb-noun pairs. The authors argued that if MWEs are decomposable, then they are more likely to be compositional. To prove this claim, the authors measured the relative compositionality of verb-noun pairs with collocation- and context-based features. They also proposed an SVM-based ranking function to rank verb-noun pairs based on their relative compositionality, by integrating both their proposed features and features from previous research.

Bannard *et al.* (2003) and McCarthy *et al.* (2003) concentrated exclusively on English VPCs, in measuring compositionality, and in the case of Bannard *et al.*, also the semantic contribution of the parts. The assumption of Bannard *et al.* was that if a VPC is similar to its verb in isolation, then the verb contributes its simplex meaning, and similarly for particles. The proposed approach uses the semantics of the verb and particle in a given VPC, based on a range of semantic models. To measure the similarity, the authors modeled the similarity via a number of variants of the overlap of the top N neighbors. The first technique they employed is that of Lin (1999). The second one (called "knowledge-free similarity measure") is similar to the first method but uses synonyms obtained from a distributional thesaurus. The third method uses corpus-based semantic similarity between the original VPC and its 10-nearest neighbor word-substituted expressions. Finally, the last method measures similarity based on the pair-wise comparison of all VPCs with all verbs and all particles using distributional similarity. The authors also carried out comparative evaluation with a thesaurus, hand-made resources and collocation measures.

McCarthy *et al.* (2003), on the other hand, modeled VPC semantics holistically (i.e. if the VPC semantics is predicted in part or in whole by its parts, it is considered to be compositional), but used a finer-grained scale of 0 to 10 to model composition-

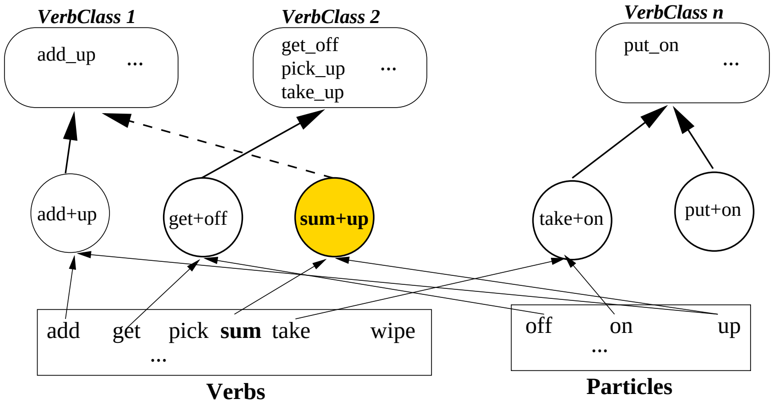

Figure 5.17: Verb classes in predicting the compositionality of verb-particle constructions

ality. The authors used a wide range of similarity measures based on distributional similarity to model compositionality (see Section 3.2.3).

5.2.2 Motivation for our VPC Compositionality Method

In prior work relating to VPCs, Villavicencio (2005) claimed that VPCs with similar verbs and particles tend to have similar semantics as a whole. Our method for modeling the compositionality of VPCs is based on this same intuition. For example, assuming *call up* is COMPOSITIONAL, when we replace *call* with the synonym *ring*, *ring up* will also be COMPOSITIONAL. In this way, we can predict the compositionality of unseen VPCs by looking at their semantic similarity with VPCs which we know the compositionality of.

Figure 5.17 depicts our method for predicting the compositionality of the VPC *sum up*. Here, VPCs are grouped by semantics, and each group is tagged with its relative compositionality. For example, *put on* and *take on* in verb class n are semantically similar. Note that the semantics of the component verb and particle determine the semantics of the overall VPC, and that the same VPC can vary in its relative compositionality across different interpretations (c.f. *make up the answer* vs.

VPC	Compositionality	Contribution
lie down	YES	both
close off	YES	verb
get down	YES	particle
chicken out	NO	none

Table 5.16: VPCs with compositionality judgements, based on the semantic contribution of the components

kiss and make up). For the purpose of this research, however, we make the simplifying assumption that a given VPC has fixed compositionality irrespective of semantics.

The main step in classifying VPCs in our method is based on semantic similarity. In order to calculate the semantic similarity of a test instance with each of the training instances, we consider the semantic contribution of the verb and particle independently. This gives us four possibilities: verb contribution only, particle contribution only, both verb and particle contribution, or no semantic contribution from with component.

Table 5.16 describes the relative semantic contribution of the parts in determining the semantics of the overall VPC, and the compositionality of the VPC. Similarly to McCarthy *et al.* (2003), we consider the VPC to be compositionality when one or both parts contribute towards the overall semantics of the VPC. Hence the first three examples are compositional while the last one is non-compositional.

The examples in Table 5.16 were retrieved from two lexical resources: Bannard (2006) and McCarthy *et al.* (2003). We use Bannard (2006) to model the semantic contribution of each of the verb and particle, and McCarthy *et al.* (2003) as the source of the degree of compositionality of each VPC. We consider a VPC to be compositional if a majority of the annotators rate the VPC as having 5 (out of 10) or higher. Our decision to binarize the compositionality judgments was based on the realization that the annotation of compositionality is subjective, and the desire to maximize agreement over the data.

We claim that VPC compositionality can be predicted using semantic similarity, as calculated based on the semantic contribution of each of the components of the VPC. We use WordNet v2.1 *hypernym*s to represent the semantics of the VPCs and their components. In particular, we use direct *hypernym* (1_{st} *hypernym*) and root *hypernym* (N_{th} *hypernym*) to represent VPC and verb semantics, in order to capture semantic generalizations. Note that since our calculation of compositionality is sense-independent,[2] we use all senses of the verb/VPC to calculate the overlap in semantics. In order to satisfy the need for particle semantics, we adopt the classification presented in Bannard *et al.* (2003), as described in Section 5.2.4. In order to evaluate our approach, we built two types of classifiers: one lists the features in a conventional linear format, while the other lists the features in the form of a 2-dimensional matrix.

5.2.3 Classifier Design

In this work, we built two different types of classifiers, based on TiMBL v5.1 (Daelemans *et al.* 2004) and a Maximum Entropy learner. The first uses a linear listing of the semantics of the VPCs and verbs (**C1**). In order to detect the same or similar semantic combinations of verbs and particles, we mark the co-occurrences of verb and particle semantics. We use the *verb*, *particle* and *collocation* properties of the verb and particle as features. The input features are described in Table 5.20. The second classifier design uses a 2-dimensional matrix of verbs and particles, with each cell indicating the existence of a VPC with that combination of verb and particle, and a given VPC is represented as the concatenation of the feature vector for the verb and particle contained in it. This style of classification was introduced in Uchiyama *et al.* (2005) (**C2**). Uchiyama *et al.* (2005) showed that by concatenating the column and row features of the matrix, they were able to achieve strong performance for a semantic classification task over Japanese compound verbs which relates to compositionality. Since English VPCs are similar to Japanese compound verbs, we expect similar performance improvements in using this classifier design.

In Figure 5.18, C1 is the one-dimensional classifier with only particle information,

[2] There are certainly cases where compositionality varies greatly across senses of a VPC, but in our experience these are rare.

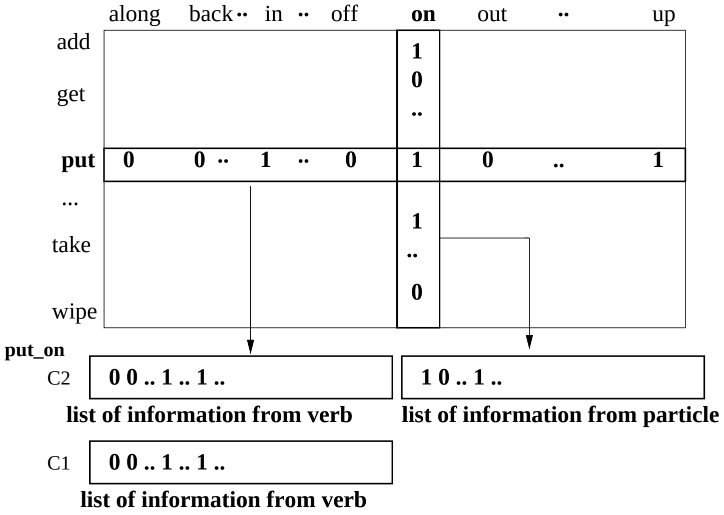

Figure 5.18: 2-dimensional matrix classifier

and C2 is the 2-dimensional classifier based on the concatenation of the verb and particle features of the 2-dimensional matrix. For each of C1 and C2, we experiment with two different representations for each cell: `collocational` and `semantics`. The collocational classifier has a 1 for a given combination of verb and particle if it has been observed as a VPC in the training data, and otherwise has a value of 0. The semantics classifier, on the other hand, uses the *hypernyms* (i.e. 1_{st} and N_{th} *hypernyms*) of the VPCs and verbs, as well as the semantics of the particles in isolation. As a result, we have four distinct classifiers.

5.2.4 Data Collection for Modeling VPC Compositionality

We used data from two sources for our experiments: data used in Bannard (2006), providing the semantic contribution of each of the verb and particle in the form of binary classes; and data used in McCarthy *et al.* (2003)[3], binarized as described above.

[3] http://mwe.stanford.edu/resources/

Compositionality	Contribution	VPC	V
Compositional	Verb	52	59
	Particle	12	11
	Both	50	62
Non-compositional	None	22	25
Total		134	158

Table 5.17: Breakdown of the compositionality data from Bannard (2006)

VPCs with Semantic Contribution

160 VPCs were manually tagged with the semantic contribution of the verb and particle in Bannard (2006). The tagging involved 29 human annotators who tagged "yes" if the VPC entailed the verb, and "no" otherwise, and similarly for particles. If they could not decide on the contribution of verb or particle, they were allowed to assign a tag of "don't know". We consider the VPCs in the data set to be COMPOSITIONAL if at least half of the human annotators answered "yes" for either the verb, particle or both. We also specified the relative semantic contribution of the verb and particle when the VPC is compositional: VERB-CONTRIBUTION, PARTICLE-CONTRIBUTION or VERB&PARTICLE-CONTRIBUTION. Table 5.17 describes the number of compositional and non-compositional VPCs in the data set.

Note that in Table 5.17, the total number of instances is less than the original number since some of the VPCs (VPC) or verbs (V) were not found in WORDNET v2.1. As a result, the final size of the test and training data set is only 134 for the compositionality task and 158 for the semantic contribution task.

VPCs with Degree of Compositionality

McCarthy *et al.* (2003) manually scored 117 VPCs for compositionality, assigning scores between 0 and 10. A score of 10 indicates a fully compositional VPC, while a score of 0 indicates a fully non-compositional VPC. We binarized this to simple

Compositionality	VPC	V
Compositional	35	58
Non-compositional	32	39
Total	67	97

Table 5.18: Breakdown of the compositionality data from McCarthy *et al.* (2003)

COMPOSITIONAL and NON-COMPOSITIONAL, based on the average compositionality score across all three annotators.

Due to some of the instances not occurring in WordNet v2.1, we extracted a total of 67 VPCs and 97 verbs in isolation from the original McCarthy *et al.* (2003) dataset.

Semantics of Particles

In order to extract the semantic contribution of particles in the VPCs, two human annotators tagged the semantics of all particles in our data set. The tags used were TEMPORAL, SPATIAL-DIRECTION and SPATIAL-POSITION as described in Bannard *et al.* (2003). We also added a class of NONE for particles where the semantics in the VPC did not correspond to either of these three. Two human annotators tagged each of 256 VPCs where both the verbs and the VPC were found in WordNet v2.1, and combined their annotations. The initial agreement was 51.71%, and all instances of disagreement were resolved based on mutual discussion. The final distribution of VPCs is described in Table 5.19.

In our data, we observed that the particle semantics did not always match the prediction of particle contribution. We found that particles in 22 VPCs were tagged as NONE despite the particle being marked as contributing to the VPC semantics (either particle only or both of particle and verb). This discrepancy may be due to the fact that we determined the compositionality based on the majority of the annotations (binary compositionality judgment), and that the annotators for the particle and

Semantics	Number of VPCs
SPATIAL-DIRECTION	86(33.59%)
SPATIAL-POSITION	49(19.14%)
TEMPORAL	0(0%)
NONE	121(47.27%)
Total	256

Table 5.19: Semantics of particles in the data

VPC compositionality were different. As a result, although the semantics of particles were tagged as NONE, they still seem to contribute to the semantics of the VPC. For example, the compositionality of *knock out* from the data of Bannard *et al.* (2003) is a borderline case. 14 human annotators marked it as particle compositional and 12 annotators marked it as particle-non-compositional. In our case, the particle semantics of *knock out* is considered to be NONE based on our final annotation. We also noticed that there were no instances of SPATIAL particles in this dataset, as SPATIAL particles tend to be fully productive and the VPC thus doesn't occur in WordNet v2.1 (e.g. *eat up*).

5.2.5 Evaluation of VPC Compositionality Modeling

Evaluation Set

In order to to evaluate the success of our approach at modeling the compositionality of VPCs, we used the semantics (senses) of both the VPC and verb in isolation for the test and training instances. We used two different representations of *hypernyms*: first *hypernym* (1_{st}) and root *hypernym* (n_{th} *hypernym*). Additionally, we added the semantics of the particles in order to test the influence of the particles on compositionality. To evaluate, we built two different classifiers, C1 and C2, based on the two feature representations. The total number of discrete data sets in our experiments was 25, as listed in Table 5.20.

Test	Training	ExperimentID	Classifier	Features
VPC	VPC	E1	C1	Vs,Vs+P,Vs+P+Ps
		E2	C2	C–C,C–S,S–C,S–S
V	V	E3	C1	Vs,Vs+P,Vs+P+Ps
V	VPC	E4	C1	Vs,Vs+P,Vs+P+Ps

Table 5.20: Different data sets used in our experiments

In Table 5.20, *Vs* signifies verb semantics, *P* signifies the particle word form, *Ps* signifies particle semantics, *C* signifies collocation, and *S* signifies semantics. For E2, we adapt the two feature–value representations from Uchiyama *et al.* (2005) and end up with the following four possibilities: (1) C–C, where both the verb and particle are represented by way of simple (binary) co-occurrence; (2) C–S, where the verb is represented via co-occurrence with different particles, but the particle is represented by its semantics (e.g. SPATIAL-POSITION, or 0 if it doesn't produce a VPC in combination with the given verb); (3) S–C where the verb is represented via semantics (i.e. WordNet synsets) and the particle via simple co-occurrence; and (4) S–S where both the verb and particle are represented via semantics. All semantic features take the form of 1_{st} or N_{th} *hypernym*s (both verb in isolation and VPCs).

We evaluated via 10-fold cross-validation for experiments 1, 2, and 3 (E1, E2 and E3), while we used simple holdout evaluation for experiment 4 (E4) due to the different sources of semantics of VPCs and verbs.

Modeling Compositionality

We used two different learners, TiMBL v5.1 and Maximum Entropy, to build our classifiers. We report the best performance for each experiment, along with the baseline, in terms of precision, recall and F-score here, and include the full results in Section D of the appendix. The baseline is computed by majority vote, and varies as the final number of instances differs depending on occurrence in WordNet v2.1, as mentioned in Section 5.2.4. We also report on correlation between the classifiers

ExperimentID	Feature	Semantics	Compositional	Precision	Recall	F-score
E1	base	–	Y	.731	–	–
	line	–	N	.269	–	–
	V	N_{th}	Y	.743	.728	.736
			N	.298	.315	.307
E2	base	1_{st},N_{th}	Y	.761	–	–
	line		N	.239	–	–
	S-C	1_{st}	Y	.791	.895	**.843**
			N	.429	.250	**.339**
E3	base	–	Y	.749	–	–
	line	–	N	.251	–	–
	V	1_{st}	Y	.771	.864	.818
			N	.366	.234	.300
E4	base	–	Y	.598	–	–
	line	–	N	.402	–	–
	V	1_{st}	Y	.632	.948	.759
			N	.700	.180	.286

Table 5.21: Results of experiments with TiMBL v5.1

and the human judges, as in McCarthy *et al.* (2003). Note that in Table 5.21 and Table 5.22, *ExperimentID* is the experiment number (E1-E4), *Feature* is the feature combination, *Semantics* is the source of the semantics, and *Compositional* (**Y**es or **N**o) is compositionality. Note that Table 5.25 shows the accuracies corresponding to Table 5.21 and 5.22.

The results in Table 5.21 are from TiMBL v5.1 and the results in Table 5.22 are from Maximum Entropy. The first result of note is that our method works well when using both the semantics of VPCs and verbs. However, in general, we observed that the particle semantics diminished the performance of our method, since we did not have enough instances to generalize over. We also found that 1_{st} *hypernym*s are the most reliable feature representation for the verbs.

Secondly, we compared the two classifier designs, C1 and C2. As we expected, C2 produced the best performance for both TiMBL and Maximum Entropy, mir-

ExperimentID	Feature	Semantics	Compositional	Precision	Recall	F-score
E1	base	–	Y	.731	–	–
	line	–	N	.269	–	–
	VPS	N_{th}	Y	.751	.905	.828
			N	.435	.185	.310
E2	base	CC	Y	.761	–	–
	line		N	.239	–	–
	C-S	1_{st}	Y	.759	.993	**.876**
			N	.000	.000	.000
E3	base	–	Y	.749	–	–
	line	–	N	.251	–	–
	VPS	1_{st}	Y	.773	.895	.836
			N	.429	.234	.332
E4	base	–	Y	.598	–	–
	line	–	N	.402	–	–
	VP,VPS	1_{st}	Y	.604	1.000	.802
			N	1.000	.026	.513

Table 5.22: Results of experiments with MAXIMUM ENTROPY

roring the results of Uchiyama *et al.* (2005). The highest F-score was 84.3% and 87.6%, respectively, for TIMBL and MAXIMUM ENTROPY, at modeling compositional VPCs. However, the performance at modeling non-compositional VPCs was considerably lower, with C1 producing the highest F-score of 51.3% in experiment 4, based on MAXIMUM ENTROPY. We claim that this was due to the relative lack of non-compositional VPCs in the data set. However, overall, all results for the classifiers based on both TIMBL and MAXIMUM ENTROPY exceeded the baseline.

Thirdly, we compared the results for F-score and accuracy. Unlike the F-score figures, the accuracy was low. Among experiments E1, E2 and E3 (using 10-fold cross validation), only C2 with MAXIMUM ENTROPY exceeded the baseline. On the other hand, C1 with experiment 4 (E4, based on holdout evaluation) outperformed the baseline. The results show that the relative sparsity of non-compositional VPCs in our data set reduces the overall accuracy.

Feature	Compositional	Non-Compositional	Total
VPC	112 Verb: 52 Particle: 10 Both: 50	22	134
V	133 Verb: 59 Particle: 62 Both: 12	25	158

Table 5.23: Number of instances with verb and/or particle semantic contribution where V is verb in isolation

Modeling the Semantic Contribution

In this experiment, we attempt to predict the relative semantic contribution of the verb and particle within compositional VPCs. The data for this experiment is described in Table 5.23.

For the purposes of evaluation, we built three classifiers using TiMBL v5.1: (1) two classifiers using C1 but containing different semantics, one from the VPC and the other from the verb in isolation, and (2) one classifier using C2. Based on evaluation with 10-fold cross-validation, we achieved the highest accuracy of 81.95% with classifier C2 using the feature semantics-semantics (S-S), deriving semantics from the 1_{st} *hypernym* of the VPC and using particle semantics. Table 5.24 shows the best-performing classifiers for the semantic contribution task.

Of the three classifiers based on C1 and C2, we found that our approach produced the best performance using C2, with the S-S feature representation. The F-score exceeded the baseline except for the case of particle contribution, once again due to the relative infrequency of instances of particle contribution.

Feature	Compositionality	Combination	Precision	Recall	F-score
base line	yes	VP	.373	–	–
		V	.388	–	–
		P	.075	–	–
		All	.836	–	–
	no		.164	–	–
S-S	yes	VP	.489	.460	.475
		V	.474	.745	.610
		P	.000	.000	.000
		All	**.843**	**.964**	**.903**
	no		**.333**	**.091**	**.212**

Table 5.24: Result of the verb and particle contribution to compositionality

Measuring Correlation

We checked the Pearson correlation between our methods and human judgment over the data set of McCarthy *et al.* (2003), by extracting the marginal probabilities of each instance from Maximum Entropy and using the original continuous average of annotations in the gold-standard. Table 5.25 shows the accuracy from TiMBL and Maximum Entropy, and also the correlation for Maximum Entropy. Figures in **boldface** indicate where the results exceed the baseline. Note that in Table 5.25, *Semantics* is the source of the semantics, *C* is the level of *hypernym*s, *Vs* is verb semantics, *P* is the particle word form, and *Ps* is particle semantics.

To compare the correlation results, the range reported in McCarthy *et al.* (2003) is between −0.115 and 0.490, whereas ours is between −0.070 and 0.274. Although the best correlation achieved by our models is lower than that of McCarthy *et al.* (2003) (0.490 vs. 0.274), it is worth reflecting that McCarthy *et al.* (2003) used distributional similarity over the entire British National Corpus (i.e. a data- and computationally-intensive approach), as contrasted with our experiments which use a relatively simple feature representation based on lexical semantic resources (i.e. a knowledge-intensive approach). Out of our classifiers, C1 using direct *hypernym*s to represent verb semantics (i.e. E3, with Verb$_{SEM}$(1st)) produced the highest correlation. We observed that the particle semantics reduced the overall performance as they became noises in the data. We also observed it due to the lower agreement on annotating particle semantics.

5.2.6 Summary of VPC Compositionality Modeling

We proposed a method based on semantic similarity to model the compositionality and semantic contribution of the components in VPCs. The motivation for our method is that similar combinations of verb and particle tend to have similar semantics and compositionality.

We took into account the combination of verb and particle as our primary feature in determining semantic similarity. We also used the semantics of the verbs and particles. In evaluating our approach, we built two different styles of classifier, one using features from only the verb, and the other using a 2-dimensional representation based on the verb and particle. We also experimented with the use of direct and unique-beginner hypernyms as our verb semantic representation. Finally, we tested our method by building classifiers using TiMBL and a Maximum Entropy learner. We further tested the correlation between the marginal probability estimates of the Maximum Entropy model and human judgments.

From our investigation, we found that not only do the same or similar pairings of verb and particle share the same semantics, but they also have same compositionality. We also found that direct *hypernym* is the optimal semantic representation for verbs. Comparing TiMBL and Maximum Entropy, there was relatively little difference in performance, although Maximum Entropy had the advantage of generating marginal probabilities to use in evaluating the correlation between the system predictions and human judgments.

5.3 Chapter Summary

In this chapter, we investigated two MWE-based computational tasks based on semantic similarity and the intuition that similar instances have similar semantics. Our computation of similarity was based on semantic modeling of the components. To interpret NCs, we computed the semantic similarity of each of the modifier and head noun, and combined them to compute the final similarity. To model the compositionality and semantic contribution of English VPCs, we checked the semantic overlap of VPCs and their parts (verb and particle).

Our main finding is that semantic similarity effectively captures the semantic diversity of NCs and VPCs when combined with hand-crafted resources such as WordNet and CoreLex. As WordNet has very fine-grained semantics, we heuristically extracted the semantics for NCs and VPCs by looking at first sense and its three direct hypernyms. In our experiments, this heuristic proved to be reliable. Based on these rich semantic representations of NCs and VPCs, semantic similarity was able to model the semantics (i.e. semantic relations for NCs and compositionality for VPCs) relative to similar NCs and VPCs.

In summary, semantic similarity is a useful statistical technique for semantically classifying MWE lexical items. It interprets semantic relations in NCs by referring to similar MWEs based on similarity. It also model compositionality and the semantic contribution of the components in VPCs by checking the semantic similarity between a VPC and its components (either the verb or particle in isolation, or both).

We summarize the individual findings of this chapter as follows.

Summary of the constituent similarity method for interpreting noun compounds:

- We proposed a constituent similarity based on 1-nearest neighbour matching over the union of senses of the modifier and head noun, with distance defined by word-level similarity in WordNet;
- We tested the relative contribution of constituents with respect to SR prediction, and found that contribution was surprisingly even;
- We tested the constituent similarity method over both 2- and 3-term NCs;

- We extended the basic method with bootstrapping & k-NN, with mixed success;
- We showed the utility of SR interpretation by integrating it into a combined bracketing method.

Summary of our method for modeling the compositionality of verb-particle constructions:

- We confirmed the utility of semantic similarity in modeling the compositionality and semantic contribution of parts in VPCs;
- We adapted a 2-dimensional matrix classifier for use over English VPCs;
- We proposed heuristics based on direct hypernyms and root hypernyms for representing verb semantics, and confirmed that direct hypernyms (1_{st}) is an effective representation;
- We checked the correlation between our system predictions and the gold-standard labels based on the method proposed by McCarthy *et al.* (2003).

Exp.ID	Semantics(C)	TiMBL	Maximum Entropy	
		Accuracy	Accuracy	Correlation
E1	base	73.10%		
	Vs(1_{st})	52.70%	64.00%	.181
	Vs(N_{th})	61.70%	67.00%	.148
	Vs+P(1_{st})	59.70%	66.50%	.196
	Vs+P(N_{th})	61.70%	68.00%	.148
	Vs+P+Ps(1_{st})	54.70%	68.00%	.127
	Vs+P+Ps(N_{th})	59.20%	71.50%	.085
E2	base(C)	73.60%		
	base(1,2)	75.10%		
	CC	71.09%	**73.91**%	.100
	CS(1_{st})	70.60%	**75.50**%	-.070
	CS(N_{th})	70.60%	**75.50**%	-.070
	SC(1_{st})	74.00%	67.00%	.029
	SC(N_{th})	73.50%	66.50%	.020
	SS(1_{st})	68.50%	72.00%	.155
	SS(N_{th})	66.00%	72.00%	.152
E3	base	74.90%		
	Vs(1_{st})	70.60%	70.31%	**.274**
	Vs(N_{th})	70.60%	70.40%	.183
	Vs+P(1_{st})	70.20%	71.85%	.274
	Vs+P(N_{th})	67.10%	69.16%	.163
	Vs+P+Ps(1_{st})	67.10%	73.10%	.250
	Vs+P+Ps(N_{th})	66.70%	69.93%	.140
E4	base	59.80%		
	Vs(1_{st})	**63.90**%	**62.89**%	.237
	Vs(N_{th})	**61.90**%	57.79%	.072
	Vs+P(1_{st})	**62.90**%	58.76%	.204
	Vs+P(N_{th})	**60.80**%	**60.82**%	.121
	Vs+P+Ps(1_{st})	57.70%	**60.82**%	.157
	Vs+P+Ps(N_{th})	**60.80**%	**60.82**%	.106

Table 5.25: Accuracy and correlation over the compositionality task

Chapter 6

MWEs and Ellipsed Predicates

Prior research on NC (noun compound) interpretation has approached the problem primarily using two approaches: semantic similarity (Rosario and Marti 2001; Moldovan *et al.* 2004; Kim and Baldwin 2005; Nastase *et al.* 2006; Girju 2007), and paraphrasing based on an ellipsed predicate (Levi 1978; Vanderwende 1994; Lapata 2002; Kim and Baldwin 2006b). By "ellipsed predicate" we mean that there is a binary predicate which can be used to paraphrase the NC into a form which explicitly describes the underlying semantics. For example, we can interpret *peanut butter* as "butter made from peanut(s)" using the ellipsed predicate *made from*.

In this chapter, we present a novel method for interpreting NCs based on the prediction of an *ellipsed predicate*. The method uses sentential contexts for the components of a given NC to identify predicates which relate them, which it then translates back into semantic relations.

In the following sections we take a brief look at related research work based on identification of ellipsed predicates, present the motivation for our approach, describe the details of our NC interpretation method, and finally evaluate our method.

6.1 Motivation and Method for Interpreting Noun Compounds via Ellipsed Predicates

As defined in Section 5.1, semantic relations (SRs) are directed binary predicates, generally in the form of either a verb or a preposition. (6.1) shows two NCs interpreted using both verbal (V) and prepositional (P) predicates.

(6.1) *peanut butter* = MAKE

- (V) "butter made from peanut"
- (P) "butter from peanut"

abortion problem = TOPIC

- (V) "problem concerned with abortion"
- (P) "problem about abortion"

Previous research on interpretation using ellipsed predicates has focused on verbal forms (Levi 1978; Vanderwende 1994; Lapata 2002) and/or prepositions (Levi 1978; Lauer 1995). Levi (1978) defined SRs using a mixture of 9 verbs and prepositions. Vanderwende (1994) defined SRs based on *WH*- questions and attempted to interpret NCs via verb semantics automatically extracted from an online dictionary. Figure 6.1 shows the NC interpretation method, using examples taken from the original paper.

In Figure 6.1, *keep* acts as the predicate which links *bird* and *cage* via the indicated pseudo-dependency structure. The links are determined based on analysis of a machine-readable dictionary, and act as a guide in narrowing down the interpretation choices. By using the verb semantics and direction of the SR, we can interpret the NC as WHERE(LOCATION), i.e. "cage where (one) keeps bird(s)".

Another example of using verb semantics is Lapata (2002), where compound nominalizations were interpreted via the verbal form of deverbal head nouns, focusing exclusively on the two SRs of subject and object.

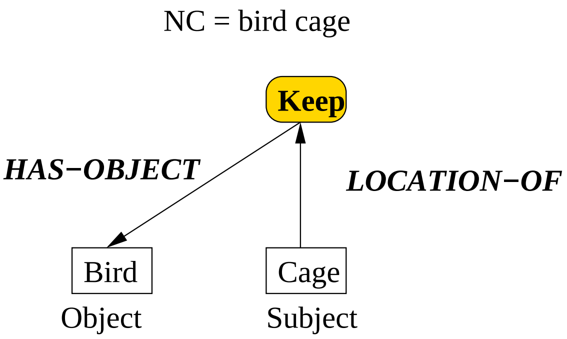

Figure 6.1: An example of the NC interpretation of Vanderwende (1994)

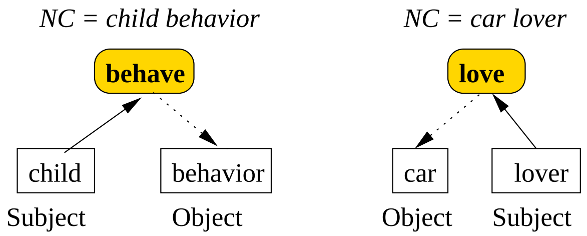

Figure 6.2: Interpreting compound nominalizations (Lapata 2002)

In Figure 6.2, the head noun *behavior* has base verb *behave*, on the basis of which *child behavior* is interpreted as SUBJECT ("child behaves"). Similarly, the head noun *lover* has base verb *love*, leading to the interpretation OBJECT ("love the car").

As seen above, the SRs in NCs can be represented via the argument structure of a related verb, with the modifier and head noun as arguments. In our method, in order to link the head noun and modifier, we use verbs associated with the definitions of general-purpose SRs instead of extracting the verbs manually (Vanderwende 1994) or restricting our method to a particular subset of NCs (Lapata 2002).

In Figure 6.3, the verbs *have* and *produce* are automatically extracted from the definitions of the SRs used in Chapter 5. The grammatical roles of the head noun and modifier are also extracted from the SR definitions.

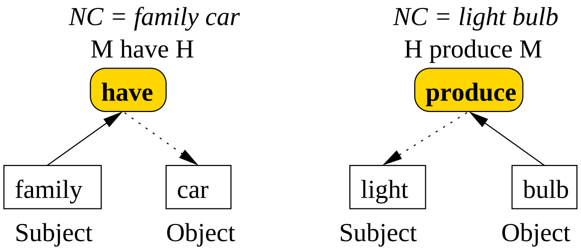

Figure 6.3: Interpreting NCs via underlying predicates in our method

The following examples show how the SR expresses the relationship between the modifier (M) and head noun (H) of the given NCs. As with the method of Vanderwende (1994), this SR can be represented as a binary predicate with the head noun and modifier as arguments.

(6.2) **NC:** *family car*
relation: POSSESSOR
logical form: `own(M,H)`
gloss: "family owns car"

(6.3) **NC:** *student protest*
relation: AGENT
logical form: `perform(M,H)`
gloss: "student performs protest"

(6.2) and (6.3) show how the SR links the head noun and modifier. Using the grammatical role of the head noun and modifier and verb in the sentences, *family car* is interpreted as POSSESSOR and *student protest* is interpreted as AGENT.

In addition, we observe that a given SR can be associated with a range of verbs, possibly with variation in the argument structure:

(6.4) **NC:** *family car*
relation: POSSESSOR
logical form: `own(M,H)`
synonyms: `have/possess/belong to`
gloss: "car belongs to family"

(6.5) **NC:** *student protest*

relation: AGENT
logical form: `perform(M,H)`
synonyms: `perform/act/execute/carry out/do`
gloss: "student carries out protest"

An example of the argument structure changes is *belong to* in relation to the POSSESSOR SR, which we analyze as a binary predicate but with the head noun as the first argument (subject) and modifier as the second argument (prepositional object). That is, the logical form is `belong-to(H,M)`, corresponding to *car belongs to family.* Note that the third verb (*have*) is identical to *own* in argument structure.

The mapping of SRs onto verbs and semantic structures via their associated logical forms was a simple and effective process, but suffers from low coverage (similarly to other pattern-based knowledge acquisition techniques (Hearst 1992; Widdows 2003; Snow *et al.* 2005)). To overcome this issue, we introduce a verb mapping mechanism. That is, in order to expand the coverage of the seed verbs extracted from the SR definitions, we map the seed verbs onto synonyms. The seed verbs are expected to have high precision in terms of SR interpretation, but low recall. Additionally, within the set of seed verbs, we chose two sets of seed verbs of size 57 and 84, in order to examine how the coverage of actual verbs by seed verbs affects the performance of our method. Initially, we chose a set of 60 seed verbs manually. We then added synonyms from the MOBY THESAURUS for some of the 60 verbs. Finally, we filtered verbs from two expanded sets that occur very frequently in the corpus (as this might overly skew the results). The verbs *own, possess, belong_to* are from the set of 57 for the SR POSSESSOR, and the verbs *acquire,grab,occupy* are added in the set of 84. In order to map the verbs onto seed verbs, we used WORDNET::SIMILARITY and MOBY THESAURUS.

6.2 System Architecture for the Ellipsed Predicate Method

Figure 6.4 shows the basic architecture employed in our ellipsed predicate interpretation method. First, we identify token instances of each verb in the combined set

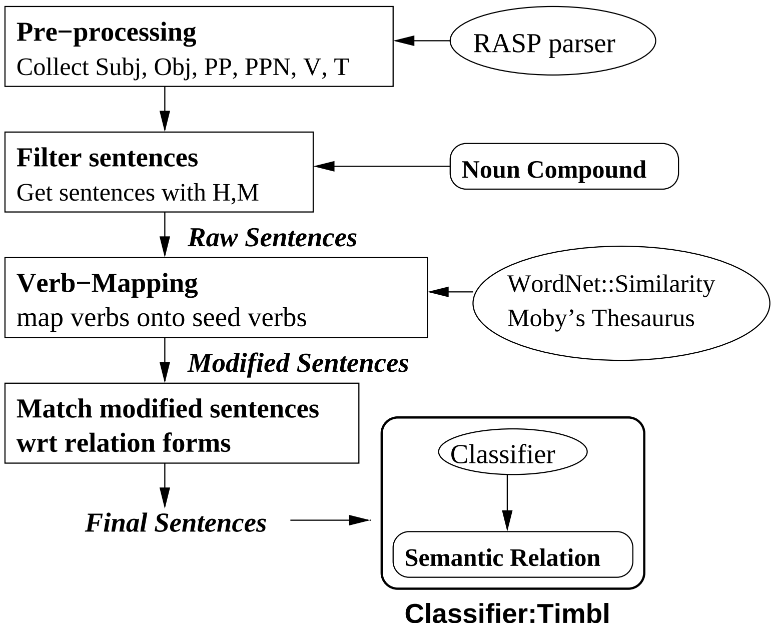

Figure 6.4: System Architecture

of SRs, and identify the argument structure for each token instance of a given verb via the dependency representation produced by RASP. As our corpus, we used the combination of the Brown Corpus, Wall Street Journal and British National Corpus. Second, we determine the relative fit of the argument structure of each instance with the template(s) associated with the given verb. Third, we map actual verbs onto SRs. In practice, the computation of which verbs are associated with which SRs (in either the original set of seed verbs or the expanded set) is carried out offline, and the mapping takes the form of a simple table lookup. Finally, assuming correct fit, we translate the relevant arguments onto a logical form based on the template associated with the SR. We then feed the instances into a supervised classifier based on TiMBL v5.1.

In evaluation, we employ two data representations for our token instances: *Count* and *Weight*. *Count* is based on the raw number of corpus instances, while *Weight* employs the seed verb weight described below.

	Sentences (# of unique verbs)	NC types
Overall	9013 (1,328)	2,166
Test dataset	7714 (1,213/1,165)	453

Table 6.1: Data for the ellipsed predicate interpretation method: sentences and NCs

6.3 Data

6.3.1 Data Processing

To test our method, we extracted 2,166 NC types from the Wall Street Journal (WSJ) component of the Penn Treebank. We additionally extracted sentences containing the head noun and modifier in pre-defined constructional contexts from the amalgam of: (1) the Brown Corpus subset contained in the Penn Treebank, (2) the WSJ portion of the Penn Treebank, and (3) the British National Corpus (BNC).

Two annotators tagged the 2,166 NC types independently at 52.3% inter-annotator agreement, and then met to discus all contentious annotations and arrive at a mutually-acceptable gold-standard annotation for each NC. The Brown, WSJ and BNC data was pre-parsed with RASP, and sentences were extracted which contained the head noun and modifier of one of our 2,166 NCs in subject or object position, or as (head) noun within the NP of an PP. (e.g. *The car belonged to that family.* when the NC is *family car*). After extracting these sentences, we counted the frequencies of the different modifier–head noun pairs, and filtered out: (a) all constructional contexts not involving a verb contained in WordNet 2.0, and (b) all NCs for which the modifier and head noun did not co-occur in at least five sentential contexts. This left us with a total of 453 NCs for training and testing. The combined total number of sentential contexts for our 453 NCs was 7,714, containing 1,165 distinct main verbs.

We next randomly split the NC data into 80% training data and 20% test data. The final number of test NCs is 88; the final number of training NCs varies depending on the verb-mapping method.

The relational forms for all SRs except for PROPERTY, TIME and EQUATIVE—a total of 17 SRs—were manually defined according to the SR definitions provided by Barker and Szpakowicz (1998). The relational forms are based on the definition of the seed verbs and argument structure. For TIME and EQUATIVE, we do not have any corresponding sentential forms from which to extract seed verbs. As a result, we need a separate mechanism for interpreting the TIME and EQUATIVE SRs. For TIME, we employed CoreLex. That is, when the modifier belongs to the TIME class in CoreLex, we assign the TIME class to the NC. For the EQUATIVE SR, we separately gathered NPs where the head noun and modifier are both heads in a coordinate structure, and calculate a score based on Equation 6.6:

$$NC_i = -log_2(\frac{\sum NC_i\ in\ Conjunction}{\sum M\ in\ NC_i * \sum H\ in\ NC_i}) \tag{6.6}$$

NC_i in Conjunction means that the modifier and head noun occur in conjunction form such as *coach and player*. Note that due to the coordinated semantics, this is the only SR where the order of the head noun and modifier is flexible (e.g. *player coach* is considered semantically identical to *coach player*):

(6.7) **NC:** *player coach*
relation: EQUATIVE
logical form: `eq(M,H) = eq(H,M)`
gloss: "player and coach"

The relation PROPERTY can be considered to be a supertype of several relations including MATERIAL (e.g. *apple pie*), and no instances were found of an NC which was PROPERTY and not another SR. As a result, we discard the SR in our evaluation, and restrict the set of SRs to 19.

As mentioned above, in one form of evaluation, we weight the instance based on the following equation:

$$Weight(V_j) = \frac{\sum_{i=1}^{n}(H_i, V_j)}{\sum_{k=1}^{m}\sum_{i=1}^{n}(H_i, V_k)} \tag{6.8}$$

where V_j is a verb from either the 57 or 84 verb set, and H_i is one of the 231 unique head nouns from the test NCs. Our motivation behind *Weight* is that we observed

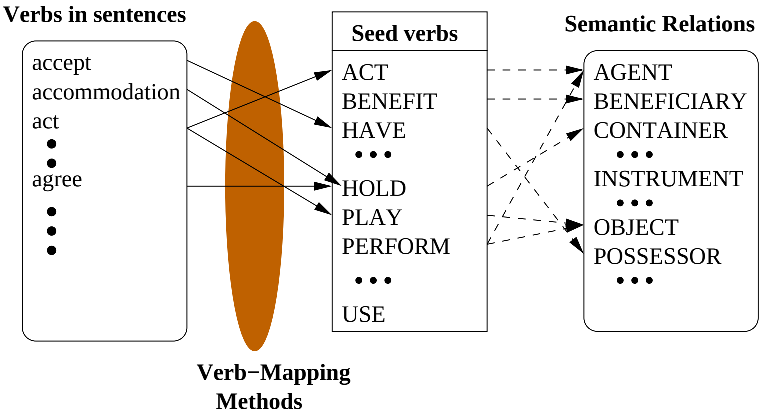

Figure 6.5: Verb-mapping method

that certain head nouns occur more often with particular seed verbs. For example, head nouns with semantics of FOOD are often used with seed verbs such as *make* (e.g. *apple pie*) and *locate* (e.g. *California wine*). That indicates the correlation between the head noun and seed verb. Hence, we calculate the weight by checking the distribution of a pair of a head noun and a seed verb over the total number of these pairs.

6.3.2 Verb-mapping method

The sentential contexts gathered from corpus data contain a wide range of verbs, not just the seed verbs. To map the verbs onto seed verbs, and hence estimate which semantic relation(s) each is a predictor of, we experimented with two different methods. First we used the WordNet::Similarity package to calculate the similarity between a given verb and each of the seed verbs, experimenting with the 5 similarity methods of WUP, JCN, RANDOM, LESK and VECTOR. Second, we used the Moby Thesaurus to extract both direct synonyms a combination of direct and second-level indirect synonyms of verbs, and from this, calculate the closest-matching seed verb(s) for a given verb.

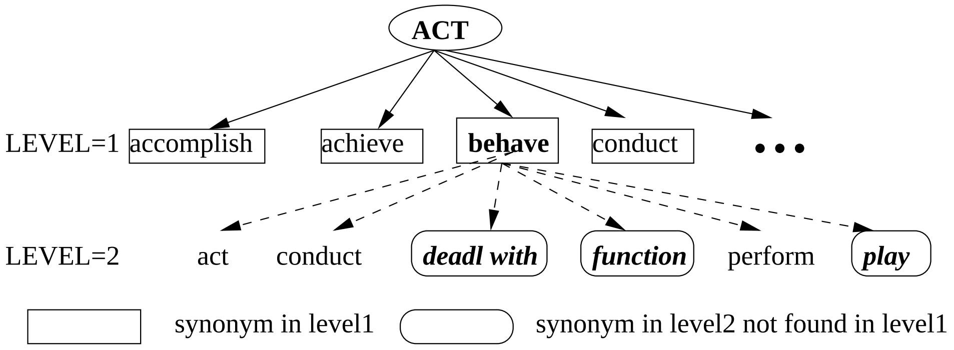

Figure 6.6: Expanding verbs via synonyms

Figure 6.5 depicts the procedure for mapping verbs in constructional contexts onto the seed verbs. Verbs found in the various contexts in the corpus (on the left side of the figure) map onto one or more seed verbs, which in turn map onto one or more semantic relations.[1] We replace all non-seed verbs in the corpus data with the seed verb(s) they map onto, potentially increasing the number of corpus instances.

Out of the similarity methods experimented with, WUP, JCN and RANDOM (WordNet), as well as Direct-Synonyms and Direct- and Indirect-Synonyms (Moby Thesaurus) return one or more seed verbs close to the original verb, while LESK and VECTOR return a unique seed verb. As a result, after mapping the actual verbs onto seed verbs, the number of modified sentences potentially increases with the first five methods.

Since direct (i.e. level 1) synonyms from Moby Thesaurus are not sufficient to map all verbs onto seed verbs, we also include second-level (i.e. level 2) synonyms, expanding from direct synonyms. Table 6.2 shows the coverage of sentences for test NCs, in which **D** indicates direct synonyms and **I** indicates indirect synonyms.

[1] There is only one instance of a seed verb mapping to multiple semantic relations, namely *perform* which corresponds to the two relations AGENT and OBJECT.

# of seed verbs	D-synonyms	D+I-synonyms
57	6,755 (87.57%)	7,388 (95.77%)
84	6,987 (90.58%)	7,389 (95.79%)

Table 6.2: Coverage of direct (D) and indirect (I) synonyms in WordNet

			57		84	
Verb Semantics	Method	Baseline	Count	Weight	Count	Weight
Similarity	WUP	42.35%	32.47%	32.05%	40.66%	42.47%
	JCN	42.35%	40.85%	40.85%	**47.06**%	42.62%
	RANDOM	42.35%	37.97%	37.18%	18.46%	25.97%
	LESK	42.35%	41.67%	41.67%	43.90%	45.71%
	VECTOR	42.35%	**46.67%**	**46.67%**	41.38%	**52.63**%
Thesaurus	D-Synonym	42.35%	33.75%	33.75%	31.76%	34.18%
	D,I-Synonym	42.35%	33.78%	34.21%	33.33%	40.62%

Table 6.3: Results with 17 SRs

6.4 Experiment (I): Ellipsed Predicate Method

We evaluated our method over both 17 semantic relations (without EQUATIVE and TIME) and the full 19 semantic relations, due to the low frequency and lack of verb-based constructional contexts for EQUATIVE and TIME. Note that the test data set is the same for both sets of semantic relations, but that the training data in the case of 17 semantic relations will not contain any instances for the EQUATIVE and TIME relations, meaning that all such test instances will be misclassified. The baseline for all verb mapping methods is a simple majority-class classifier, which leads to an accuracy of 42.35% for the TOPIC relation. We experiment separately with the two data representations for our instances (*Count* and *Weight* — see Section 6.2). Tables 6.3 and 6.4 show the performance of our method over the two sets of SRs.

Overall, VECTOR, JCN and LESK are the strongest-performing methods. In particular, the combination of *Weight* and VECTOR performed best, suggesting that it

			57		84	
Verb Semantics	Method	Baseline	Count	Weight	Count	Weight
Similarity	WUP	40.91%	31.58%	31.17%	41.38%	43.94%
	JCN	40.91%	42.03%	42.03%	**47.06%**	44.64%
	RANDOM	40.91%	38.46%	37.66%	20.00%	28.00%
	LESK	40.91%	44.00%	44.00%	41.46%	48.65%
	VECTOR	40.91%	**46.67%**	**46.67%**	41.38%	**52.63%**
Thesaurus	D-Synonym	40.91%	35.06%	35.06%	32.14%	35.62%
	D,I-Synonym	40.91%	33.78%	34.21%	33.33%	39.34%

Table 6.4: Results with 19 SRs

is an advantage to be attuned to the corpus of interest.

While an increased number of seed verbs generates more training instances under verb mapping, it is imperative that the choice of seed verbs be made carefully so that they do not introduce noise into the classifier and reduce overall performance. Figure 6.7 is an alternate representation of the numbers, with results for each individual method over 57 and 84 seed verbs juxtaposed for each of *Count* and *Weight*. From this, we get the intriguing result that *Count* generally performs better over fewer seed verbs, while *Weight* performs better over more seed verbs. The overall finding is that although more seed verbs tends to increase the dataset noise, the larger number of instances helps to increase the performance.

6.5 Experiment (II): Comparison with Constituent Similarity Method

As noted in Section 6.3.1, we excluded all NCs for which we were unable to find at least 5 instances of the modifier and head noun in an appropriate sentential context. This exclusion reduced the original set of 2,166 NCs to only 453, meaning that the proposed method is unable to classify up to 80% of NCs. For real-world applications,

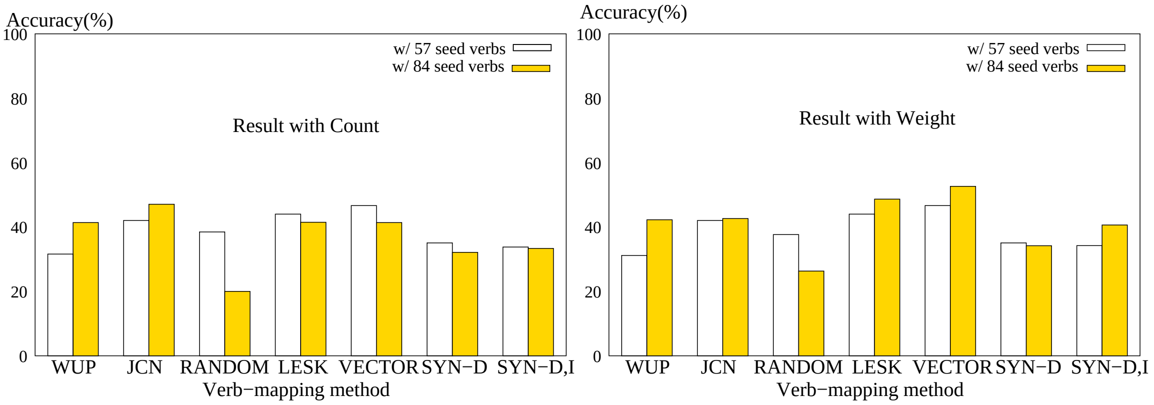

Figure 6.7: Performance with 57 vs. 84 seed verbs over the two data representations

Group	Measure	Baseline	57	Baseline	84
Path	WUP	43.37%	44.99%	43.37%	47.67%
	LCH	43.37%	42.17%	43.37%	41.67%
Information-content	JCN	44.15%	41.56%	43.37%	40.93%
	LIN	44.15%	33.77%	43.37%	34.94%
Random	RANDOM	43.37%	40.96%	42.85%	22.62%
Relatedness	LESK	**47.76%**	46.97%	43.83%	46.58%
	VECTOR	42.85%	34.48%	44.44%	33.33%

Table 6.5: Results in combination with the constituent similarity method with 17 SRs

a method which is only able to arrive at a classification for 20% of instances is clearly of limited utility, and we need some way of expanding the coverage of the proposed method. This is achieved by adapting the constituent similarity method introduced in Section 5.1 to our task, wherein we use lexical similarity to identify the nearest-neighbour NC for a given NC, and classify the given NC according to the classification for the nearest-neighbour.

Tables 6.5 and 6.6 show the performance of the combined method. The overall results pattern similarly to those for the standalone method method. Namely, vector and lesk achieve the best performance, with minor variations in the absolute per-

Group	Measure	Baseline	57	Baseline	84
Path	WUP	41.86%	46.51%	41.38%	47.13%
	LCH	41.86%	41.86%	41.38%	41.38%
Information-content	JCN	43.03%	41.77%	41.86%	40.70%
	LIN	43.03%	34.18%	41.86%	34.88%
Random	RANDOM	41.86%	23.26%	41.38%	21.84%
Relatedness	LESK	**47.76%**	46.27%	43.83%	46.58%
	VECTOR	41.38%	34.48%	42.67%	32.00%

Table 6.6: Results in combination with the constituent similarity method with 19 SRs

formance relative to the original method but the best results for each relation set actually dropping marginally over the original method. This drop is not surprising when we consider that we use an imperfect method to identify the nearest neighbour for an NC for which we are unable to find corpus instances in sufficient numbers, and then a second imperfect method to classify the instance. We predict that the method would produce better results over larger datasets. Even the performance of the constituent similarity method over the same data set supports this analysis. That is, the best performance using the constituent similarity method is 47.6% with the WUP measure. As the performance of the constituent similarity method is less than that of ellipsed predicate method, when it is combined with the ellipsed predicate method, the overall performance of the combined method decreases due to the lower performance of the integrated method (i.e. the constituent similarity method).

Compared to previous work, our method produces reasonably stable performance when operated over the open-domain data with small amounts of training data. Rosario and Marti (2001) achieved about 60% using a neural network in a closed domain, Moldovan *et al.* (2004) achieved 43% using word sense disambiguation of the head noun and modifier over open domain data, and Kim and Baldwin (2005) produced 53% using lexical similarities of the head noun and modifier (using the same relation set, but evaluated over a different dataset). The best result achieved by our system was 52.63% over open-domain data, using a general-purpose relation set. Note that Girju (2007) achieved considerably higher performance of 74-78% using a cross-lingual corpus and 22 SRs. However, her method requires considerable effort to construct the corpus and cross-lingual linguistic mappings.

6.6 Chapter Summary

In this chapter, we have introduced a method for automatically interpreting NCs based on analysis of the ellipsed predicate associated with a given noun pairing. Our method literally interprets NCs using their original definitions.

Our proposed method is heavily reliant on verb semantics, in the form of sets of seed verbs associated with a given SR, and argument structure, to map the arguments of a given verb onto the SR. Since the seed verb sets lacked coverage, we introduced a verb mapping method to map the actual verbs onto the seed verbs and their associated relational mappings. In this, we tested two methods: WordNet::Similarity, and Moby Thesaurus. Finally, we built a supervised classifier to evaluate our method. The results showed that our method worked best in combination with the VECTOR, JCN and LESK verb-mapping methods (best performance = 52.63%).

Our contributions in this chapter can be summarized as follows:

- We proposed a method for interpreting noun compounds using ellipsed predicates retrieved from verb semantics;

- We proposed a verb mapping method in order to map actual verbs onto seed verbs, and investigated the effectiveness of different verb mapping methods;

- We achieved an accuracy of 52.63% with 84 seed verbs using the VECTOR similarity measure and instance weighting (i.e. *Weight*);

- We combined our method with the constituent similarity method in order to achieve full coverage.

Chapter 7

MWEs and Substitutability

As introduced in Section 3.3, *substitutability* is a widely-used NLP technique of substituting in candidate words for a given target word (usually in context) to analyse the relative fit of each. Substitutability relates closely to semantic similarity: when one lexical item is substitutable for another, it is often the case that the two items are semantically and syntactically similar. The degree of substitutability can be used as a tool to classify lexical items.

In this chapter, we present two modeling tasks that employ substitutability. The first, called the *constituent substitution* method, interprets NCs using substitutability and bootstrapping. We apply this method to NC interpretation tasks, including using data from SemEval-2007. The second task disambiguates word senses of component nouns in NCs. We present details of these two tasks including the techniques and algorithms used. We finally summarize the approaches and how they utilize substitutability.

7.1 Constituent Substitution for Interpreting Noun Compounds

We first look at the constituent substitution method in the context of interpreting NCs using substitutability and bootstrapping. Based on sense collocations, we replace one component at a time to generate similar NCs which share the same SR with the original NC. We then evaluate semantic similarity-based methods to gauge the genuine performance relative to a benchmark system. Further, we investigate the performance of existing methods based on semantic similarity over data from SemEval-2007. The evaluation of the selected methods takes place in the context of both 2-way and 7-way classification. We also provide an analysis of hybrid approaches and why they have superior performance.

7.1.1 Motivation for Constituent Substitution

NCs have notoriously high productivity, which presents difficulties for NLP (Lapata 2002). Despite this, however, naturally-occurring NCs tend to be restricted by the way in which semantic relations are constructed. Moldovan *et al.* (2004) exploited this property of NCs to interpret them. We combine this with the method of semantic similarity (i.e. implicit sense collocation) to investigate the interaction between sense collocation and NC interpretation.

In this study, we hypothesize that when one component of an NC is replaced by a similar word, its SR remains the same. This is because the newly-generated NC has the same or similar sense collocation to the original NC. For example, the SR of *horse doctor* is OBJECT where the modifier *horse* is acted upon by the head noun *doctor*. When the modifier *horse* is replaced by a *hypernym* such as *animal*, we are able to predict the SR of the resultant NC *animal doctor* as OBJECT from the original NC. We hypothesize that the same should hold for *synonym*s and *sister word*s. For example, the SR of *lemon juice* is MAKE which denotes that the head noun *juice* is made from the modifier *lemon*. When we replace the modifier *lemon* with its *sister word lime*, we can still correctly predict the SR as MAKE.

(7.1) shows the sense collocation of the original NC and new NCs which are generated by replacing similar words, namely *synonym*s, *hypernym*s and *sister word*s respectively:[1]

[1] Based on synset data in WordNet2.1.

(7.1)

(a) automobile factory

- **Synonym:** car factory
- **Hypernym:** vehicle factory
- **Sister:** truck factory

(b) automobile factory

- **Synonym:** automobile mill
- **Hypernym:** automobile plant
- **Sister:** automobile mint

In (7.1), we replace one noun at a time since we want to restrict the degree of semantic variation. In (7.1a), we replace the modifier with related words, and in (7.1b) we replace the head noun with related words. By replacing a constituent with a synonym, we generate *car factory* and *automobile mill*, both of which have exactly the same sense collocation as the original NC (as the replaced constituents belong to the same semantic class as the original words). On the other hand, *vehicle factory*, *automobile plant*, *truck factory*, and *automobile mint*, whose constituents are replaced by a hypernym in the first two instances and sister word in the second two instances, have different sense collocations to the original NC and yet maintain the same semantic relation of TOPIC.

Notice in (7.1) that with *automobile mint*, we generate what is a pragmatically-marked NC with a semantically-plausible interpretation. That is, our world knowledge about how mints operate and what they produce makes the generated NC sound unnatural, but if we can get beyond this to interpret the NC, the semantic relation we come up with is plausible. To deal with cases like this, our proposed method checks all generated NCs against a pre-compiled list of attested NCs and filters out anything which has not been observed in corpus data.

A second more serious concern is that we will produce an NC which violates our basic assumption about the semantic relation being preserved across controlled semantic variation of the constituent words. For example, an alternate sister of *factory* is *recycling plant*, leading to *car recycling plant*. Here, the most natural interpretation is OBJECT rather than TOPIC. We detail a filtering method for dealing with such examples in the next section.

7.1.2 System Architecture for the Constituent Substitution Method

In this section, we take a detailed look at the proposed architecture for the constituent substitution method. We also present two hybrid approaches we developed by combining selected NC interpretation methods with the constituent substitution method, based on the observation that the constituent substitution method can be used to automatically expand the number of training instances.

The two hybrid approaches utilize existing frameworks or techniques. The first hybrid method (hybrid) combines the sense collocation method of Moldovan *et al.* (2004), constituent similarity method of Kim and Baldwin (2005) (Section 5.1) and constituent substitution co-training of Kim and Baldwin (2007d). We do not manipulate the original methods but cascade them, processing training data generated by the constituent substitution method on each step. Note that constituent substitution co-training is essentially the application of constituent substitution to the generation of training instances. Unlike constituent substitution, the procedure is repeated only once to avoid overgeneration.

The second hybrid method (called constituent similarity co-training: $\mathsf{CSim}_{\mathrm{CT}}$) combines the constituent similarity method with bootstrapping. A detailed architecture of these hybrid methods is presented below.

Architecture of Constituent Substitution Method

Figure 7.1 shows the system architecture of constituent substitution. In the first step, we start with a small amount of seed NCs whose SRs are manually tagged. We generate new NCs by replacing a component of each NC with corresponding word(s) (*synonyms*, *hypernyms* or *sister words*). Note that on each iteration, we replace only one component in order to avoid extreme semantic variation (to keep the sense collocation within a certain range). We then filter the new NCs based on a *word*-level similarity threshold between the original noun and its substitute, using WordNet::Similarity. For those NCs that remain after filtering, we apply a second filter to remove false positive NCs, simply by requiring that the newly-

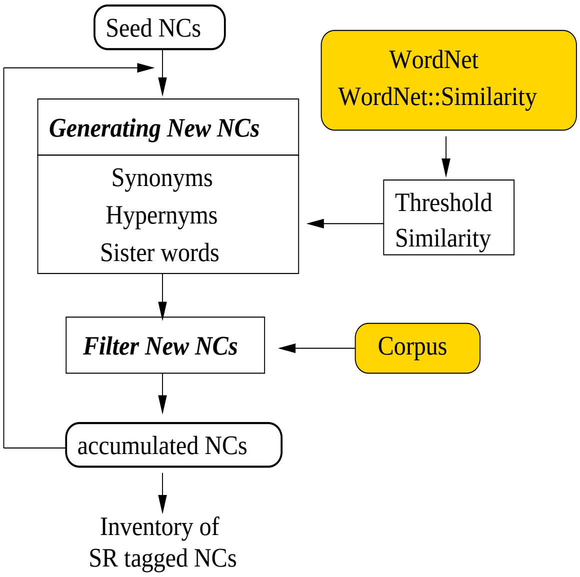

Figure 7.1: Architecture of the constituent substitution method

generated NCs are instantiated in corpus data in sufficient numbers. The details of the corpus used for the secondary filtering can be found in Section 7.1.3.

To set the similarity threshold, we analyzed the distribution of noun pair similarity values. We randomly selected 200 nouns from our dataset and calculated the word-level similarity for all resulting noun pairs based on the method described in Section 5.1.3 and the WUP similarity measure. Figure 7.2 graphs the resulting similarity distribution, and Table 7.1 details the average and selected threshold similarities. The graphs show the range of the similarity with the WUP method across the three relation types.

Hybrid Method I

Figure 7.3 presents the architecture of the first hybrid method, which interleaves sense collocation and constituent similarity, and includes co-training for each. There are five steps in total. In each step, we have *tagged* and *not tagged (NT)* NCs as the

Relation	Average	Threshold
Synonym	.94	.90
Hypernym	.89	.85
Sister word	.83	.80

Table 7.1: Average similarity and selected threshold value for each ontological relation type

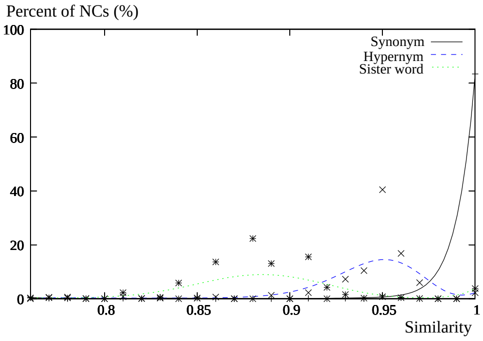

Figure 7.2: Range of similarity values over different ontological relation types

output.

First, we apply the basic sense collocation method (which uses the word sense pair of modifier and head noun) relative to the original training data. If a given test instance has the same sense collocation as a training instance, we judge the predicted SR to be correct.

Second, we apply the constituent similarity method described in Section 5.1 over the original training data, with the qualification that we only classify test instances where the final similarity is above a threshold of 0.8.

Third, we apply the constituent substitution co-training method and re-run the sense collocation method over the expanded training data from the first two steps. Since the sense collocations in the expanded training data have been varied through the advent of *hypernyms* and *sister words*,[2] the number of sense collocations in the

[2] Note that *synonym*-derived training instances do not provide extra sense collocation data, due

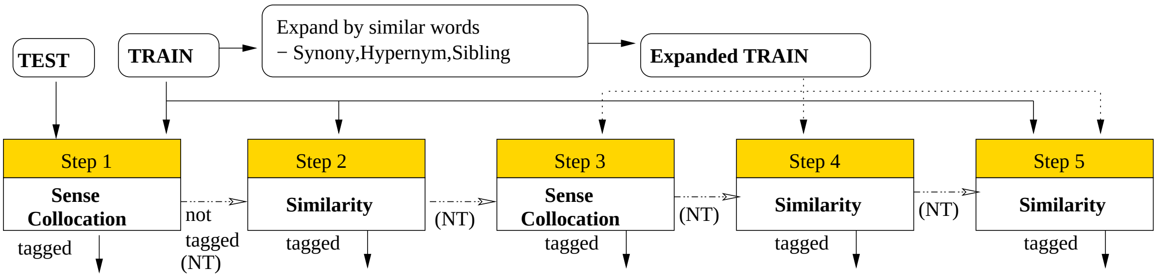

Figure 7.3: Architecture of Hybrid Method I

expanded training data is much greater than that of the original training data (937 vs. 16,676). For details of the data used in evaluation, see Section 7.1.3.

Fourth, we apply the constituent similarity co-training method over the consolidated training data (i.e. the original and expanded training data) from both the constituent similarity method and bootstrapping, with the threshold unchanged at 0.8.

Finally, we apply the constituent similarity method over the combined training data, without any threshold (to guarantee an SR prediction for every test instance). However, since the generated training instances are more likely to contain errors, we decrement the similarity values for generated (i.e. expanded) training instances by 0.2, as we prefer predictions based on the original training instances.

Hybrid Method II : Constituent Similarity Co-training Method

Figure 7.4 depicts our second hybrid system, which is based solely on the constituent similarity method with co-training via bootstrapping.

We perform iterative co-training using bootstrapping, with the slight variation that we hold off reducing the threshold if more than 10% of the test instances are tagged in a given iteration, giving other test instances a chance to be tagged at a higher threshold level relative to newly generated training instances. The residue of test

to the nature of WordNet: they simply provide additional examples of an existing combination of WordNet synsets.

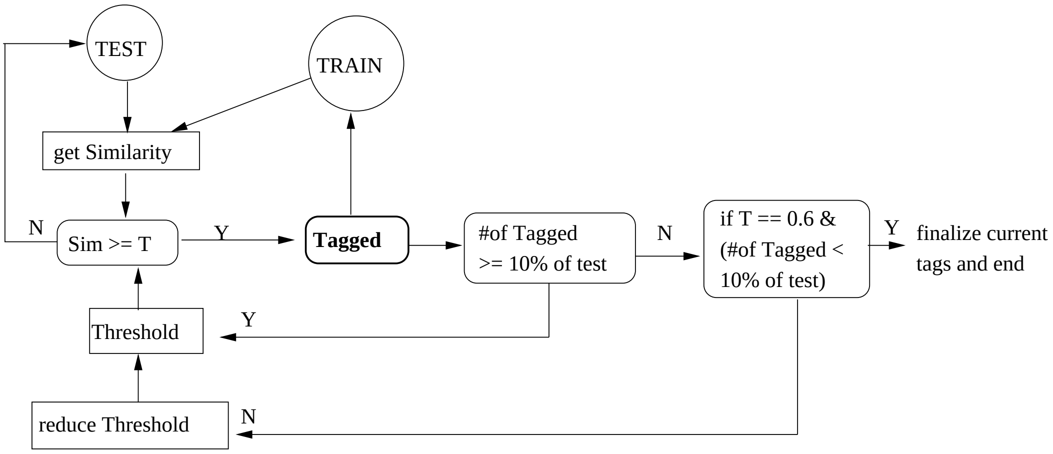
TEST
TRAIN
get Similarity
N
Sim >= T
Y
Tagged
#of Tagged >= 10% of test
N
if T == 0.6 & (#of Tagged < 10% of test)
Y
finalize current tags and end
Threshold
Y
reduce Threshold
N

Figure 7.4: Architecture of Hybrid Method II: the constituent similarity co-training method

instances on completion of the final iteration (threshold = 0.6) are tagged according to the best-matching training instance, irrespective of the magnitude of the similarity.

7.1.3 Data Collection

Data for Constituent Substitution Method

As our first dataset, we randomly gathered and annotated two sets of NCs: a 200 NC seed set and a 400 NC test set. The seed set is used to automatically generate NCs. The test set (which is disjoint with the seed set) is used to examine the quality of the NC interpretations derived by our method. That is, we use the test data to evaluate the performance of existing interpretation methods over the NC data we generate. The NCs in both sets were sourced from the Wall Street Journal. SENSEVAL 2/3 and SEMCOR were obvious alternate candidates given the sense annotations were available, but were unsuitable due to the small number of NC types.

To get the SRs, two human annotators tagged both the SRs and word senses for the NCs in the two sets and met to resolve any annotation disagreements (these NCs were included in the experiments). Over the two data sets, the initial inter-annotator agreement for tagging SRs was 52.3% and the inter-annotator agreement for tagging word senses was 58.8%.

The corpus filter used to exclude unattested NCs uses NCs extracted from the combination of the British National Corpus and Reuters Corpus using a full text chunker (Baldwin and Tanaka 2004). From these NCs, we took only noun-noun bigrams adjoined by non-nouns (e.g. ... *the apple pie was* ...) to ensure that they were not part of a larger compound nominal. We additionally measured the entropy of the left and right contexts for each noun-noun type, and filtered out all compounds where either entropy value was < 1.[3] This was done in an attempt to, once again, exclude NCs which were embedded in larger multiword expressions, such as *service department* in *social service department*. The combined total number of unique NC types extracted from the two corpora was $389,400$.

Data for Benchmarking NC interpretation

As our second dataset, we used data from SemEval-2007. The original data contains 7 relations: Cause-Effect (CE), Instrument-Agency (IA), Product-Producer (PP), Origin-Entity (OE), Theme-Tool (TT), Part-Whole (PW), and Content-Container (CC). The task in the SemEval-2007 competition was to identify the compatibility of a given SR with each test instance. That is, for each NC an SR candidate is provided for which a binary compatibility prediction must be made. The data additionally provides word senses from WordNet 3.0. It also provides a query per instance, that is a manually-generated patterns used to perform a web search for sentences that are positive examples of the given relation.

In this work, we used the SemEval-2007 data for two sets of experiments. First, we used the original data directly as defined in the SemEval-2007 task (**binary** or **2-way classification**). Second, we combined together the positive instances for each SR into a single set of training and test instances, and carried out evaluation in a manner compatible with our datasets (**multiple** or **7-way classification**).

[3] For the left token entropy, if the most-probable left context was *the*, *a* or a sentence boundary, the threshold was switched off. Similarly for the right token entropy, if the most-probable right context was a punctuation mark or sentence boundary, the threshold was switched off.

SR	Definition	Examples
CE	N_1 is the cause of N_2	*virus flu, hormone growth*
IA	N_1 is the instrument of N_2, N_2 uses N_1	*laser printer, ax murderer*
PP	N_1 is a product of N_2, N_2 produces N_1	*honey bee, music clock*
OE	N_1 is the origin of N_2	*bacon grease, desert storm*
TT	N_2 is intended for N_1	*reorganization process, copyright law*
PW	N_1 is part of N_2	*table leg, daisy flower*
CC	N_1 is store or carried inside N_2	*apple basket, wine bottle*

Table 7.2: The set of 7 semantic relations, where N_1 is the head noun and N_2 is a modifier

	Binary			Multiple		
SR	Test	Training	Ext. Training	Test	Training	Ext. Training
CE	80	136	2,588	36	71	1,854
IA	78	135	1,400	36	68	1,001
PP	93	126	2,591	55	78	2,089
OE	81	136	3,085	35	52	1,560
TT	71	129	2,994	27	50	1,718
PW	72	138	2,577	28	64	1,510
CC	74	137	2,378	37	63	1,934
Total	549	937	17,613	254	446	11,664

Table 7.3: Breakdown of the SemEval-2007 data across the two tasks and 7 SRs

Table 7.3 describes the number of test and training instances for each of the binary and multiple classification datasets.

We further analyzed the data and found that in the original dataset, only 5 NCs occur as candidate instances for multiple SRs. Among these 5 instances, only two (i.e. *axe carpenter* and *drill dentist*) occur as positive instances for more than one SR. That is, (at least) these two NCs can be interpreted differently according to context, while we can assume that the remainder of the NCs have a single interpretation relative to the provided SRs. Another issue with the SemEval-2007 dataset is that some of the NCs are part of ternary or higher-order NCs (81 training and 40 test NCs for

modifier	head noun		our NC
billiard table	room	→	*table room*
body	bath towel	→	*body towel*

Table 7.4: Extracting binary NCs from higher-order NCs

the binary dataset, and 42 training and 24 test NCs for the multiple dataset). With these NCs, we extracted a binary NC based on bracketing, in the manner depicted in Table 7.4.

Despite these slight reservations about the comparability of the dataset with other data used in this research, we experiment with it in the interests of maximizing comparability with other research.

7.1.4 Experiment (I): Interpreting NCs by the Constituent Substitution Method

We evaluate the constituent substitution method from two perspectives. First, we evaluate the quantity of NCs which we are able to correctly interpret. Second, we focus on the usability of the acquired NCs. Since the constituent substitution method is based on sense collocation, the newly-acquired NCs are also annotated with sense information. Hence, we can confirm the utility of the acquired NCs relative to the accuracy of the sense predictions. Finally, we benchmark two existing non-hybrid and hybrid methods using the data from SemEval-2007.

The first experiment tries to automatically acquire SRs using the constituent substitution method. We start with a small seed set of manually-tagged NCs. Tables 7.5–7.7 show the performance of the constituent substitution method over the three different lexical relations: *synonym*s, *hypernym*s and *sister word*s. Note that for *synonym*s and *sister word*s, the number of new NCs becomes saturated when we iterate a number of times.[4]

[4] The fact that we generate new NCs at all on the second iteration is a product of us only

In evaluating the method, we calculate the following numbers, which we report in Tables 7.5–7.7: Note that the errors are checked manually from the positive NCs (A) and sample (D).

- New NCs (A) = number of newly-derived NCs
- Positive NCs (B) = derived NCs positively attested in the corpus, i.e. found in the NC set collected from British National Corpus and Reuters Corpus
- Error in (B) = ratio of incorrect tagged NCs in (B)
- Filtered NCs (C) = NCs not attested in the corpus and filtered out, i.e. (A) - (B)
- Sample (D) = test sample (sub-sample of (C)); about 20% of the data was selected as the test sample
- Negative NCs in (D) = ratio of non-NCs in (D), i.e. bogus NCs due to overgeneration
- SR Error in (D) = ratio of incorrectly tagged NCs in (D)
- All Error in (D) = total error ratio in (D).
- Total Error in (A) = ratio of NCs tagged incorrectly in (A), i.e. errors in (B) and errors in (C), factoring out errors in (D)

We predictably derived the largest number of NCs using *sister word*s. After the first iteration, starting with 200 seed NCs, we generated 384,616 and 10,728 NCs using *synonym*s, *hypernym*s and *sister word*s, respectively. Even when we used only the positive NCs found in the corpus (strict attestation filtering), the numbers of newly generated NCs are 69, 100, and 686, respectively. As such, we can see that the constituent substitution method is successful in automatically deriving a large number of NCs.

substituting for one word at a time, such that *automobile factory* and *car plant* will be generated from *car factory* on the first iteration, from each of which we will generate *automobile plant* on the second iteration. Similar combinatorics apply in the case of sister words.

	I1	I2
New NCs (A)	384	229
Positive NCs (B)	69	13
Error In (B)	21.74%	7.69%
Filtered NCs (C)	315	216
Sample (20%) (D)	63	43
Negative NCs in (D)	7.94%	13.95%
SR error in (D)	27.59%	18.92%
Negative NCs & SR error in (D)	33.33%	30.23%
Total error In (A)	31.25%	30.56%

Table 7.5: Analysis of the performance of constituent substitution with *synonym*s across two iterations

When we look at the error rate of the newly derived NCs in Tables 7.5–7.7, we observe that the NCs generated via *synonym*s contain less errors than the other two ontological relations. However, from an interpretation point of view, the accuracy in using *hypernym*s and *sister word*s is relatively high compared to the accuracy reported for previous methods, namely 43.2% (Moldovan *et al.* 2004) and 53.0% (Kim and Baldwin 2006b), as compared to between 64.7% and 70.8% after the first iteration of the constituent substitution method. In addition, this method needs only a small number of manually-tagged NCs (200 NCs in this experiment) compared to prior approaches. As a result, we strongly believe that the constituent substitution method can efficiently derive NCs tagged with SRs with lower error rates and reduced human effort. Also, it has the potential to significantly increase the volume of automatically interpreted NCs.

Note that when we use *hypernym*s and *sister word*s we go beyond the scope of the original semantic collocation, so that we can significantly increase the number of tagged NCs by iterating repeatedly. In Tables 7.5–7.7, we can see that the increase in error rate as we get further and further away from the original semantic collocations is gradual (linear). Also, in the filtered (unattested) NCs, the error increases gradually, and that in excluding these in constituent substitution method, the error rate is kept

	I1	I2	I3	I4
New NCs (A)	616	997	1,202	1,079
Positive NCs (B)	100	87	70	58
Error in (B)	25.00%	22.99%	40.00%	43.10%
Filtered NCs (C)	516	910	1,132	1,021
Sample (20%) (D)	103	182	226	204
Negative NCs in (D)	0.97%	6.04%	11.06%	18.14%
SR error in (D)	30.39%	40.94%	44.28%	40.08%
Negative NCs & SR error in (D)	30.10%	44.51%	50.44%	47.90%
Total error in (A)	29.22%	42.63%	49.75%	42.74%

Table 7.6: Analysis of the performance of constituent substitution with *hypernym*s across 4 iterations

down, especially with *synonym*s and *sister word*s. Considering that the number of generated NCs is exponential, especially with *hypernym*s and *sister word*s, the total error rate is relatively low. Finally, the results in Table 7.8 show that the error rate for NCs derived via *synonym*s and *sister word*s is linear even after several iterations. However, despite our best efforts at filtering, it is apparent that we require more efficient methods to effectively avoid noise in the generated data. We leave this as an area for future work.

7.1.5 Experiment (II): Constituent Substitution Co-training Method

In a second experiment, we tested the usability of NCs acquired by the constituent substitution method in NC interpretation using two methods: the sense collocation method of Moldovan *et al.* (2004) and the constituent similarity method from Section 5.1. Table 7.9 shows the amount and quality of the data used for the second experiment. Note that *Combined* puts all three similar words together but filters out duplicated NCs.

Table 7.10 shows the performance of the two interpretation methods ("Moldovan"

	I1	I2	I3	I4
New NCs (A)	10,728	192,917	19,275	44
Positive NCs (B)	686	1,751	106	0
Error in (B)	24.05%	30.26%	19.81%	0.00%
Filtered NCs (C)	10,042	191,202	19,169	44
Sample (20%) (D)	502	956	958	44
Negative NCs in (D)	7.97%	14.33%	7.31%	68.18%
SR error in (D)	30.52%	24.79%	36.37%	100.0%
Negative NC & SR error in (D)	36.06%	35.56%	41.02%	100.0%
Total error In (A)	35.28%	35.52%	40.89%	100%

Table 7.7: Analysis of the performance of constituent substitution with *sister word*s across 4 iterations

= the method of Moldovan *et al.* (2004); "Kim" = the constituent substitution method, as presented in Section 5.1) over the generated NCs. Although the number of NCs generated by the constituent substitution method is large, the range of sense collocations we capture is restricted, especially for *synonym*s. The baseline for each method is based on using only the 200 seed NCs as training data, which in itself outperformed a simple majority class baseline (i.e. Zero-R) for both methods tested. Using additional training data generated by the constituent substitution method, we were able to improve over this baseline performance. Although the number of generated NCs is 5 times more than the 200 seed NCs, the performance of Moldovan *et al.* (2004) does not increase significantly. This confirms that idea that Moldovan *et al.* (2004) relies not only on the amount of training data but also on the range of sense collocations in the training data.

We also analyzed the performance of the two methods using error-free (Correct) and "error-full" (All) training data. Despite the smaller amount of training data, the best performance is achieved using the Correct data with Moldovan *et al.* (2004). This is disappointing in terms of the general-purpose NC interpretation task targeted in this thesis, but very promising for domain-specific NC interpretation. Our reasoning here is that the filtering could be tailored to a given domain, such that we could

	Iteration	Error In (B)	Error In (C)	Total error In (A)
Synonym	1	21.74%	33.33%	31.25%
	2	19.51%	32.08%	30.56%
Hypernym	1	25.00%	30.10%	29.22%
	2	24.06%	39.65%	37.51%
	3	28.40%	44.42%	57.26%
	4	31.11%	48.11%	42.74%
Sister word	1	24.05%	36.06%	35.28%
	2	28.49%	35.73%	35.50%
	3	28.12%	37.83%	35.97%
	4	28.12%	38.94%	35.67%

Table 7.8: Accumulated error in Experiment I

Relation	Correct	All	SR Error
Synonym	338	813	22.88%
Hypernym	668	3,015	39.90%
Sister	1,042	10,928	34.64%
Combined	1,648	14,356	32.47%

Table 7.9: Dataset sizes and quality in Experiment II

generate NCs with higher lexical similarity to NCs in that domain, and hence have a higher chance of producing the correct classification.

Finally, we evaluated the performance of interpretation methods relative to comparable amounts of training data generated using the three different ontological relation types (Table 7.11). We generated the respective training datasets by stopping the given algorithm at a point where it had generated around 600 NC instances, in each case using only corpus-attested NCs. For reference, we reproduce the accuracy of the constituent similarity method using the original dataset of 1,769 hand-tagged instances.[5] We achieved the best performance from *hypernyms* using the constituent

[5] Note that the dataset is not sense-tagged, and as a result, we have shown the performance for our method only, based on word similarity.

Relation	Method	Seed NCs	Correct	All
Baseline	Zero-R	23.00%	–	–
	Moldovan	33.25%	–	–
	Kim	29.75%	–	–
Synonym	Moldovan	–	33.25%	32.50%
	Kim	–	29.00%	29.50%
Hypernym	Moldovan	–	32.50%	32.75%
	Kim	–	30.50%	29.00%
Sister	Moldovan	–	35.50%	34.25%
	Kim	–	29.00%	28.00%
Combined	Moldovan	–	34.75%	34.00%
	Kim	–	29.75%	29.00%

Table 7.10: Results of NC interpretation

Dataset	Size of data (error rate)	Moldovan	Kim
Manual	1769 (0.00%)	–	42.00%
Synonym	613 (30.34%)	32.50%	29.50%
Hypernym	616 (29.32%)	31.75%	**32.75%**
Sister	686 (24.05%)	29.00%	28.00%

Table 7.11: Results of interpretation over differing datasets for the methods of Moldovan and Kim

similarity method, although the difference between methods is not significant. As a result, we conclude that there is no significant difference between the three relation types with respect to the task of NC interpretation for a given quantity of training data.

7.1.6 Experiment (III): Benchmarking NC Interpretation

To directly compare ourselves with prior research on NC interpretation, we test two semantic similarity based methods from Moldovan *et al.* (2004) and Section 5.1 over the SemEval-2007 dataset, as described in Section 7.1.3. Also, we hybridize

System	Method
M	Sense collocation (SC)
ME	Sense collocation (SC) + constituent substitution co-training
K	Constituent similarity (CS)
KE	Constituent similarity (CS) + constituent substitution co-training
hybrid	SC + CS + constituent substitution co-training
CSim$_{\mathrm{CT}}$	Constituent similarity + bootstrapping

Table 7.12: Systems tested over the SemEval-2007 data

the two systems to utilize the constituent substitution method. We then tested the six systems, as detailed in Table 7.12 over both the 2-way and 7-way tasks. Here, group **A4** and **B4** were system categories used in the SemEval-2007 competition, where A4 contains systems that use neither word sense information nor the query, and B4 contains systems that use word sense information but not the query.[6] We categorized our systems into these two groups in order to evaluate the impact of sense information on system performance. For the baseline, we use Zero-R (i.e. a majority class vote). Note that in Table 7.12, M is the sense collocation method, K is the constituent similarity method, E is constituent substitution co-training method, *hybrid* is the first hybrid method and $CSim_{CT}$ is the second hybrid method, as detailed in Table 7.12.

7-way Classification Task

We discuss the overall results for the 7-way classification task in Section 7.1.6 below.

Tables 7.14 and 7.15 show the results at each step for the hybrid and CSim$_{\mathrm{CT}}$ methods, respectively. As each method proceeds, the amount of tagged data increases but the classification accuracy of the system decreases, due to the inclusion of increasingly noisy training instances in the previous step.

[6] In the original task, there were also groups C4 and D4 which made use of the query. As none of our methods use the query, they are irrelevant to our discussion here.

Group	Method	Precision	Recall	F-score	Accuracy
–	Majority	–	–	–	21.65
A4	K	51.80	52.21	52.00	52.76
	$\mathsf{CSim}_{\mathrm{CT}}$	51.66	51.13	51.39	52.17
B4	M	70.46	44.37	54.45	49.61
	ME	64.64	46.61	54.16	50.79
	KE	52.32	51.97	52.14	52.76
	hybrid	50.04	50.49	50.26	51.57

Table 7.13: Results for the 7-way classification task

Step	Method	Tagged	Accuracy in Tagged	Untagged
1	SC	12	100.0%	242
2	CS	57	71.93%	185
3	extSC	0	00.00%	185
4	extCS	78	46.15%	107
5	CStt	107	39.25%	0

Table 7.14: A breakdown of results for hybrid over the 7-way classification task, across the component methods (SC = sense collocation; CS = constituent similarity; extSC = SC + SC co-training; extCS = constituent similarity + SC co-training; and CStt = the final step over both the original and expanded training data)

We additionally show the performance over the individual SRs by hybrid in Figure 7.5. Note that $\mathsf{CSim}_{\mathrm{CT}}$ had a similar performance profile across the different SR types.

2-way Classification Task

Once again, we discuss the overall results for the 2-way classification task in Section 7.1.6 below.

As with the first experiment, we analyzed the number of tagged instances and the accuracy of the hybrid and $\mathsf{CSim}_{\mathrm{CT}}$ methods, as shown in Table 7.17 and 7.18 respectively. The overall results are similar to those for the 7-way classification task.

Iteration	Threshold	Tagged	Accuracy in Tagged	Untagged
1	0.90	29	89.66%	225
2	0.85	12	75.00%	213
3	0.80	31	61.29%	182
4	0.75	43	53.49%	139
5	0.70	63	53.97%	76
6	0.65	26	34.62%	50
7	<0.65	49	24.49%	1

Table 7.15: A breakdown of results for $\mathsf{CSim}_{\mathrm{CT}}$ over the 7-way classification task, across the individual iterations

Figures 7.6 and 7.7 show the performance for positive and negative classifications for each individual relation. The performance when the classifier outputs are mapped on to the 7-way classification task are similar to those in Figures 7.6 and 7.7.

Discussion of Results

We show a performance comparison of the 6 systems over the 7-way and 2-way classification tasks in Table 7.13 and Table 7.16 respectively. The performance of all the methods exceeded the baseline. The constituent similarity (K) system performed the best in group A4 and the constituent similarity + SC co-training (KE) system performed the best in group B4 for both classification tasks. In general, the performance of constituent similarity is marginally better than that of sense collocation. Also, the utility of constituent substitution co-training is validated by its better performance, outperforming both constituent similarity and sense collocation.

In order to compare the original methods with the hybrid methods, we observed that the original methods, **M** and **K**, and their co-training variants, **ME** and **KE**, performed consistently better than the hybrid methods, hybrid and $\mathsf{CSim}_{\mathrm{CT}}$. That is, the combination of the methods lowers overall performance. We also found that greater numbers of training instances contribute to improved performance, as all methods are supervised. As expected, the step-wise performance of hybrid and $\mathsf{CSim}_{\mathrm{CT}}$ degrades

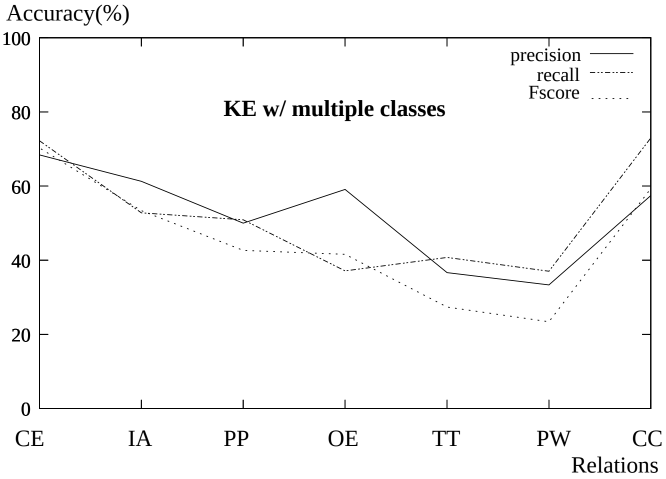

Figure 7.5: Performance across the individual SRs for the 7-way classification task

with each iteration. In the case of $\mathsf{CSim}_{\mathrm{CT}}$, the performance is nearly monotonically decreasing across the iterations (except for iteration 3 = 59.46% to iteration 4 = 72.23%). This shows that the performance of $\mathsf{CSim}_{\mathrm{CT}}$ is heavily influenced by the quality of training instances, which decreases from iteration to iteration.

Comparing our systems with those in the original binary classification in SemEval-2007, the 6 systems were below the best performing system in the respective group, although their performance is competitive. (See the full results of SemEval-2007 in Section C). This is partly because the methods were originally designed for multi-way (positive) classification and have not been customized to the binary task reformulation.

Comparing the SC and CS methods, we found that the methods interpret SRs with 100% accuracy when the sense collocations are found in both the test and training data. However, the CS method is more sensitive to variations in sense collocations than the SC method. This in turn leads to better performance. The CS method interprets NCs with high accuracy when the computed similarity is sufficiently high (e.g. with similarity ≥ 0.9, the accuracy is 89.7%). Another benefit of this method is

Group	Method	Precision	Recall	F-score	Accuracy
–	All True	48.5	100.0	64.8	48.5
–	Probability	48.5	48.5	48.5	51.7
–	Majority	81.3	42.9	30.8	57.0
A4	Best	66.10	66.70	64.80	66.00
	K	63.22	62.79	62.74	65.03
	CSim$_{\mathrm{CT}}$	61.50	55.70	57.80	62.70
B4	Best	79.70	69.80	72.40	76.30
	M	67.21	58.39	54.48	63.41
	ME	60.18	57.09	55.38	61.88
	KE	65.99	65.70	65.42	66.85
	hybrid	61.70	56.80	58.70	62.50

Table 7.16: Results for the 2-way classification task

that it interprets NCs without word sense information. As a result, we can glean that the CS method is a more flexible and robust approach, although one of its weakness is its reliance on the similarity measure.

7.1.7 Summary of the Constituent Substitution Method

We proposed a novel method for automatically interpreting NCs. Our inspiration in carrying out this work is the sense collocation method of Moldovan *et al.* (2004)—i.e. when the sense collocation is the same or similar, NCs share the same SR—which we extend by combining it with bootstrapping. Unlike existing NC interpretation methods, the constituent substitution method derives SR-tagged NCs using sense collocation and bootstrapping, based on the notion of substitutability.

We implemented our method by replacing one of the components in a given NC with a similar word (i.e. *synonym*, *hypernym* or *sister word*) to generate a new NC which we assume to have the same SR as the original. We then check for an occurrence of the newly-generated NC relative in a corpus to verify that it is a plausible noun combination. This process is repeated, including the newly-acquired NCs as training

Step	Method	Tagged	Accuracy in Tagged	Untagged
1	SC	21	80.95%	526
2	CS	106	68.89%	420
3	extSC	0	00.0%	420
4	extCS	61	60.66%	359
5	CStt	359	61.88%	0

Table 7.17: A breakdown of results for hybrid over the 2-way classification task, across the component methods (SC = sense collocation; CS = constituent similarity; extSC = SC + SC co-training; extCS = constituent similarity + SC co-training; and CStt = the final step over both the original and expanded training data)

Iteration	Threshold	Tagged	Accuracy in Tagged	Untagged
1	0.90	21	81.00%	526
2	0.85	73	72.60%	474
3	0.80	56	71.43%	418
4	0.75	74	59.46%	344
5	0.70	101	72.23%	243
6	0.65	222	57.21%	21
7	<0.65	21	99.60%	0

Table 7.18: A breakdown of results for $\mathsf{CSim}_{\mathrm{CT}}$ over the 7-way classification task, across the individual iterations

instances, through the process bootstrapping.

During evaluation, the constituent substitution method showed good results over the NC interpretation task (64.72% - 70.8%). Also, it acquired large numbers of NCs with no human labor. Despite the benefits of the method, one of the problems with this approach is the difficulty in interpreting significantly outside the known range of sense collocations. In future work, we will further examine the proposed method to expand the range of interpreted NCs and the variation of sense collocations.

We also evaluated a number of constituent substitution-based methods over the SemEval-2007 dataset, in the form of both a 2-way and 7-way classification task. We found that the constituent similarity method performed slightly better due to sense

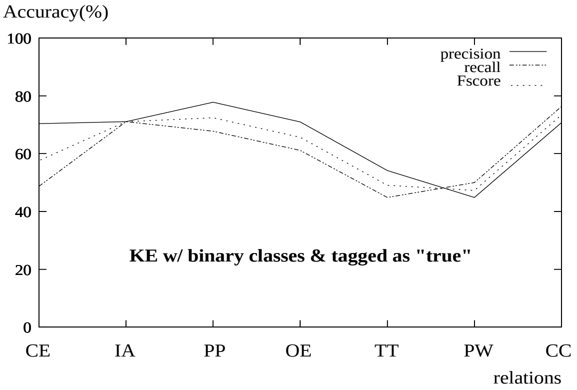

Figure 7.6: Positive task

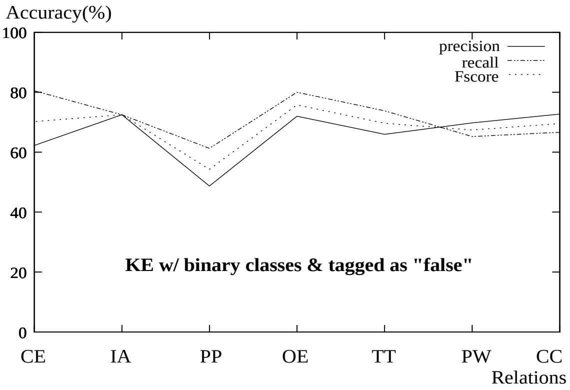

Figure 7.7: Negative task

insensitivity. We also found that constituent substitution co-training is an effective way of increasing the training instances. The performance of our hybrid methods was no better than that of the original methods, implying that more work needs to be done to build effective hybrid approaches.

7.2 Disambiguating Noun Compounds

In this section, we present a novel method for word sense disambiguation (WSD) of components in NCs in order to acquire knowledge for NC interpretation and ultimately improve NC interpretation performance.

Ever since Sparck Jones (1983) showed the importance of word sense in NC interpretation, studies have taken into account the word senses of component nouns in interpreting NCs, and achieved compelling results (Moldovan *et al.* 2004; Kim and Baldwin 2005). However, all the previous work on WSD has focused almost exclusively on simplex words and ignored MWEs. Hence, our primary interest here is to investigate a method for automatically disambiguating word sense in NCs, and analyze the differences between WSD of NCs and simplex words. We also attempt to build a successful real-world NLP application by using the acquired word sense data in NC interpretation tasks.

7.2.1 Background of Word Sense Disambiguation

Word sense disambiguation is the task of resolving the sense of word instances, usually relative to a predefined sense inventory (Agirre and Edmonds 2006). It has been recognized as one of the hardest tasks in NLP (referred to as an *AI-complete* problem). WSD has been suggested as an intermediate task for NLP applications, although at current performance levels, most attempts to incorporate WSD in actual applications have been unsuccessful. For example, Sanderson (1996) identified the potential for WSD to enhanced the performance of information retrieval (IR), but in practice found that it was impossible to achieve the required level of accuracy to achieve that potential gain. Vickrey *et al.* (2005) and Carpuat and Wu (2005), on the other hand, got mixed results for WSD in the context of machine translation (MT).

WSD research is categorized into two primary categories: corpus-based and knowledge-based (Ide and Veronis 1998). Corpus-based methods use features based on neighboring (content) words in a fixed word window of the target word, while knowledge-based methods extract features from lexical resources such as dictionaries. Yarowsky (1995) famously developed a corpus-based WSD method based on bootstrapping based on collocations, while McCarthy *et al.* (2004) proposed a method for learning the first sense of a word based on grammatical context, content words in a word window, and ontological semantics. Leacock and Chodorow (1998) identified monosemous *hypernyms* and *hyponyms* of a given target word, and from these acquired sense-annotated examples automatically, in a knowledge-based technique. Banerjee and Pedersen (2003) showed the usefulness of *hypernyms* in WSD based on dictionary definition overlap. Mihalcea and Moldovan (1999) and Agirre and Martinez (2000) used lexical substitution to perform WSD.

As stated above, however, all mainstream work on WSD has been carried out for simplex words only. Analysis of WSD specifically over MWEs is needed to develop a robust and accurate WSD method. A particular motivation of this claim is that word sense has been shown to be useful in interpreting NCs (Moldovan *et al.* 2004; Kim and Baldwin 2005).

7.2.2 Motivation for Disambiguating Nouns in NCs

The need for sense disambiguation of the elements of NCs motivated this research on WSD of NCs. We observed that the sense distribution of nouns in NCs doesn't necessarily correspond to that in simplex contexts. Additionally, by focusing on the elements of the NC, we can bring the **one sense per collocation** heuristic (Yarowsky 1993) into play, in assuming that the elements of a given NC will always occur with the same sense irrespective of context, just as we explored the hypothesis above that the SR for a given combination of senses in an NC is fixed.

We compare this approach with a standard corpus-based approach, using a state-of-the-art WSD system called SenseLearner (Mihalcea and Faruque 2004). We show that our disambiguation method based solely on word sense combinatorics is more successful at disambiguating word sense than existing methods. Note that we do not dispute the claim that context influences NC interpretation (e.g. (Girju *et al.* 2007)). Rather, for our current purposes we focus exclusively on word sense at the type-level for NCs out of context and leave the harder task of token-level interpretation/WSD for future research.

Sense Distribution of Polysemous Elements in NCs

Our motivating observation for the task is that the word sense distribution of NC constituents is both different to that for simplex usages, and varies across head noun and modifier usages (which we will refer to as "grammatical roles"). For example, *art* as a modifier has a different sense distribution to when it occurs as a head noun. Table 7.19 describes the sense distribution for the words *art* and *day* for each grammatical role within an NC, and also across all usages (NC or otherwise). The word senses and sense glosses are based on WordNet2.1, and the sense distributions are taken from SemCor. According to WordNet2.1, *art* has a total of 4 senses and *day* has 10 senses.

In Table 7.19, the majority of usages of *art* as a modifier occur with sense$_1$ ("the products of human creativity; works of art collectively") while it has a more uniform distribution across all senses as a head noun but still with sense$_1$ as predominant sense.

	art			*day*		
Sense	Modifier	Head noun	Overall	Modifier	Head noun	Overall
ws_1	**.85**	**.62**	**.67**	.13	.04	**.41**
ws_2	.11	.04	.22	.02	.04	.20
ws_3	.00	.03	.08	**.80**	.00	.12
ws_4	.04	.31	.03	.00	**.91**	.20
ws_5	—	—	—	.04	.01	.05
ws_6	—	—	—	.00	.00	.03
ws_7	—	—	—	.00	.00	.00
ws_8	—	—	—	.01	.00	.00
ws_9	—	—	—	.00	.00	.00
ws_{10}	—	—	—	.01	.00	.00

Table 7.19: Sense distribution for *art* and *day* as an NC modifier, head noun and overall in SemCor

This agrees with the majority sense overall in SemCor. *Day* as a modifier is used mostly with sense_3 ("daytime, daylight"), while sense_4 ("a day assigned to a particular purpose or observance") is the majority sense when used as a head noun, and sense_1 ("twenty-four hour period, solar day") is the majority sense overall. These two nouns are representative of the two extremes observed in SemCor: a largely identical sense distribution both within NCs and overall (as with *art*), and a radically different sense distribution across the two grammatical relations within NCs, and also overall (as with *day*). Most nouns are somewhere between these two extremes, that is the three sense distributions vary to some degree. This observation provides the motivation to carry out WSD in a manner specific to NCs and also the grammatical relation with an NC.

Sense Restrictions in NCs

As discussed above, Moldovan *et al.* (2004) used sense collocation to interpret NCs, based on the hypothesis that when the sense collocation of two NCs is same, their SR is most likely also the same. Moldovan *et al.* encoded this hypothesis in the

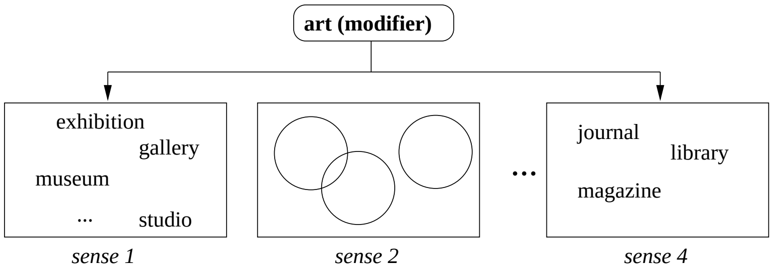

Figure 7.8: Sense restriction due to the semantics of the disambiguated element in NCs

following formulation, based on conditional probability:

$$sr^* = \text{argmax}_{sr_i} P(sr_i | ws(n_1), ws(n_2)) \tag{7.2}$$

where $ws(n_*)$ is the word sense of noun n_*, n_1 is the modifier, n_2 is the head noun and each sr_i is an SR.

Based on the above, we formulate our probabilistic model to disambiguate the word sense of polysemous element in NCs:

$$ws^*(n_i) = \text{argmax}_{ws(n_i)} P(ws(n_i) | ws(n_j), sr) \tag{7.3}$$

where n_i is the target noun to disambiguate, n_j is the remaining noun in the NC (which we are assuming has already been disambiguated), and sr is the SR between the modifier and head noun. Note that we assume we know the sense of the non-target noun n_j (either the head noun or the modifier) as well as the SR in this formulation, and use this to determine the word sense of the target noun n_i (either the modifier or the head noun, respectively). As it is unlikely that we will have reliable access to the sr for a given NC, we modify Equation 7.3 by replacing sr with the grammatical role (either modifier or head noun), to encode the observation from above that the sense distribution of a noun can vary greatly across the two roles. Hence, our final formulation is Equation 7.3 with sr replaced by the grammatical role (as in Equation 7.4). We further experiment with the inclusion of sr to attest the contribution of sr to WSD.

Figure 7.8 depicts sets of nouns which co-occur with *art* as a modifier across the different senses (i.e. 4) of *art*. When *art* occurs with *museum* or *gallery*, its sense is “art, fine art, creation”. On the other hand, when it occurs with words such as *journal* or *magazine*, the sense of *art* is ws_4 (“artwork, art, graphics, non-textual matter”). Note that the semantic groups in each word sense of *art* can be overlapped by words which are shown in sense 2.

One Sense per Collocation

The **one sense per collocation** heuristic was proposed by Yarowsky (1993) as a general bootstrapping method for WSD. It assumes that a word will be used with the same sense within a given word collocation, such as NCs and adjective-noun collocations, across all token occurrences. Yarowsky claims that this heuristic is effective at disambiguating words which occur in collocational contexts, and showed that the accuracy over a range of binary disambiguation bootstrapping tasks was between 90% and 99%. The work also showed that the heuristic was more successful in certain contexts than others. In the case of nouns, the best disambiguating context was directly adjacent adjectives or nouns, underlying the effectiveness of the heuristic for our work.

We draw on the one sense per collocation heuristic to disambiguate constituents in NCs. However, in our case, the heuristic is applied slightly differently to the original in Yarowsky (1995), in that we are seeking to disambiguate both nouns in NCs rather than one element based on what the words it co-occurs with. We also apply it to the full WordNet sense inventory rather than coarse-grained binary distinctions. Hence, we do not expect as high accuracy as reported by Yarowsky. We also do not make any linguistic claims about the potential for a given NC to have different senses based on context. Our basic claim is that the majority of token occurrences of a given NC will conform to a given sense combination.

7.2.3 System Architecture to Disambiguate Word Sense

To disambiguate the word senses of constituents in NCs, we built two classifiers, one supervised and one unsupervised.

Supervised Method

The first classifier is supervised and uses the sense collocation method of Moldovan *et al.* (2004), modified to model the grammatical roles of target nouns as described above, in the form:

$$ws^*(n_i) = \text{argmax}_{ws(n_i)} P(ws(n_i)|ws(n_j), grammatical\_role) \tag{7.4}$$

Here, the semantics of n_j, which is assumed to be sense-determined, is extracted from two sources. We experiment with the use of two types of semantics: noun classes from CoreLex, and the first sense from WordNet. We also test the method under various conditions, as presented below. First, we use only the predetermined semantics of the non-target noun as a feature but underspecify the grammatical role; second, we use both the semantics of the non-target noun and the SR; and third, we use all of the semantics of the non-target noun, the SR and the grammatical role. The first two variants are intended to test the contribution of the grammatical role relative to the original method. The second and third variants are intended to check the importance of the SR in disambiguating the NC.

$$ws^*(n_i) = \text{argmax}_{ws(n_i)} P(ws(n_i)|ws(n_j)) \tag{7.5}$$

$$ws^*(n_i) = \text{argmax}_{ws(n_i)} P(ws(n_i)|ws(n_j), sr) \tag{7.6}$$

$$ws^*(n_i) = \text{argmax}_{ws(n_i)} P(ws(n_i)|ws(n_j), grammatical\_role, sr) \tag{7.7}$$

Unsupervised Method

We also built an unsupervised classifier based on lexical substitution, similarly to Mihalcea and Moldovan (1999) and Agirre and Martinez (2000). We replace the target

Sense	Substituted NCs
1	**craft/artifact** museum
2	**artistic production/creative activity** museum
3	**artistry/superior skill** museum
4	**artwork/graphics/visual communication** museum

Table 7.20: Example of the substitution method for each word sense of a polysemous NC element

noun with its *synonym*s from WordNet synsets, then compute the probability of each underlying word sense by calculating the frequency of the substituted NCs in a corpus. We used a web corpus (via the Google search engine) since it provides a large amount of data to compute the probabilities, despite noise (Lapata and Keller 2004). Note that as each word sense can have more than one *synonym*, we normalize the frequency across all *synonym*s to compute the final probability. Finally, we assign the word sense which has highest substitution-based frequency to the target noun.

Equation 7.8 shows how to compute the probability when the target noun (n_1) is the modifier and the non-target noun (n_2) is the head noun:

$$ws^*(n_1) = \text{argmax}_{s_i \in ws(n_1)} \frac{\sum_{n_j \in ss(s_i) \setminus \{s_i\}} \mathit{freq}(n_j, n_2)}{|ss(s_i) \setminus \{s_i\}|} \tag{7.8}$$

where each s_i is a word sense of n_1, and $ss(s_i)$ returns the synset containing sense s_i. The calculation in the case that the target noun is a head noun is analogous, with the only change being in the calculation of the corpus frequency.

7.2.4 Data Collection for Disambiguating Word Sense

To evaluate our method, we initially collected the top-20 frequent polysemous nouns from SemCor and Senseval 2. Then we identified binary NCs (i.e. noun-noun sequences) in the British National Corpus which contained each of 20 randomly-selected nouns in either the modifier or head noun position (but not both). From this, we extracted polysemous nouns which occurred as both modifier and head noun over

noun	# of sense	noun	# of sense	noun	# of sense
art	4	authority	7	bar	14
channel	8	child	4	circuit	6
day	10	nature	5	stress	4

Table 7.21: Target noun set, and the polysemy of each

at least 50 NC token instances. Finally, we selected 9 nouns which occurred in the most NCs, as described in Table 7.21.

As the final dataset, we randomly selected 50 NCs for each of the modifier and head noun positions of the 9 polysemous nouns. Hence, we have 100 NCs for each polysemous noun, totalling 900 instances.

To annotate the word senses of the target nouns in the 900 NCs, we hired two linguistically-trained human annotators. We extracted 50 sentences for each modifier or head noun-positioned target noun from the British National Corpus and provided them to the human annotators. These sentences were intended to help the annotators determine the word senses of target nouns in the case that a given NC was used over a range of sense assignments. Also, the set of sentences was used to take the majority class assignment for the NC type. The initial type-level inter-annotator agreement was 69.2%, and the human annotators met to discuss all instances of disagreement. A single expert annotator also annotated the 900 NCs for SR, once again with reference to their token occurrences in the British National Corpus. In post-analysis of the annotation, we observed that NC monosemy, i.e. a given noun occurring with only one sense in NCs, helps significantly when sense annotating NCs. The other fact is that although one sense tends to dominant in NCs, determining the majority sense was harder than we expected in some cases where the senses were relatively evenly distributed.

We specified the semantics of each non-target noun by either: (a) CoreLex, or (b) the first-sense and three direct *hypernyms* (similarly to Section 8.1 from WordNet 2.1). In the set of 900 NCs, 61.6% of the collocating nouns were contained in

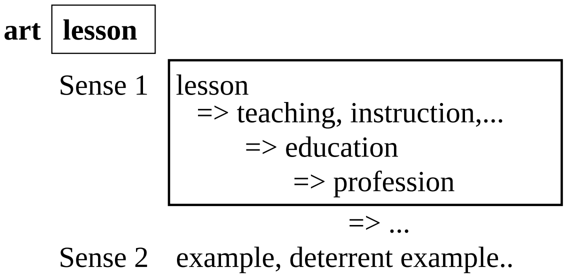

Figure 7.9: Word sense of *art* in different sense collocations

CoreLex. For the remainder, we manually assigned a CoreLex class following the CoreLex clustering methodology. For example, when *lemon* is not found in CoreLex, we assign the class food, which also contains *orange*. The determination of first sense in WordNet was based on the first sense learning method of McCarthy *et al.* (2004). Figure 7.9 shows the specification of the semantics of the non-target noun using three direct *hypernym*s. In this case, *art* is the target noun and *lesson* is the non-target noun *art lesson*.

To compute the probability for the unsupervised classifier, we calculated the web count of each synonymy-substituted NC using Google. Lapata and Keller (2004) showed that the web provides reliable probability estimates for tasks including unsupervised noun compound interpretation and bracketing. For our purpose, we generated both the singular and plural forms of each NC using morph (Minnen *et al.* 2001) to calculate the frequency of a given NC. Note that we do not include the target noun itself. If no *synonym*(s) of the target noun are found, then we use *hypernym*s (excluding substitution candidates which have lexical overlap with the target noun). For example, the synset membership of sense$_1$ of *art* is *art* and *fine art*. As *fine art* includes the word *art*, we look to the *hypernym*s, and end up with the candidates *artifact* and *craft*.

7.2.5 Experiment (I): Word Sense Disambiguation

The first experiment attempts to disambiguate the word sense of the target noun in a given NC based on each of the proposed methods. We evaluate the supervised WSD method via 10-fold cross-validation over the 900 NC instances using TiMBL5.1, and we evaluate the unsupervised method over the same 900 instances. We use two unsupervised baselines: (1) random sense assignment, and (2) the first sense prediction for the target noun by the method of McCarthy *et al.* (2004), based on the full British National Corpus (comprising both NC and non-NC instances of a given target noun). We also have one supervised baseline in the form of a majority class classifier, based on 10-fold cross-validation over the 900 instances. In order to benchmark our results, we ran SenseLearner over the dataset using the pre-trained word class models, randomly selecting one of the original sentential contexts from the British National Corpus for each NC and corresponding sense labeling. The classification accuracy for the output of each WSD method over each target noun, broken down across the modifier and head noun positions, is shown in Table 7.24 (see page 182). Results of the modified supervised methods (Equations 7.5–7.7) are presented in Table 7.22. The best-performing method is indicated in boldface in each row.

The majority-class baseline was the best-performing classifier for around half of the sub-experiments in Table 7.24, but was often only slightly better than the CoreLex and WordNet-based classifiers. There were a number of instances of the majority-class baseline falling well behind the other two supervised classifiers. According to our analysis, the lower performance is due to a lack of training data (50 instances for each target word). There was no significant difference in the performance for modifier vs. head noun WSD, but on further experimentation we were able to verify that conditioning the disambiguation on the syntactic role improved accuracy.

We found that SRs are helpful in disambiguating some of the target nouns when we look at the performance with and without SRs. However, the overall performance of the above approach is slightly less than that without SRs. We found that since the SRs we used contained errors (note that the NC interpretation method we used performed at around 55% accuracy), the noise introduced by such SRs reduces the performance of the WSD method. We expect that as the accuracy of automatic NC interpretation improves, integrating SRs into WSD will lead to higher WSD accuracy.

Comparing the performance of the different approaches using CoreLex and WordNet, the classifier that used CoreLex as the source of semantics for the non-target noun produced the best overall performance (accuracy=55%), marginally better than the performance of the same classifier using WordNet semantic features (accuracy=54%). However, considering that we manually assigned the classes of nouns which are not found in CoreLex, we acknowledge the WordNet-based method is both more general and almost as accurate.

The accuracy of SenseLearner was well below the baselines. This came as a surprise, since SenseLearner was trained over both the Senseval 2 and SemCor data. The performance of SenseLearner indicates that general-purpose WSD methods do not perform well over MWEs (in our case, NCs). This is a strong indication that the combination of sense collocation and the one-sense-per-collocation heuristic is a stronger indicator of noun sense in NCs than standard contextual features.

The unsupervised method's performance was well below that of the supervised methods (both the majority class baseline and the WordNet and CoreLex classifiers), and slightly below that of the first sense baseline with the same combined accuracy as SenseLearner.

The results of further experiments using the different variants over the basic supervised method are presented in Table 7.22.

Based on Tables 7.24 and 7.22, the grammatical role of the target noun plays an important role in disambiguating the word sense of the target noun. With regards to the importance of the SR in disambiguating NCs, we found that even though the SR predictions are noisy, they help to improve the performance of WSD, particularly when the role of the target noun is unknown (0.63 vs. 0.65 with CoreLex).

Target	Baseline			Supervised			
noun	R	F	M	P(C$_m$)	P(W$_m$)	P(C$_m$,SR)	P(W$_m$,SR)
art	.25	**.96**	**.96**	**.96**	.92	**.96**	.69
authority	.14	.14	**.42**	.28	.30	.36	.31
bar	.07	.74	.74	.74	**.76**	.67	.54
channel	.13	**.40**	**.40**	.34	.32	.38	.28
child	.25	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	.76
circuit	.17	.96	.96	.96	.94	**.98**	.66
day	.10	.24	.42	.38	.40	**.52**	.48
nature	.20	.38	**.46**	.32	.40	.38	.39
stress	.20	.10	.68	.68	**.70**	.67	.58
total	.16	.55	**.67**	.63	.64	.65	.52

Table 7.22: WSD accuracy over each target noun in the modifier and head noun positions (best-performing method in each row is indicated in **boldface**; *R* = random baseline; *F* = first sense baseline; *M* = majority class baseline; *C* = supervised classifier using CoreLex; *W* = supervised classifier using WordNet)

7.2.6 Experiment (II): NC interpretation using WSD

In the second experiment, we carried out NC interpretation with and without WSD, to evaluate the utility of WSD in interpreting NCs. Prior research suggests that noun sense information in NCs provides critical information when interpreting NCs, but has tended to be based on gold-standard sense information while our word senses are automatically generated and hence noisy.

We used the NC interpretation method of Moldovan *et al.* (2004) since it uses sense collocation. However, since we do not have sense information for the non-target noun, we used the semantic features from CoreLex and WordNet as our semantic representation. This is combined with the WordNet sense of the target noun to form the input to our supervised classifier. Finally, we performed 10-fold cross validation over the 900-instance dataset. We compared this directly to a first-sense disambiguation method (McCarthy *et al.* 2004), trained over the full British National Corpus (the same as used for the first-sense baseline in WSD). The output of the first sense classifier is combined with the CoreLex and WordNet features

Method	CoreLex	WordNet
baseline	27.3%	27.3%
similarity	34.6	34.6%
system-tagged	40.2%	**42.6%**
first-sense	**40.3%**	42.5%
hand-tagged	44.7%	54.0%

Table 7.23: Accuracy of interpreting SRs in NCs

of the collocating noun as above, producing a fully comparable classifier. We also used the approach presented in Section 5.1 as a benchmark for NC interpretation.

During the annotation of the SRs in type-level tagging, the initial type-level inter-annotator agreement was 52.31%, similarly to the agreement for our other dataset. The baseline used for the experiment was a majority-class classifier (the majority class being TOPIC).

Table 7.23 shows the performance of NC interpretation using the WordNet-based supervised method, the first sense disambiguation method of McCarthy *et al.* (2004), and hand-tagged sense data. These are compared to the results of our majority-class baseline and our method from Section 5.1. As stated before, the semantic features of collocating nouns are from CoreLex and WordNet (including the three direct *hypernym*s).

The results clearly show that NC interpretation with WSD outperforms that without WSD (54.0% vs. 42.6%). Comparing the performance of the system-tagged results, there is no significant difference between the supervised method and the first sense method. The performance of the WordNet-based supervised classifier and the first sense disambiguation method is almost identical using both the WordNet and CoreLex semantic representations. The upper-bound classifier based on hand-tagged data is predictably better than both of the automatic tagging methods, particularly for the CoreLex representation. This suggests that the features of the collocating nouns are more important than the noisy word sense features of the target noun. Additionally, in order to achieve higher accuracy (as we try to reach the upper bound accuracy), significantly higher WSD performance is required. Both automatic WSD-based methods clearly outperform both the baseline and the benchmark interpretation method. This demonstrates that word sense can indeed boost the performance of NC interpretation. Note that since all of the NCs in our dataset contain polysemous nouns, the performance of the our SR interpretation method is considerably lower than that reported in Section 5.1.

Looking at the results, we can clearly see that WSD does influence NC interpretation. We consider this a significant first step in opening up a new research direction in the field of NC interpretation.

7.2.7 Summary of Word Sense Disambiguation over NCs

We investigated a novel method to disambiguate NCs. Our motivation is that sense-disambiguated elements in NCs are a better predictor of the word sense of polysemous nouns than the context of usage of the NC. Also, we observed that the distribution of word senses can differ greatly across different grammatical roles in NCs.

In our proposed method, we combined the one sense per collocation heuristic with existing approaches from WSD (lexical substitution) and NC interpretation (sense collocation). As features for the methods, we used the word sense of the non-target noun and the role of target noun. We also experimented with an unsupervised method which replaces a constituent noun with similar words to generate new NCs, and collects web counts for each sense.

We evaluated the various methods we proposed and compared them to a number of baselines and a benchmark WSD system. We found that the grammatical role of the target noun helps to disambiguate the word sense of the noun, but that SRs were ineffectual in disambiguating the word sense of the target noun in an NC, due to noise in the SR interpretation data. However, we strongly believe that such an approach coupled with better-quality SR data will provide better results. It is also worthwhile to note that we used only 50 instances for evaluating both the supervised and unsupervised method. We expect better performance on larger datasets. We also found that a benchmark WSD system (for simplex words), namely SenseLearner, significantly underperformed our proposed methods. This is a significant finding that will motivate further work in tuning WSD methods. We also found that WSD helps significantly in interpreting NCs.

7.3 Chapter Summary

In this chapter, we presented two computational tasks based on **substitutability**. One interprets NCs based on sense collocation and bootstrapping, and the other disambiguates word senses of components in NCs based on the one sense per collocation heuristic as well as sense collocation. These methods were based on substitutability combined with sense collocation.

Substitutability can be used to model the degree of similarity between the original and new lexical item(s). In our approaches, we substituted the lexical item(s) in NCs and tested whether the semantic similarity between the original and substituted NCs is the same or similar.

Substitutability was also able to generate new NCs, based on the notion of sense collocation. In particular, sense collocation provides a means of calculating the semantic similarity between NCs, so that substitutability can be effectively used. The shortcoming of this approach is that it has limited applicability when a given sense collocation is not attested in the training data.

In order to disambiguate the word senses of constituent words in NCs, we once again adopted substitutability based on sense collocation. We used this method to build both unsupervised and supervised WSD methods for NCs.

Finally, we summarize the chapter as follows.

Summary of the constituent substitutability approach to interpreting noun compounds:

- We proposed an automatic interpretation method based on sense collocation;
- We attested that bootstrapping can generate large numbers of interpreted NCs;

- Our general performance achieved between 64.22% and 70.78% accuracy;
- We confirmed the utility of the acquired NCs for NC interpretation by using them as training data.

Summary of word sense disambiguation for NCs:

- We proposed a WSD method specifically targeting noun compounds, based on the one sense collocation heuristic and biased sense distribution on nouns in different grammatical roles in NCs;
- Our proposed methods achieved an accuracy of up to 55%;
- We employed heuristics (i.e. first sense and its hypernyms) in order to automatically model semantics of the non-target noun, and confirmed their reliability;
- We discovered that existing WSD methods do not reliably sense-tag MWEs by evaluating SenseLearner over NC data (accuracy = 30%);
- We confirmed that WSD provides key information for NC interpretation (system tagged = 42.6% vs. hand-tagged = 54.0%).

Target	Role	Baseline			Supervised				Unsupervised	SL
noun	in NC	R	F	M	C1	W1	C2	W2		
art	modifier	.25	.68	.68	.64	**.70**	.64	.65	.44	.54
	head noun	.25	**.54**	**.54**	.48	.51	.50	.56	.30	.50
	both	.25	**.61**	**.61**	.56	**.61**	.57	.61	.37	.52
authority	modifier	.14	.06	**.78**	.70	.77	.70	.75	.18	.06
	head noun	.14	.08	**.60**	.52	.54	.50	.50	.36	.08
	both	.14	.07	**.69**	.61	.65	.60	.62	.27	.07
bar	modifier	.07	.46	.46	**.54**	.47	.50	.49	.20	.46
	head noun	.07	.30	.24	**.46**	.40	.44	.43	.24	.28
	both	.07	.38	.35	**.50**	.43	.47	.46	.22	.37
channel	modifier	.13	**.24**	**.24**	**.24**	.18	.22	.20	.26	.22
	head noun	.13	.16	.26	**.28**	.24	.40	.32	.30	.12
	both	.13	.20	.25	**.26**	.21	.31	.26	.28	.17
child	modifier	.25	**.72**	**.72**	.50	.69	.74	.65	.24	.60
	head noun	.25	**.78**	**.78**	.76	.76	.76	.76	.38	.76
	both	.25	**.75**	**.75**	.63	.73	.75	.71	.31	.68
circuit	modifier	.17	**.68**	**.68**	.62	.61	.58	.61	.62	.66
	head noun	.17	.54	.54	.48	**.57**	.50	.57	.42	.52
	both	.17	**.61**	**.61**	.55	.59	.54	.59	.52	.59
day	modifier	.10	.18	**.68**	.64	.62	.62	.60	.24	.14
	head noun	.10	.06	**.90**	.88	.89	.88	.89	.16	.06
	both	.10	.12	**.79**	.76	.75	.75	.75	.20	.10
nature	modifier	.20	.04	**.70**	**.70**	**.70**	.58	.65	.30	.04
	head noun	.20	.34	.14	**.44**	.38	.40	.22	.20	.32
	both	.20	.19	.42	**.57**	.54	.49	.44	.25	.18
stress	modifier	.20	.02	.48	**.50**	.46	.40	.49	.30	.02
	head noun	.20	.08	.08	.24	**.27**	.26	.28	.28	.08
	both	.20	.05	.28	**.37**	.36	.33	.39	.29	.05
Total	modifier	.16	.34	**.60**	.59	.58	.55	.56	.31	.30
	head noun	.16	.32	.45	**.50**	**.50**	.52	.50	.29	.30
	both	.16	.33	.53	**.55**	.54	.53	.53	.30	.30

Table 7.24: WSD accuracy over each target noun in the modifier and head noun positions (the best-performing method in each row is indicated in **boldface**; *R* = random baseline; *F* = first sense baseline; *M* = majority-class baseline; *C1* = supervised classifier with CoreLex and grammatical role; *C1* = supervised classifier with CoreLex, grammatical role and SR; *W1* = supervised classifier with WordNet and grammatical role; *W2* = supervised classifier with WordNet, grammatical role and SR; *SL* = SenseLearner)

Chapter 8

MWEs and Linguistic Properties

Linguistic properties can provide strong evidence when extracting and identifying MWEs (Baldwin 2005a). We define linguistic properties to be localized information capturing syntactic and semantic properties of lexical items, represented either directly in the form of linguistic classes or via linguistic structures they are part of. It contrasts with distributional similarity, e.g. in that distributional similarity uses the neighboring words around the target lexical item as a proxy for its linguistic behaviour, whereas linguistic properties are a more direct representation of the word's linguistic properties. Examples of linguistic properties are the grammatical role of the word or its semantic class.

In this chapter, we introduce a novel method for identifying verb-particle constructions (*VPCs*) (e.g. *put on, battle on*) using linguistic properties. Our intuition is that word sense information and the grammatical roles of neighboring nouns can distinguish VPCs from simple verb–PP combinations (i.e. *Verb-PPs*) (e.g. *put* and preposition phrase, *on*). In this, we use the RASP parser to capture the argument-taking properties of a given noun and WordNet to model the ontological semantics of subject and object head nouns.

8.1 Using Linguistic Properties to Identify Verb-Particle Constructions

The key intuition underlying our proposed method is that in contexts where there is syntactic ambiguity for a given verb–preposition combination, it is possible to resolve the ambiguity via the selectional preferences of the verb vs. the VPC. For example, in the sentence *Kim ran in the room*, the object of the VPC *run in* will tend to be MACHINERY whereas the object of *in* as an adjunct of the simple verb *run* will tend to be of type PLACE. In our example sentence, *room* is semantically incompatible with the VPC analysis semantics, suggesting a verb-PP analysis. In contexts where there is a strong lexico-syntactic preference for a VPC analysis (e.g. *look it up*) or verb-PP analysis (e.g. *put it on the table*), on the other hand, syntactic parsers which are attuned to verb subcategorization and preposition valence are highly adept at predicting the correct analysis. In this light, our method takes the form of post-processing over the output of a probabilistic parser with a symbolic backbone, and attempts to identify and correctly disambiguate instances of syntactic ambiguity based on selectional preferences. The main contribution of this work is to demonstrate the utility of syntactic and semantic features for VPC identification.

As described in Section 1.2, VPC **identification** is the task of detecting individual VPC **token** instances in corpus data (Li *et al.* 2003). This contrasts with the more widely-researched task of VPC **extraction**, where the objective is to arrive at an inventory of VPC **types**/lexical items based on analysis of token instances in corpus data (Baldwin and Villavicencio 2002; Baldwin 2005a).

For the purpose of this work, we follow Baldwin (2005a) in adopting the simplifying assumption that VPCs: (a) consist of a head verb and a unique prepositional particle (e.g. *hand in*, *walk off*); and (b) are either transitive (e.g. *hand (the report) in, put on (a jumper)*) or intransitive (e.g. *battle on*). Here, we briefly recap the properties of VPCs from Section 2.3 that are relevant to the identification task. A defining characteristic of transitive VPCs is that they can generally occur with either joined (e.g. *He put on the sweater*) or split (e.g. *He put the sweater on*) word order. In the case that the object is pronominal, however, the VPC must occur in split word order (c.f. **He handed in it*) (Huddleston and Pullum 2002; Villavicencio 2003b). The semantics of the VPC can either derive transparently from the semantics of the head verb and particle (e.g. *walk off*) or be significantly removed from the semantics of the head verb and/or particle (e.g. *look up*); analogously, the selectional preferences of VPCs can mirror those of their head verbs or alternatively diverge markedly. The syntax of the VPC can also coincide with that of the head verb (e.g. *walk off*) or alternatively diverge (e.g. *lift off*).

The basic intuition behind the proposed method is that the selectional preferences of VPCs over predefined argument positions[1] provide insight into whether a verb and preposition in a given sentential context combine to form a VPC (e.g. *Kim handed in the paper*) or alternatively constitute a verb-PP (e.g. *Kim walked in the room*). That is, we seek to identify individual preposition token instances as intransitive prepositions (i.e. prepositional particles) or transitive prepositions based on analysis of the governing verb. We also analyze the application of selectional preferences over VPCs of differing levels of compositionality. Finally, we evaluate the performance of various NLP resources at VPC identification, to benchmark our method and suggest guidelines for the usage of these resources with VPCs.

8.2 Past Research on Syntactic Disambiguation of Verb-Particle Constructions

In this section, we survey relevant past research on VPCs, focusing on the extraction/identification of VPCs and the prediction of the compositionality/productivity of VPCs.

For VPC extraction and identification, Baldwin and Villavicencio (2002) proposed a method for extracting VPCs using a POS tagger, chunk parser, full syntactic parser and a combination of all three. The output of the method is a simple list of VPCs, which Baldwin (2005a) extended to propose a method for extracting VPCs with va-

[1] Focusing exclusively on the subject and object argument positions.

lence information for direct application in a grammar. Baldwin (2005a) followed Villavicencio (2003a) in assuming that VPCs: (a) have a unique prepositional particle, and (b) are either simple transitive or intransitive. Baldwin (2005a) achieved an extraction F-score of 74.9% and 89.7% for intransitive and transitive VPCs, respectively, over the British National Corpus.

Li *et al.* (2003) performed VPC identification based on hand-crafted regular expressions over the context of occurrence of verb–preposition pairs from TREC data, and reported a performance between 95.8% and 97.5%. Although these results are impressive, the adaptability of the method to new domains and languages is questionable due to employing several systems such as Named Entity recognizers and shallow and deep parsers, and the method is not directly applicable to other types of MWEs such as light verb constructions (Grefenstette and Teufel 1995; Stevenson *et al.* 2004) or determinerless PPs (Baldwin *et al.* 2006; Van Der Beek 2005).

Fraser (1976) and Villavicencio (2003b) argued that the semantic properties of verbs can determine the likelihood of their occurrence with different particles. Bannard *et al.* (2003), McCarthy *et al.* (2003) and Kim and Baldwin (2007a) proposed methods for estimating the compositionality of VPCs based largely on distributional similarity and the semantic similarity of the head verb and VPC (see Chapter 3 for details). O'Hara and Wiebe (2003) proposed a method for disambiguating the semantics of prepositions in verb-PPs. Cook and Stevenson (2006) classified the semantics of particles in VPCs using linguistic features. Katz and Giesbrecht (2006) built on the research of Baldwin *et al.* (2003a) in identifying token instances of non-compositional MWEs (particularly verb–noun idioms) in German using Latent Semantic Analysis, and further attempted to measure the compositionality of MWEs. While our interest is in VPC identification—a fundamentally syntactic task—we draw on the style of shallow semantic processing employed in these methods in modeling the semantics of VPCs relative to their base verbs.

8.3 Selectional preferences

Divergences in VPC and simplex verb semantics are often reflected in differing selectional preferences, as manifested in patterns of noun co-occurrence. That is, when verbs co-occur with particles to form VPCs, their meaning can be significantly different from the semantics of the head verb in isolation. In one example cited by Baldwin *et al.* (2003a), the cosine similarity between *cut* and *cut out*, based on word co-occurrence vectors, was found to be greater than that between *cut* and *cut off*, mirroring the intuitive compositionality of these VPCs.

(8.1) and (8.2) illustrate the difference in the selectional preferences of the verb *put* in isolation as compared with the VPC *put on*.[2]

(8.1) *put* = place

EX: $\underline{\textit{Put}}\ \textit{the}\ \boxed{\textit{book}}\ \underline{\textit{on}}\ \textit{the table.}$

ARGS: *book*$_{\text{OBJ}}$ = book, publication, object

ANALYSIS: verb-PP

(8.2) *put on* = wear

EX: $\underline{\textit{Put}}\ \underline{\textit{on}}\ \textit{the}\ \boxed{\textit{coat}}.$

ARGS: *coat*$_{\text{OBJ}}$ = garment, clothing

ANALYSIS: VPC

Put on is generally used in the context of "wearing" something, with object nouns such as *sweater* and *coat*, whereas *put* in isolation has less sharply defined selectional restrictions and can occur with any noun. In terms of the word senses of the head nouns of the object NPs, the VPC *put on* tends to co-occur with objects which have the semantics of CLOTHING. On the other hand, the simplex verb *put* in isolation tends to be used with a broader range of both concrete and abstract objects, and prepositional phrases containing NPs with the semantics of PLACE.

Also, as observed above, the valence of a VPC can differ from that of the head verb depending on the word sense in the context. (8.3) and (8.4) illustrate two different

[2] All sense definitions are derived from WORDNET 2.1.

senses of *take off* with intransitive and transitive valence, respectively. Note that *take* cannot occur as a simplex intransitive verb.

(8.3) *take off* = lift off

EX: *The* $\boxed{\textit{airplane}}$ $\underline{\textit{takes}}$ $\underline{\textit{off}}$.

ARGS: *airplane*$_{\text{SUBJ}}$ = airplane, aeroplane

ANALYSIS: VPC

(8.4) *take off* = remove

EX: $\boxed{\textit{They}}$ $\underline{\textit{take}}$ $\underline{\textit{off}}$ *the* $\boxed{\textit{cape}}$.

ARGS: *they*$_{\text{SUBJ}}$ = person, individual

cape$_{\text{OBJ}}$ = garment, clothing

ANALYSIS: VPC

In (8.3), the intransitive *take off* co-occurs with a subject of semantic class AEROPLANE. In (8.4), on the other hand, the transitive *take off* has an object noun of class CLOTHING. From the above, we can observe that head nouns in the subject and object argument positions can be used to distinguish VPCs from simplex verbs with prepositional phrases (i.e. verb-PPs).

8.4 Approach and Architecture for Identifying VPCs

The distinguishing features of our approach are: (i) it tackles the task of VPC identification rather than VPC extraction, and (ii) it uses both syntactic and semantic features, employing the WordNet 2.1 senses of the subject and/or object(s) of the verb. In the sentence *He* $\underline{\textit{put}}$ *the coat* $\underline{\textit{on}}$ *the table*, e.g. to distinguish the VPC *put on* from the verb *put* occurring with the prepositional phrase *on the table*, we identify the senses of the head nouns of the subject and object(s) of the verb *put* (i.e. *he* and *coat*, respectively). That is, VPCs are identified by looking at the semantics of the head nouns of the subject and/or object of a given verb (either VPC or verb in isolation).

Figure 8.1 depicts the complete process used to distinguish VPCs from verb-PPs.

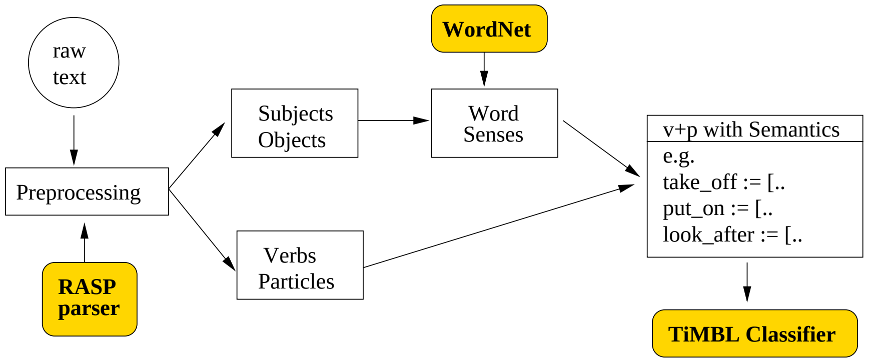

Figure 8.1: System architecture for VPC identification

First, we parse all sentences in a given corpus using the RASP parser (Briscoe and Carroll 2002), and identify verbs and prepositions in the RASP output. This is a simple process of checking the POS tags in the most-probable parse, and for both particles (tagged `RP`) and transitive prepositions (tagged `II`) reading off the governing verb from the dependency tuple output. We also retrieve the head nouns of the subject and object(s) of each verb directly from the dependency tuples. We then obtain the lexical semantics of the head nouns based on WordNet, using the first sense for that word in SemCor (see Section 8.5.2 and Figure 8.3). The final feature representation for each VPC and verb-PP takes the form of the verb lemma, preposition, and WordNet class of the subject and/or object(s). For the training instances only, we additionally generate separate instances for each of the first- to third-level hypernyms of the first sense.

Having extracted all the features, we then separate it into test and training data, and use TiMBL v5.1 (Daelemans *et al.* 2004) to learn a classifier.

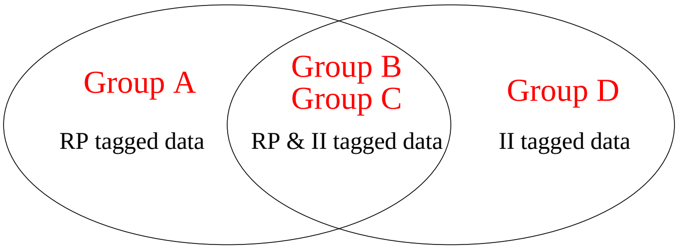

Figure 8.2: Data groups in terms of the POS tags from RASP

8.5 Data for VPC identification

8.5.1 Data Classification

Our evaluation data is made up of sentences containing prepositions tagged as either `RP` or `II`. Based on the output of RASP, the sentences are divided into four groups, as detailed in Figure 8.2.

Group A contains the verb–preposition token instances tagged exclusively as VPCs (i.e. the preposition is never tagged as `II` in combination with the given head verb). Group B contains the verb–preposition token instances identified as VPCs by RASP where there were also instances of that same combination identified as verb-PPs. Group C contains the verb–preposition token instances identified as verb-PPs by RASP where there were also instances of that same combination identified as VPCs. Finally, group D contains the verb-preposition combinations which were tagged exclusively as verb-PPs by RASP.

We focus particularly on disambiguating verb–preposition token instances falling into groups B and C, where RASP has identified an ambiguity for that particular combination. We do not further classify token instances in group D, on the grounds that: (a) for high-frequency verb–preposition combinations, RASP was unable to find a single instance warranting a VPC analysis, suggesting it had high confidence in its ability to correctly identify instances of this lexical type; and (b) for low-frequency verb–preposition combinations where the confidence of there definitively not being a

	False Positive Rate(FPR)	False Negative Rate(FNR)	Agreement
Group A	5.36%	–	94.64%
Group B	3.96%	–	99.61%
Group C	–	10.15%	93.27%
Group D	–	3.4%	99.20%

Table 8.1: False-positive rate (FPR), false-negative rate (FNR), and inter-annotator agreement across the four groups of token instances

VPC usage is low, the token sample is too small to disambiguate effectively and the overall impact would be negligible even if we tried. We do, however, return to consider data in group D in computing the precision and recall of RASP.

Naturally, the output of RASP is not error-free, i.e. VPCs may be parsed as verb-PPs and vice versa. In particular, other than the results of McCarthy *et al.* (2003) for identifying VPCs, we had no a priori sense of RASP's ability to distinguish VPCs and verb-PPs. Our only point of reference was the result of McCarthy *et al.* (2003) for RASP identifying VPCs vs. other analyses, at 0.926 precision and 0.642 recall, taking the POS tags for prepositions in the Wall Street Journal as the gold standard. While this is suggestive of the ability of RASP to identify VPCs, the evaluation is slightly problematic because: (a) the Penn POS tagset does not differentiate between particle and transitive prepositional usages of *to*, such that there is no data for this preposition; and (b) the tagset additionally attempts to differentiate between compositional (adverbial) and non-compositional particles (`RB` and `RP`, respectively), with limited success in our experience (see the definition of VPCs in Section 2.3). Therefore, we manually checked the false-positive and false-negative rates in all four groups (as defined relative to the gold-standard annotation in the Penn Treebank) and obtained the performance of the parser with respect to VPCs. The verb-PPs in group A and B are false-positives, while the VPCs in group C and D are false-negatives (we consider the VPCs to be positive examples).

To calculate the number of incorrect examples, two human annotators independently checked each verb–preposition instance. Table 8.1 details the rate of false-

	freq$_{\geq 1}$		*freq*$_{\geq 5}$	
	VPC	Verb-PP	VPC	Verb-PP
Group A	5,223	0	3,787	0
Group B	1,312	0	1,108	0
Group C	0	995	0	217
Total	6,535	995	4,895	217

Table 8.2: The number of VPC and Verb-PP token instances in groups A, B, and C at varying frequency cut-offs

Type	Groups A&B	Group C
common noun	7,116	1,239
personal pronoun	629	79
demonstrative pronoun	127	1
proper noun	156	18
who	94	6
which	32	0
No sense (*what*)	11	0

Table 8.3: The number of subject and object head nouns of different type

positives and false-negative examples in each data group, as well as the inter-annotator agreement (calculated over the entire group).

8.5.2 Data Collection

We combined together the $6,535$ (putative) VPCs and 995 (putative) verb-PPs from groups A, B and C, as identified by RASP over the corpus data. Table 8.2 shows the number of VPC tokens in groups A and B, and the number of verb-PPs in group C. *freq*$_{\geq 1}$ is the number of (VPC or V-PP) tokens which occur at least once, and *freq*$_{\geq 5}$ is the number of tokens which occur five or more times. Note that the number of (ambiguous) verb-PP tokens which occur repeatedly (in group C) is much less than that of VPCs (in groups A and B).

Initially, 8, 165 nouns were retrieved, including personal pronouns (e.g. *I, he, she*), demonstrative pronouns (e.g. *one, some, this, these*) and proper nouns (e.g. *CITI, Canada, Ford*), as the head nouns of a subject or object of a VPC or Verb-PP in the data set. Among the 8,165 nouns, we found that about 10% of the nouns were pronouns (P-PRN or D-PRN), proper nouns or WH words (*who*, *which* or *what*). We similarly retrieved 1, 343 nouns for verb-PPs in group C. Table 8.3 shows the distribution of different noun tokens across these two sets. In evaluation, we test three strategies for dealing with pronouns, proper nouns and WH words: (1) pronouns are manually resolved to the WordNet class of their antecedents and proper nouns are replaced by their hypernyms; (2) all pronouns and proper nouns are left unresolved; and (3) only proper nouns are replaced by their hypernyms. That is, in dealing with pronouns, we experimented with the option of manually resolving the antecedent and taking this as the head noun. When *which* is used as a relative pronoun, we identified if it was co-indexed with an argument position of a VPC or verb-PP, and if so, manually identified the antecedent, as illustrated in (8.5).

(8.5)

EX: *Tom likes the books which he sold off.*

ARGS: $\textit{he}_{\text{SUBJ}} = \textsf{person}$

$\textit{which}_{\text{OBJ}} = \textsf{book}$

With *what*, on the other hand, we were generally not able to identify an antecedent, in which case the argument position was left without a word sense (for detailed discussion, see Section 8.7.4).

(8.6) *Tom didn't look up what to do.*

(8.7) *What went on?*

We also optionally replaced all proper nouns with corresponding common noun *hypernyms* based on manual disambiguation since the coverage of proper nouns in WordNet is poor. Table 8.4 shows examples of manually replacing proper nouns with their corresponding common noun *hypernyms*.

Proper noun	Common noun hypernym
CITI	bank
Canada	country
Ford	company
Smith	human

Table 8.4: Example proper nouns and their common noun hypernyms

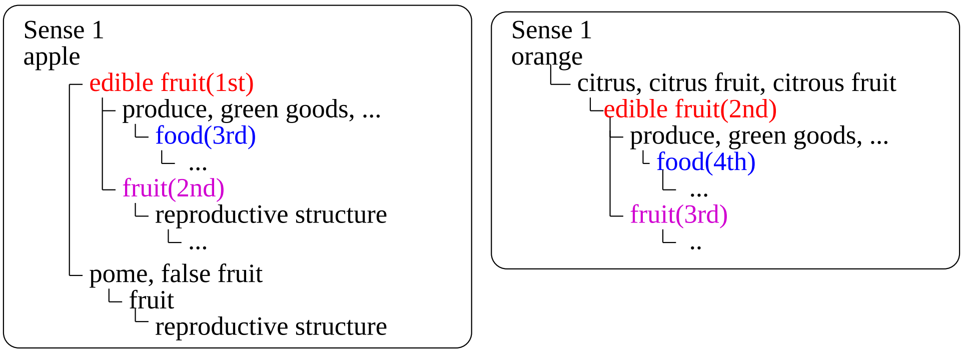

Figure 8.3: Senses of *apple* and *orange*

To estimate the word senses of the subject and object head nouns, we simply took the first sense and the associated *hypernyms* to three levels up the WordNet hierarchy.[3] This is intended as a crude form of smoothing for closely-related word senses which occur in the same basic region of the WordNet hierarchy, and enables the determination of suitable selectional preference classes in WordNet.

Figure 8.3 illustrates our method of sense disambiguation and representation, based on *apple* and *orange*. While the first (i.e. fruit) senses of *apple* and *orange* occur at different levels of the WordNet hierarchy, their semantic similarity is captured by the fact that their bag of *hypernyms* overlaps by two synsets, namely EDIBLE FRUIT and FRUIT.

Finally, we randomly selected 80% of the instances to use as training data and

[3] The choice of 3 levels was made empirically.

	Training Instances	
	Before expansion	After expansion
Group A	5,223	24,602
Group B	1,312	4,158
Group C	995	5,985

Table 8.5: Final number of training instances from the Brown Corpus and Wall Street Journal

Group	Frequency of VPCs	Size
BC	$\textit{freq}_{\geq 1}$ ($\textit{freq}_{\geq 1}$ & $\textit{freq}_{\geq 1}$)	test:498
	$\textit{freq}_{\geq 5}$ ($\textit{freq}_{\geq 5}$ & $\textit{freq}_{\geq 1}$)	train:1,809
BAC	$\textit{freq}_{\geq 1}$ ($\textit{freq}_{\geq 1}$ & $\textit{freq}_{\geq 1}$ & $\textit{freq}_{\geq 1}$)	test:1,598
	$\textit{freq}_{\geq 5}$ ($\textit{freq}_{\geq 5}$ & $\textit{freq}_{\geq 5}$ & $\textit{freq}_{\geq 1}$)	train:5,932

Table 8.6: Data set sizes at different frequency cut-offs

the remaining 20% as test data. The total number of training instances, before and after performing hypernym expansion using WordNet, is indicated in Table 8.5.

Table 8.6 describes the data fashioned from the combination of the Brown Corpus and Wall Street Journal. We collected the data from either groups B and C, or all of groups A, B and C in order to include both positive and negative instances. $\textit{freq}_{\geq N}$ means a given verb–preposition combination has been observed at least N times in the respective dataset. Note that the number of Verb-PP token instances in group C with cut-off $\textit{freq}_{\geq 5}$ is only slightly above 20%. As a result, instead of using instances with $\textit{freq}_{\geq 5}$, we chose to use instances with $\textit{freq}_{\geq 1}$.

8.6 Evaluation

Due to the differing amounts of data in A, B and C, we experimented with four different combinations of data from each, based on differing frequency thresholds over the training data. In the first two datasets, we include only instances from groups B and C (i.e. token instances of types with both VPC and V-PP instances), including all VPC instances from C (i.e. a frequency threshold of $\mathit{freq}_{\geq 1}$), and either all V-PP instances ($\mathit{freq}_{\geq 1}$) or only V-PP instances with a token frequency of 5 or greater ($\mathit{freq}_{\geq 5}$) from B. In the second two datasets, we additionally include unambiguous VPCs from group A to boost the number of positive training instances, either taking all VPC instances ($\mathit{freq}_{\geq 1}$) or only those instances with a token frequency of 5 or greater ($\mathit{freq}_{\geq 5}5$). The reason we always use all V-PP token instances ($\mathit{freq}_{\geq 1}$) from C is that the V-PPs tend to have a low frequency. Note that in all cases of VPC identification, we include all test instances, irrespective of frequency, such that the precision, recall and F-score under the different experimental settings are directly comparable.

We separately evaluated the three different strategies for resolving pronouns and proper nouns (full manual resolution, no manual resolution and manual resolution for proper nouns only).

As our baseline for VPC identification, we use the raw output of RASP. For identifying VPC and Verb-PPs, we use both the raw output of RASP, as well as zero-R (i.e. majority vote) as our baseline .

The precision and recall for VPC identification are computed as follows:

$$\mathit{Precision} = \frac{\text{Data Correctly Identified as VPC}}{\text{Data Identified as VPC}} \tag{8.8}$$

$$\mathit{Recall} = \frac{\text{Data Correctly Identified as VPC}}{\text{All VPCs in Data Set}} \tag{8.9}$$

8.6.1 Experiment (I): Fully Resolved Pronouns and Proper Nouns

Table 8.7 shows the results of our method over the Brown Corpus and Wall Street Journal using manually-resolved pronouns and proper nouns, in terms of VPC identification (i.e. identifying positive and negative VPCs). As mentioned above, we evaluate different combinations of data from A, B, and C, at different frequency thresholds. The performance of RASP at identifying VPCs is calculated based on human judge-

Data	Frequency	Precision	Recall	F-Score
RASP	$freq_{\geq 1}$	.959	.955	.957
	$freq_{\geq 5}$	.967	.962	.964
BC	$freq_{\geq 1}$	.948	.958	.952
	$freq_{\geq 5}$	.955	.979	.966
BAC	$freq_{\geq 1}$	.962	.962	.962
	$freq_{\geq 5}$	.965	.984	.974

Table 8.7: Results for VPC identification only

Data	Frequency	Type	Precision	Recall	F-Score
RASP	$freq_{\geq 1}$	PV	.933	–	–
	$freq_{\geq 5}$	PV	.941	–	–
zero-R(BC)	$freq_{\geq 1}$,$freq_{\geq 5}$	PV	.546	–	–
zero-R(BAC)	$freq_{\geq 1}$,$freq_{\geq 5}$	PV	.859	–	–
BC	$freq_{\geq 1}$	PV	.807	.803	.805
	$freq_{\geq 5}$	PV	.865	.853	.859
BAC	$freq_{\geq 1}$	PV	.866	.866	.866
	$freq_{\geq 5}$	PV	.927	.884	.9054

Table 8.8: Results for VPC and Verb-PP identification

ment over all token instances in groups B and C. When RASP identifies a verb and particle correctly, we consider it to have identified the VPC correctly irrespective of whether the argument structure is correct or not. Also, we ignore ambiguity between particles and adverbs (e.g. *hand out* vs. *walk out*), leading to higher performance than that reported by McCarthy *et al.* (2003).

Table 8.7 shows that the performance over high-frequency data from groups A, B and C is the highest (F-score = 0.974). As a general trend, the best results are achieved over the high-frequency VPCs, including data from A.

Table 8.8 shows the precision, recall and F-score of VPC and Verb-PP identification (i.e. positive and negative VPCs and positive and negative Verb-PPs). For both of VPC and Verb-PP identification, we have four different data groups with frequency based cut-offs. We calculated the precision of RASP in Table 8.8 by checking both correctly tagged VPCs and Verb-PPs. This precision measures how well RASP distinguishes between VPCs and Verb-PPs for ambiguous verb-preposition combinations in group B and C. Note that, apart from the performance of RASP, the baseline (i.e. *zero-R*) of identifying both VPCs and Verb-PPs is computed by majority class over instances in groups A, B and C.

Table 8.7 shows that the performance over high-frequency data identified from groups A, B and C is the highest (F-score = 0.974). Generally, the experiment with the data set containing high frequency and both positive and negative examples produced the highest performance. Encouragingly, we achieve a slightly higher result than the 0.958–0.975 claimed by Li *et al.* (2003) with relatively little manual intervention (to resolve the semantic class of each pronoun and proper noun). The approach of Li *et al.* (2003) is encumbered by the cost of generating hand-written rules for each individual verb–preposition combination. Our method has the benefit of achieving comparably higher performance with considerable less annotation cost.

In Table 8.8, our method achieved lower performance than RASP does due to ignoring the data from group D. However, our method still outperformed the majority baseline.

In Tables 8.7 and 8.8, we see that our method performs better at VPC identification than verb-PP identification. The fundamental reason is that we ignore the instances in group D, which means our method lacks instances to disambiguate verb-PPs. In terms of the token frequency of a given verb–preposition combination, our method predictably performs better at both VPC identification and verb-PP identification over high-frequency instances.

8.6.2 Experiment (II): No Resolved Pronouns or Proper Nouns

We next repeat the experiment using the same data set as above but without manual resolution of the antecedents of pronouns and proper nouns. Here, every pronoun and proper noun (and common noun not found in WordNet) is represented not as a synset but as a coarse-grained feature describing the noun type (common

Data	Frequency	Precision	Recall	F-Score
RASP	$\mathit{freq}_{\geq 1}$	.959	.955	.957
	$\mathit{freq}_{\geq 5}$	.967	.962	.964
BC	$\mathit{freq}_{\geq 1}$	.936	.958	.946
	$\mathit{freq}_{\geq 5}$	.940	.956	.948
BAC	$\mathit{freq}_{\geq 1}$	.949	.969	.959
	$\mathit{freq}_{\geq 5}$	.951	.966	.958

Table 8.9: Results for VPC and verb-PP identification without resolving pronouns and proper nouns

Data	Frequency	Type	Precision	Recall	F-Score
RASP	$\mathit{freq}_{\geq 1}$	PV	.933	–	–
	$\mathit{freq}_{\geq 5}$	PV	.941	–	–
zero-R(BC)	$\mathit{freq}_{\geq 1}$,$\mathit{freq}_{\geq 5}$	PV	.546	–	–
zero-R(BAC)	$\mathit{freq}_{\geq 1}$,$\mathit{freq}_{\geq 5}$	PV	.859	–	–
BC	$\mathit{freq}_{\geq 1}$	PV	.792	.787	.788
	$\mathit{freq}_{\geq 5}$	PV	.798	.794	.795
BAC	$\mathit{freq}_{\geq 1}$	PV	.881	.846	.862
	$\mathit{freq}_{\geq 5}$	PV	.874	.847	.859

Table 8.10: Results for VPC and verb-PP identification without resolving pronouns and proper nouns

noun, pronoun, or proper noun). Common nouns are automatically assigned WordNet synsets as before, whereas pronouns and proper nouns are sub-classified into the human and non-human classes. All of these features are automatically derived, and based on POS tags and dictionaries.

Our interest in this experiment is to determine the relative drop when we take away the rich ontological semantics we manually annotated in the first experiment.

Table 8.9 and 8.10 show the results without manually resolving pronouns and proper nouns. Due to the relative sparsity of semantic information, the performance of this method is below that of the manually-resolved nouns in our first experiment, but still slightly above the F-score of RASP (0.959 vs. 0.957) at VPC identification.

Data	Frequency	Precision	Recall	F-Score
RASP	$freq_{\geq 1}$	.959	.955	.957
	$freq_{\geq 5}$	.967	.962	.964
BC	$freq_{\geq 1}$	.938	.960	.948
	$freq_{\geq 5}$	.938	.957	.947
BAC	$freq_{\geq 1}$	.951	.967	.959
	$freq_{\geq 5}$	.951	.966	.958

Table 8.11: Results for VPC identification only when partially resolving pronouns and proper nouns using WordNet

8.6.3 Experiment (III): Partially Resolved Pronouns and Proper Nouns using WordNet

Our third experiment is identical to the previous two experiments except that proper nouns are (partially) resolved using WordNet, in that if a proper noun is found in WordNet it is resolved in an identical manner to common nouns, and if not we fall back to the human vs. non-human binary distinction from experiment 2. As such, this experiment still requires no manual effort to resolve the semantics of head nouns, but lacks semantics for pronouns and proper nouns which do not occur in WordNet.

Our expectation is that despite WordNet having poor coverage of proper nouns, we will still manage to retrieve word senses of many commonly-occurring proper nouns automatically. Note that around 28% of the proper noun token instances in our data were found in WordNet.

Tables 8.11 and 8.12 describe the performance of our method with partially-resolved semantics for proper nouns. The F-score is almost identical to that for unresolved semantics (experiment 2), suggesting that the primary gain in performance in experiment 1 was for pronouns rather than proper nouns.

Data	Frequency	Type	Precision	Recall	F-Score
RASP	$freq_{\geq 1}$	PV	.933	–	–
	$freq_{\geq 5}$	PV	.941	–	–
zero-R(BC)	$freq_{\geq 1}$,$freq_{\geq 5}$	PV	.546	–	–
zero-R(BAC)	$freq_{\geq 1}$,$freq_{\geq 5}$	PV	.859	–	–
BC	$freq_{\geq 1}$	PV	.800	.795	.795
	$freq_{\geq 5}$	PV	.797	.792	.793
BAC	$freq_{\geq 1}$	PV	.880	.851	.865
	$freq_{\geq 5}$	PV	.874	.849	.861

Table 8.12: Results for VPC and verb-PP identification when partially resolving pronouns and proper nouns using WordNet

Freq	Type	#	Precision	Recall	F-score
$freq_{\geq 1}$	V	4WS	.962	.962	.962
		1WS	.958	.969	.963
$freq_{\geq 1}$	P	4WS	.769	.769	.769
		1WS	.800	.743	.770
$freq_{\geq 5}$	V	4WS	.964	.983	.974
		1WS	.950	.973	.962
$freq_{\geq 5}$	P	4WS	.889	.783	.832
		1WS	.813	.614	.749

Table 8.13: Results with hypernym expansion (4WS) vs. only the first sense (1WS)

8.6.4 Experiment (IV): 4 Word Senses vs. 1 Word Sense

Finally, we compare the performance of: (a) the proposed method with manual sense resolution and hypernym expansion (4WS); with (b) manual sense resolution but without hypernym expansion (1WS). Note that for all experiments reported so far, we have used hypernym expansion, and as such, the numbers for hypernym expansion are identical to those from Table 8.7. The results presented in Table 8.13, suggest that using hypernyms improves performance over frequent verb–preposition combinations.

8.7 Discussion

8.7.1 Performance Analysis

We proposed an automatic method for identifying English VPCs based on the selectional preferences of different argument positions. We experimented with three different strategies for resolving the semantics of pronouns and proper nouns, and found that while an oracle co-reference resolution and proper noun interpretation system improved performance slightly, the relative increment over a fully-automated method with partial coverage is slight. Overall, our method exceeded the performance reported in Li *et al.* (2003) and the RASP baseline.

In Table 8.7, we can see that the performance over high-frequency data identified from groups B, A and C is the highest (F-score = 0.974). In general, we would expect the data set containing the high frequency positive and negative examples to give us the best performance at VPC identification.

Combining the results for Tables 8.7 and 8.9, we see that our method performs better at VPC identification than verb-PP identification. Since we do not take into account the data from group D with our method, the performance at verb-PP identification is low compared to that for RASP, which in turn leads to a decrement in the overall performance.

Among the experiments over different options of resolving pronouns and proper nouns, we found that our method achieved the best performance given manually-resolved pronouns and proper nouns. That is, detailed semantic information about the subject and object head nouns helps to distinguish VPCs from verb-PPs, although the relative increment in identification performance is perhaps incommensurate with the effort required to provide this extra information. This result begs the question of how well the method would perform given automatically resolved pronouns and automatically interpreted proper nouns, which we leave for future research.

Note that our word sense disambiguation is based on simple first sense information, and doesn't rely on a word sense disambiguation system or hand tagging. The method proved superior to simple lexical probabilities (as are used by RASP), and gained

Parser	Precision	Recall	F-Score
RASP	.959	.955	.957
fnTBL	.703	.632	.668
Charniak parser	.659	.694	.676
Minipar	.364	.429	.397

Table 8.14: Performance of different parsers at VPC identification

from semantic smoothing via three levels of hypernyms.

8.7.2 Performance of Resources

Our method takes the form of a post-processing step after parsing, with all experiments based on RASP. Clearly the performance of the post-processing is predicated on the quality of the parser output, as we rely on the parser to identify the argument structure and head nouns. To evaluate the relative performance of RASP at VPC identification relative to other existing parsers, we evaluated a full text chunk parser based on fnTBL (Ngai and Florian 2001), Charniak parser (Charniak 2000) and Minipar (Lin 1993) over the same task. Note that we did not retrain any of the parsers, just as we did not retrain RASP in our original experiments. Table 8.14 shows the performance of the different parsers at VPC identification.

Table 8.14 shows the VPC identification performance of the three parsers, relative to the performance for RASP. RASP outperformed all three parsers, suggesting that it was a well-chosen parser for the task at hand. Note that as we did not retrain the parsers or do any post-processing, the performance of Minipar is much less than that reported in McCarthy *et al.* (2003).

8.7.3 Reflections on Compositionality

To investigate the correlation between the compositionality of each VPC and our ability to identify token instances of it, we took 117 VPCs of varying semantic compositionality that occurred in the data set of McCarthy *et al.* (2003). Recall that McCarthy *et al.* (2003) provide compositionality judgements for VPC types from three human judges on a scale of 0 to 10 (0 = non-compositional, 10 = fully compositional).

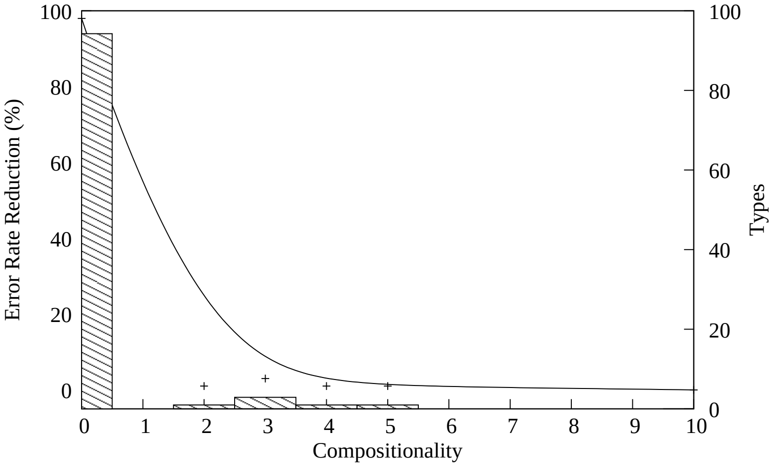

Figure 8.4: Error rate reduction for VPCs of varying compositionality

Figure 8.4 is a plot of the F-score for both RASP and our method at different levels of compositionality, for those VPCs in our data set which also occur in the data of McCarthy *et al.* (2003). The goal to compare the compositionality with McCarthy *et al.* (2003) is to see how our proposed method performed with human scores. The points are the actual number of VPCs with respect to the compositionality scores and bars and line are projection of the scores based on the actual VPC numbers. From the graph, we see that we our method actually degrades the VPC identification performance of RASP over low-compositionality VPCs, but greatly increases performance over high-compositionality VPCs, with the combined effect being a modest increase in F-score. The reason for this is that low-compositionality VPCs (e.g. *drag on*) are often easy for parsers to identify, as their subcategorization properties diverge from the simplex verb or there is no corresponding simplex verb at all (c.f. *chicken out*). RASP performs predictably well over these VPCs. High-compositionality VPCs (e.g. *call in*), on the other hand, tend to be less easy to distinguish from V-PPs based on syntax alone, and the semantic modeling underlying our method comes to the fore. Given a reliable method for predicting the compositionality of a given verb–preposition combination, we could consider evoking our method only for high-compositionality VPCs, and more effectively hybridise our method with the raw RASP outputs. Current research on compositionality prediction, however, is far from reliable (McCarthy *et al.* 2003; Baldwin *et al.* 2003a), making this an unrealistic expectation at present.

8.7.4 Factors for Further Study

Clearly there are more possibilities for exploiting semantic features than what we have explored in this work. As future research, we are particularly interested in including distributional similarity and semantic features for other argument types

From our method, we found several factors that require further study. In manually analyzing the data, some data instances were missing head nouns, leading to nouns without word senses. If only a small number of token instances are available, missing word senses can influence the performance of the method since the classifier relies on training data to disambiguate VPCs against verb-PPs. Particular instances of missing nouns are imperative and abbreviated sentences such as the following:

(8.10) *Come in.*

(8.11) *(How is your cold?) Broiled out.*

Another factor is the lack of word sense data, particularly in WH questions. We cannot simply replace it with an antecedent common noun to retrieve the noun semantics:

(8.12) *What do I hand in?*

(8.13) *You can add up anything.*

Also, the method is clearly dependent on the base performance of RASP, and any improvement in the base parser has the potential to improve our method (or even make our approach redundant!). We observed that among the false positive VPCs, there were occurrences of the particle occurring before the verb. An example of this is the following sentence:

(8.14) Help me up, I feel kind of stiff.

from which RASP identified the VPC *feel up*. Linguistically speaking, the particle must always appear after the verb (except with non-selected adverbial uses of prepositions such as *up he got*), a constraint which could be built in to RASP.

In other cases, the particle was attached to the wrong verb, e.g. in the following sentence:

(8.15) Lucy drew out the chair and sat down.

from which RASP identified the VPC *draw down*. Once again here, a constraint on the degree of separation between the verb and its particle (similarly to (Baldwin 2005a)) could prevent such misanalyses.

8.8 Chapter Summary

In this chapter, we utilized the linguistic properties of VPCs—in the form of the selectional preferences of different arguments for a given verb/VPC—in order to identify them in context. We proposed a method to identify VPCs automatically from raw text data based on these linguistic properties. We first used RASP to identify verb-preposition token instances as possible VPCs or verb-PPs. Then, we extracted the argument structure for each verb and derived the word senses of the subject and/or object head nouns. Finally, we built a supervised classifier using TiMBL v5.1 to relabel false positive VPCs as verb-PPs and vice versa. Over a small data set extracted from the Brown Corpus and Wall Street Journal, our classifier achieved an F-score of 0.974 for the task of VPC identification. We also tested the proposed method over various representations of noun semantics, and showed that automatic methods can approach the performance of methods which assume full coreference resolution and proper noun interpretation. Finally, we demonstrated a direct correlation between the degree of compositionality and the ability of our method to correctly identify VPCs.

The main advantage of our method is that it is fully automated and makes active use of existing resources. We suggest that our proposed approach is a reliable, stable method for automatic VPC identification.

A summary of the findings of the chapter is:

- We proposed a VPC identification method based on selectional preferences.
- The performance of the method (F-score of 97.4%) was shown to:
    - exceed the baseline RASP performance;
    - exceed previously-published results for VPC identification;
    - benefit from hypernym expansion (4WS vs. 1WS);
    - correlate (somewhat) with the relative compositionality of the VPC.

Chapter 9

Conclusions

9.1 Summary and Findings of this Thesis

This thesis has been an attempt to model the syntax and semantics of English MWEs based on the following statistical approaches: semantic similarity, substitutability, ellipsed predicates, co-occurrence properties and linguistic properties. A detailed discussion of the experiments to automatically acquire the syntax and semantics of MWEs was presented in Chapters 5–8. In each experiment, we trialled different approaches to resolve the problems focused, such as evaluating statistical approaches that are suited to specific MWE types and tasks.

The approaches presented in this thesis were:

1. Automatic interpretation of NCs (Chapters 5, 6 and 7)

2. Disambiguating word senses of NCs (Chapter 7)

3. Modeling the compositionality of VPCs (Chapter 5).

4. Identifying VPCs (Chapter 8)

We will present a brief summary of the approaches and how they performed in the following sections.

9.1.1 Interpreting Noun Compounds

We proposed three different models to automatically interpret noun compounds. We undertook a study based on statistical approaches, namely: semantic similarity, substitutability (in Chapters 5 and 7) and ellipsed predicate (in Chapter 6).

Constituent Similarity Method

The first method (called the constituent similarity method, based on semantic similarity from Section 5.1) was founded on lexical similarity between test and training instances. The intuition behind this method is that when the union of senses of constituents in NCs is similar to that of seen NCs (training NCs), their semantic relation (SR) is the same. It is based on the same motivation as the sense collocation method, but uses implicit sense collocation through the union of the senses of the NC components.

Our primary method to interpret noun compounds based on semantic similarity was to use the highest similarity with the test instance in the form of a 1-nearest neighbor classifier. In detail, first, we compute the similarities of the head noun and modifier, respectively, for a given test and training instance pairing, and combine the two similarities by multiplying them. For a given test instance, we select the training instance with the highest similarity and assign its SR to the test instance.

We experimented with this basic method in the following ways:

1. evaluate the method over 2-term NCs

2. evaluate the method over 3-term NCs

3. test the contribution of each NC component (i.e. head noun and modifier)

4. combine the proposed method with bootstrapping in order to acquire more NCs at no manual overhead

5. apply the k-nearest neighbor algorithm to smooth the results of classification

6. utilize the acquired NCs for bracketing NCs with 3 or more terms.

Our finding with the constituent similarity method was that it identifies “sense collocations” implicitly at the word level. We observed that when sense collocations were found in both the test and train dataset, the constituent similarity method performed comparably with the sense collocation method of Moldovan *et al.* (2004), but if these are not available, our method performed better since it has the ability to “fuzzy-match” relative to WordNet. We also attested the contribution of the individual component (i.e. head noun and modifier) over different SRs. Furthermore, we confirmed that the constituent similarity method is easily implementable and adoptable in other hybrid approaches (as shown in Sections 5.1 and 7.1). Finally, we showed the potential to use semantic relations in applied contexts with the bracketing task.

Ellipsed Predicate Method

The ellipsed predicate method was built on the hypothesis that SRs in NCs could be disambiguated via ellipsed predicates using verb semantics. The main idea of this method is to use the hidden semantics within NCs to interpret semantic relations. That is, the SRs are implicitly encoded in NCs, and can be extracted in the form of directed predicates corresponding to the definition/description of SRs. This method resembles previous studies based on ellipsed predicates (Levi 1978; Vanderwende 1994; Lapata 2002), and literally uses the definition of SRs to disambiguate them. However, in our approach, we expand the set of corpus-based evidence through a novel verb mapping method using lexical similarity to automatically acquire the verb semantics in the form of related predicates.

In detail, we first defined the clausal form of each SR using verbs and the directed templates. The verbs in these templates were labelled seed verbs, and describe the semantics of each SR. In the second step, we used a verb-mapping method in order to map the non-seed verbs onto seed verbs in clauses including both the head noun and modifier. Once individual clauses were classified in terms of SRs, we finally built a supervised classifier using TiMBL v5.1.

Our main contribution with this method was to overcome the data sparseness experienced by earlier implementations of the ellipsed predicate method, by introducing a verb mapping method. The verb-mapping method provides a simple and low-cost approach to obtaining the verb semantics between the head noun and its modifier(s). We also observed that this idea can be applied in an unsupervised context using the constituent substitution method described in Section 7.1. To overcome data sparseness, we expanded the set of training NCs using the constituent substitution method and extracted clauses containing both the original NCs and an expanded set of NCs, and used both as evidence to compute the probability of SRs.

Constituent Substitution Method

The constituent substitution method used the semantic similarity and substitutability of NCs as indicators of sense collocation to disambiguate the SR. The motivation behind this work is similar to that for the semantic similarity approach to NC interpretation (Rosario and Marti 2001; Moldovan *et al.* 2004; Kim and Baldwin 2005; Nastase *et al.* 2006; Girju 2007). That is, we hypothesize that the production of NCs is restricted by the way in which SRs are constructed, despite the potential for unlimited combination of nouns. Hence, as long as the sense collocation of two NCs is the same or similar, they have the same SR.

To model this intuition, our basic method is to replace one component at a time in a given NC by a similar word in order to modify the NC but preserve the same or similar sense collocation. This method was also combined with bootstrapping in order to enlarge the number of interpreted NCs. That is, we used the newly-generated NCs as a new set of base NCs in generating/interpreting NCs on the next iteration. The process can be repeated as long as new NCs are generated or the pool of similar words is exhausted.

In our experiments, we confirmed that the NCs acquired by the constituent substitution method have higher accuracy than those from previous methods. We also confirmed that, as expected, bootstrapping provides a significant performance boost to the constituent substitution method. We were able to automatically expand the set of interpreted NCs with nearly exponential growth under the same or similar scope of sense collocation. Hence, we claim that this approach provided a partial solution to the knowledge acquisition bottleneck by simple bootstrapping, in that we could automatically acquire a large number of training instances for NC interpretation. One drawback of this method, however, was that the generated NCs were all from the same or similar sense collocations. Hence, the variation of sense collocation in the acquired NCs was small. We also found that existing methods can be combined with this approach to provide marginally better results. We found that the variation and scale of training data are key factors in improving NC interpretation performance, as most of methods presented are supervised.

9.1.2 Disambiguating the Word Sense of Noun Compounds

We proposed a word sense disambiguation method for NCs, to aid in NC interpretation as well as to enhance the performance of a state-of-the-art WSD system. The intuition underlying this work is that sense-disambiguated NC elements are better predictors of the word sense of polysemous nouns in NCs than the context of usage. Also, by focusing on the elements of the NC, we were able to bring the one sense per collocation heuristic into play, in assuming that the elements in a given NC will always occur with the same sense.

We proposed WSD methods using supervised and unsupervised techniques, by employing semantic similarity, substitutability and co-occurrence properties. For both the supervised and unsupervised methods, we used sense collocation. However, for the supervised method, we integrated the roles of the NC constituents since the distribution of word senses of the constituents can vary considerably according to their roles in the NC. For the unsupervised method, we employed substitutability to replace one of the constituents at a time and generate new NCs within the same or similar sense collocation, wherein we used web co-occurrence data to disambiguate the word sense.

In evaluation, we compared our methods with the start-of-art SenseLearner WSD system. Our approach is, to the best of our knowledge, the first method to provide reliable disambiguation of word senses of MWEs (in our case, NCs). Our methods showed that not only could we successfully harness techniques developed for the WSD of simplex words in disambiguated the constituents of an NC, but that the WSD of MWE constituents could take advantage of MWE-specific features. We further confirmed that grammatical role plays an important role in the disambiguation of NC word sense, as the sense distribution is biased according to the grammatical role. In addition, we found that existing WSD techniques/methods for simplex words are ineffectual at sense disambiguating MWEs.

9.1.3 Modeling the Compositionality of VPCs

Our method for modeling the compositionality of English VPCs in Chapter 5 was motivated by the claim by Villavicencio (2005) that similar verbs and particles combine to form VPCs with similar semantics. We apply this to the task of VPC compositionality modeling in assuming that compositional VPCs can be predicted from regular patterns of combination of verbs and particles.

To implement this idea, we represented VPCs by simple collocational information and semantic classes, for each of the verb and particle, and classified VPCs accordingly. To represent the semantics of each verb, we used the first sense and its 3 hypernyms from WordNet. For particle semantics, we used the set of particle semantics defined by Bannard (2003). Based on these features, we built both a 1-dimensional and a 2-dimensional matrix representation, based on previous work on Japanese compound verbs (Uchiyama *et al.* 2005).

In our evaluation, we found that semantic similarity can effectively detect the compositionality of VPCs. Our method was also able to detect the semantic contribution of the VPC component. We also provided evidence for the cross-lingual utility of the method developed over Japanese compound verbs. Finally, we confirmed that the performance measure suggested by McCarthy *et al.* (2003) is reliable for testing the compositionality of MWEs.

9.1.4 Identifying VPCs

Our VPC identification method was based on linguistic properties of VPCs. That is, the selectional preferences of VPCs over predefined argument positions provided insight into whether a verb and preposition in a given sentential context combine to form a VPC or alternatively constitute a verb-PP.

We implemented the proposed method using the syntactic and (shallow) semantics of nouns governed by the target verbs. We retrieved the syntactic features of candidates (either VPCs or Verb-PPs), and also extracted the semantics of subject and object(s) of the verb in order to model the selectional preferences of the VPC. In detail, in order to capture the semantics of the nominal arguments, we once again used a simple first-sense method (McCarthy *et al.* 2004) with *hypernym* expansion, based on WordNet. For pronouns and proper nouns, we experimented with a number of methods for capturing their common noun semantics. Our approach is based centrally on RASP, in that VPC candidates are extracted from the output of RASP. Further, we tested the importance of resolving pronouns and proper nouns to acquire the semantics of nouns.

In our evaluation, we found that our proposed method performed well, and considerably reduced the reliance on manual annotation required in prior research (Li *et al.* 2003). Our main finding was that syntactic and semantic features of VPCs and verb-PPs, including the semantics of nominal arguments, are good predictors for differentiating VPCs from Verb-PPs. We also attested the performance of existing resources on VPC identification. Finally, we attested that resolving the pronouns and proper nouns is a critical component in improving the performance of the proposed method.

9.2 Directions for Further Research

In this thesis, we have presented a variety of techniques for modeling MWEs. Although some of the addressed issues are resolved partially or in full, many of the tasks remain unsolved. In this section, we wrap up by pointing out the remaining tasks and mapping out a future direction for MWE research, focused particularly on the target MWEs in this thesis: NCs and VPCs.

9.2.1 Future Research on Noun Compounds

Our future work on NCs will focus on four major issues:

- develop more unsupervised methods
- propose a reliable set of semantic relations (SRs) along with comparison methods
- adopting existing methods to other types of English MWEs, and NCs in other languages (cross-over and cross-lingual study)
- apply the research in real-world NLP applications

The first focus is on unsupervised methods for automatically interpreting NCs. There has been some prior work on using automated approaches, but with the notable exceptions of Lapata and Keller (2004), most of the research has been founded on supervised methods. Such supervised approaches are heavily reliant on large amounts of training data, resulting in a high cost in terms of both time and computation. Additionally, the true performance of supervised methods is hard to measure since it depends crucially on the context and the size of the test data.

While Lapata and Keller (2004) proposed an interesting approach, the performance of the method is seriously lacking compared to supervised methods. The approach relied on a small set of (coarse-grained) semantic relations and noisy web counts based on queries to Google, which might have caused the method to underperform. Apart from this work, there has not been any serious attempt to construct unsupervised methods for NC interpretation. More focus should be placed on the development of unsupervised approaches to NC interpretation, to avoid the human labor overhead and achieve higher utility.

We secondly need a standardized set of semantic relations (SRs) and benchmarking frameworks to objectively evaluate different approaches. So far, all proposed methods have been evaluated under certain assumptions and experimental setups that makes comparison difficult. The field is still a long way from agreeing on a standardized set of SRs, and arguments for and against particular SR analyses are generally highly subjective. Ó Séaghdha (2007) succinctly described six desiderata for an ideal SR set: *coverage*, *coherence*, *balance*, *generalization*, *ease of annotation* and *utility*. Coverage of SRs is a measurement of the quantity of NCs which can be successfully interpreted by the defined SRs. In terms of this scheme, the sets of SRs defined by Levi (1978) and Lauer (1995) suffer from under-coverage, while Finin (1980) exaggerated the number of SRs. Coherence is the requirement for clear differentiation among the SRs, and relates to ease of annotation, as clear demarcation of SRs avoids ambiguity when annotating NCs. Balance refers to the uniformity of NCs across the set of SRs. As many (supervised) NC interpretation methods are based on machine learning techniques, this criteria relates to the "machine tractability" of a given SR set. Generalization means that the concepts underlying the categories should be generalizable to other linguistic phenomena. It is related to the quality of SRs which capture the linguistic nature of NCs. Ease of annotation refers to the provision of adequate annotation guidelines to human annotators to support consistent annotation. Finally, utility is the requirement that the SRs capture useful semantic distinctions in terms of how the annotated data will be applied.

Our ideal of a standard set of SRs takes all of these into consideration. However, emphasis must be placed on the quality and quantity of SRs, i.e. to coherence and generalization for quality assurance, and coverage for quantity (= scalability). Our proposed solution is to resolve this trade-off using multivalued semantic relations (SRs) that can group several closely-related SRs. Research on word sense disambiguation (WSD) has shown that using super-tags (in the form of unique beginners or lexicographic files in the context of WordNet) is helpful in resolving the granularity problem. Nastase *et al.* (2006) recently presented such an approach using five parent classes of 30 SRs, and Ó Séaghdha (2007) similarly proposed a hierarchical SR set. We propose to look at the cross-over of several existing sets of SRs to create a set of abstract SRs. Human annotators can then be used to test the reliability of the groupings (over small set of NCs). We believe that using such abstract "top-level" SRs will help resolve problems such as fine-grained granularity and overly-strict evaluation.

The third area for future research is to utilize existing methods to model different types of MWEs, as well as NCs in other languages. As we described in Section 1.2.6, cross-over and cross-lingual research is essential in order to resolve overall issues. Some research has shown that existing methods can be adopted from one type of MWE to another, as well as form one language to another (Katz and Giesbrecht 2006; Kim and Baldwin 2007a). Hence, more attempts at cross-over and cross-lingual research will bring benefits to not only one specific language/MWE type, but across the board for MWE research in various languages.

As part of our research on NC interpretation, we have shown that heuristics to extract noun semantics using WordNet developed for VPC identification (Figure 8.3), can be successfully applied to word sense disambiguation (Figure 7.9). Hence, we strongly believe that our proposed methods can be successfully applied to other types of English MWEs. Furthermore, we witnessed the benefit of cross-lingual research in Girju (2007) and several attempts to interpret NCs in other languages (Yoon *et al.* 2001; Zhao *et al.* 2007). Our intuition in cross-lingual studies is that languages in the same family have similar linguistic phenomena and can be analyzed using similar techniques. In addition, the same or similar linguistic features can be easily detected within the same language family.

Our final proposed area of future research in the applications of the proposed work on NCs to real-world NLP applications. As many researchers have suggested, NC interpretations could be used to populate a knowledge base and improve the robustness and fluency of NLP applications. Systems that should particularly benefit from this are question-answering (QA) and summarization systems, particularly in the areas of topic detection and text generation. Although there have been few attempts to utilize SRs, Venkatapathy and Joshi (2006) showed the potential for employing SRs in NLP applications, and Moldovan *et al.* (2004) showed specific examples where SRs appear to aid QA. For example, with the question, *What did the factory in Howell Michigan make?*, the verb *make* suggests that SR MAKE or PRODUCE may be relevant to the answer. In addition, Nakov and Hearst (2006) made suggestions about indirect applications of SRs in the medical domain. For example, *migraine treatment* as a query would prefer documents containing verbs such as *relieve* and *prevent*, based on the SR of *migraine treatment* (as *treatment* which relieves or prevents a *migraine*). In summary, we aim to furnish evidence that will employ these direct and indirect clues from interpreted NCs to improve the performance of NLP applications.

9.2.2 Future Research on Verb-Particle Constructions

Future research directions for verb-particle constructions are:

- develop unsupervised methods for extraction and identification
- adopt existing methods to other types of English MWEs and VPCs in other languages (cross-over and cross-lingual study)
- apply the research in real-world NLP applications

We are interested in developing unsupervised methods for extracting/identifying VPCs, and in utilizing the data in modeling the compositionality of multiword expressions (MWEs). Previous research on extracting (Baldwin 2005a) and identifying VPCs (Li *et al.* 2003; Kim and Baldwin 2006a) has been relatively successful. However, most current systems are based on supervised methods that require large amounts of training data and are hence not easily applicable to lexical-tuning over novel domains or new languages. Existing methods also rely heavily on parsers, which ties their performance to the performance of parsers. Hence, we suggest that future research should focus on unsupervised methods that do not depend heavily on parsers and large training data sets.

We also propose that existing methods be applied to other types of MWEs in English as well as other languages. Much research has been done to extract verb–complement pairs (e.g. verb-particle constructions, light-verb constructions, prepositional verbs or verb-noun combinations) in English and other languages (Bame 1999; Baldwin *et al.* 2003a; Bannard *et al.* 2003; Baldwin 2005b; Venkatapathy and Joshi 2005; Uchiyama *et al.* 2005; Cook and Stevenson 2006). While the target of this research has been varied, the basic methodology that has been used is relatively constant. Previous research (this thesis included) has predominantly used machine learning techniques in combination with relatively similar features relating to the linguistic properties of the target construction. There is potential to generalize methods over related linguistic phenomena.

The last proposed research direction relating to VPCs is to apply the outcomes of VPC research in real-world NLP applications. VPCs can be used in formulating rich queries for applications such as question answering and general-purpose information retrieval. They can also improve the fluency of text generation, and a detailed notion of MWE compositionality can be used to emphasize or add nuance to lexical items. It can also provide lexical variety, and particles can be used to "nuance" simplex verbs. Although the semantic contribution of particles in compositional VPCs is weak, they are able to manipulate the semantics of VPCs to a certain degree. For example, *up* in *eat up* adds the degree of completion to the verb *eat*. Finally, such applications have high real-world applicability and would benefit immensely from performance improvements in compositionality modeling.

Bibliography

ABEILLÉ, ANNE. 1988. Light verb constructions and extraction out of NP in a tree adjoining grammar. In *Papers of the 24th Regional Meeting of the Chicago Linguistics Society*.

AGIRRE, ENEKO, and PHILIP EDMONDS. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht, Netherlands: Springer.

——, and DAVID MARTINEZ. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of COLING workshop on Semantic Annotation and Intelligent Content*, 11–19, Saarbrucken, Germany.

ALBA-SALAS, JOSEP, 2002. *Light Verb Constructions in Romance. A Syntactic Analysis*. Cornell University dissertation.

BALDWIN, TIMOTHY. 2005a. The deep lexical acquisition of English verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions* 19.398–414.

——. 2005b. Looking for prepositional verbs in corpus data. In *Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, 115–126, Colchester, UK.

——, COLIN BANNARD, TAKAAKI TANAKA, and DOMINIC WIDDOWS. 2003a. An empirical model of multiword expression decomposability. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 89–96, Sapporo, Japan.

——, JOHN BEAVERS, LEONOOR VAN DER BEEK, FRANCIS BOND, DAN FLICKINGER, and IVAN A. SAG. 2003b. In search of a systematic treatment of determinerless PPs. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, 145–156, Toulouse, France.

——, John Beavers, Leonor Van Der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. 2006. In search of a systematic treatment of determinerless PPs. In *Syntax and Semantics of Prepositions*, ed. by Patrick Saint-Dizier. Springer.

——, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, 2047–2050, Lisbon, Portugal.

——, and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, 24–31, Barcelona, Spain.

——, and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, 98–104, Taipei, Taiwan.

Bame, Ken, 1999. Aspectual and resultative verb-particle constructions with up. Handout for talk presented at the Ohio State University Linguistics Graduate Student Colloquium.

Banerjee, Satanjeev, and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003)*, 805–810, Acapulco, Mexico.

Bannard, Colin, 2003. Statistical techniques for automatically inferring the semantics of verb-particle constructions. Master's thesis, University of Edinburgh.

——, 2006. *Acquiring Phrasal Lexicons from Corpora*. University of Edinburgh, UK dissertation.

——, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 65–72, Sapporo, Japan.

Barker, Ken, and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998)*, 96–102, Montreal, Canada.

Bauer, Laurie. 1983. *English Word-formation*. Cambridge, UK: Cambridge University Press.

Bolinger, Dwight, 1976a. Meaning and memory.

——. 1976b. *The Phrasal Verb in English*. Boston, USA: Harvard University Press.

Bond, Francis, 2001. *Determiners and number in English, contrasted with Japanese, as exemplified in machine translation*. Brisbane, Australia: University of Queensland dissertation.

Borthen, Kaja, 2003. *Norwegian bare singulars*. Norwegian University of Science and Technology dissertation.

Brinton, Laurel. 1985. Verb particles in English: Aspect or aktionsart. *Studia Linguistica* 39.157–168.

Briscoe, Ted, and John Carroll. 2002. Accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, 1499–1504, Las Palmas, Canary Islands.

Buckeridge, Alan M., and Richard F. E. Sutcliffe. 2002. Disambiguating noun compounds with Latent Semantic Indexing. In *Proceedings of the 2nd International Workshop on Computational Terminology*, Patras, Greece.

Buitelaar, Paul, 1989. *CoreLex: Systematic Polysemy and Underspecification*. Brandeis University dissertation.

Burnard, Lou, 1995. User guide for the British National Corpus.

Butt, Miriam. 2003. The light verb jungle. In *Proceedings of the Workshop on Multi-verb Constructions*, 1–49, Trondheim, Norway.

Calzolari, Nicoletta, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, 1934–1940, Las Palmas, Canary Islands.

Cao, Yunbo, and Hang Li. 2002. Base noun phrase translation using web data and the EM algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, 37–40, Taipei, Taiwan.

Carpuat, Marine, and Dekai Wu. 2005. Word sense disambiguation vs. staitstical machine translation. In *Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics*, 387–394, Ann Arbor, USA.

Chafe, Wallace L. 1968. Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language* 4.109–127.

Chander, Ishwar, 1998. *Automated postediting of documents*. University of Southern California dissertation.

Charniak, Eugene. 2000. A maximum entropy-based parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics*, Seattle, USA.

Choueka, Yaacov. 1988. Lookin for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of RIAO*, 43–38.

——, Shmuel T. Klein, and E. Neuwitz. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal for Literary and Linguistic* 4.34–38.

Church, Kenneth W., and Patrick Hanks. 1989. Word assication norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the Association of Computational Linguistics (ACL-1989)*, 76–83, Vancouver, Canada.

Cook, Paul, and Suzanne Stevenson. 2006. Classifying particle semantics in English verb-particle constructions. In *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 45–53, Sydney, Australia.

——, and ——. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, 41–48, Prague, Czech Republic.

Copestake, Ann, and Alex Lascarides. 1997. Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proceedings of the 35th Annual Meeting of the Association of Coomputational Linguistics and 8th Conference of the European Chapter of Association of Computational Linguistics (ACL/EACL-1997)*, 136–143, Madrid, Spain.

Cruse, Alan D. 1986. *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, 2004. TiMBL: Tilburg memory based learner, version 5.1, reference guide.

Dehe, Nicole. 2002. *Particle Verbs in English: syntax, information structure and intonation*. Amsterdam/Philadelphia: John Benjamins Publishing.

——, Ray Jackendoff, Andrew McIntyre, and Silke Urban (eds.) 2001. *Verb-Particle Explorations*. Berlin/New York: Mounton de Gruyter.

Dias, Gaël, S. Guilloré, and J. G. Pereira Lopes. 1999. Multilingual aspects of multiword lexical units. In *Workshop on Language Technologies in the Framework of the 32rd Annual Meeting of the* Societas Linguistica Europaea, Ljubljana, Slovenia.

Dirven, René. 2001. The metaphoric in recent cognitive approaches to English phrasal verbs. *metaphorik.de* 1.39–54.

Downing, Pamela. 1977. On the creation and use of English compound nouns. *Language* 53.810–842.

Dras, Mark, and Mike Johnson. 1996. Death and lightness: Using a demographic model to find support verbs. In *Proceedings of the 5th International Conference on the Cognitive Science of Natural Language Processing*, 165–172, Dublin, Ireland.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19.61–74.

Evert, Stephen, 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. University of Stuttgart dissertation.

Fan, James, Ken Barker, and Bruce W. Porter. 2003. The knowledge required to interpret noun compounds. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1483–1485, Acapulco, Mexico.

Fazly, Afsaneh, Ryan North, and Suzanne Stevenson. 2005. Automatically distinguishing literal and figurative usages of highly polysemous verbs. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, 38–47, Ann Arbor, USA. Association for Computational Linguistics.

——, and Suzanne Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, 9–16, Prague, Czech Republic.

Fellbaum, Christiane (ed.) 1998. *WordNet, An Electronic Lexical Database*. Cambridge, Massachusetts, USA: MIT Press.

Fernando, Chitra, and Roger Flavell. 1981. *On idioms*. Exeter: University of Exeter.

FILLMORE, CHARLES, PAUL KAY, and MARY C. O'CONNOR. 1988. Regularity and idiomaticity in grammatical constructions. *Language* 64.501–538.

FININ, TIMOTHY WILKING, 1980. *The semantic interpretation of compound nominals*. University of Illinois, Urbana Champaign dissertation.

FIRTH, JOHN RUPERT. 1957. A Synopsis of Linguistic Theory, 1933-1955. In *Studies in Linguistic Analysis*, ed. by J. R. Firth, 1–32. Oxford: Blackwell.

FOLLI, RAFFAELLA, HEIDI HARLEY, and SIMIN KARIM. 2003. *Determinants of even type in Persian complex predicates*. Cambridge Working Papers in Linguistics.

FRASER, BRUCE. 1976. *The Verb-Particle Combination in English*. The Hague: Mouton.

GATES, EDWARD. 1988. *The treatment of multiword lexemes in some current dictionaries of English*. Snell-Hornby.

GIBBS, RAYMOND W. 1980. Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition* 8.149–156.

GIRJU, ROXANA. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 568–575, Prague, Czech Republic.

——, DAN MOLDOVAN, MARTA TATU, and DANIEL ANTOHE. 2005. On the semantics of noun compounds. *Computer Speech and Language* 19.479–496.

——, PRESLAV NAKOV, VIVI NASTASE, STAN SZPAKOWICZ, PETER TURNEY, and DENIZ YURET. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th Semantic Evaluation Workshop(SemEval-2007)*, 13–18, Prague, Czech Republic.

GREFENSTETTE, GREGORY, and SIMUAL TEUFEL. 1994. What is a word, what is a sentence? problems of tokenization. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research*, 79–87, Budapest, Hungary.

——, and ——. 1995. A corpus-based method for automatic identification of suport verbs for nominalizations. In *Proceedings of the 7th European Chapter of Association of Computational Linguistics (EACL-1995)*, 98–103, Dublin, Ireland.

GRIES, STEFAN T. 1999. Particle movement: A cognitive and functional approach. *Cognitive Linguistics* 10.105–145.

GRISHMAN, RALPH, CATHERINE MACLEOD, and ADAM MYERS, 1998. COMLEX syntax reference manual.

GROVER, CLAIRE, MARIA LAPATA, and ALEX LASCARIDES. 2004. A comparison of parsing technologies for the biomedical domain. *Journal of Natural Language Engineering* 1.1–38.

HASPELMATH, MARTIN. 1997. *From Space to Time in The World's Languages*. Munich, Germany: Lincorn Europa.

HEARST, MARTI. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING '92)*, Nantes, France.

HERMJAKOB, ULF, EDUARD HOVY, and CHIN-YEW LIN. 2002. Automated question answering in Webclopedia. In *Proceedings of the ACL-02 Demonstrations Session*, 98–99, Philadelphia, USA.

HIRST, GRAEME, and DAVID ST-ONGE. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In (Fellbaum 1998), 305–332.

HOSHI, H., 1994. *Passive, Causive, and Light Verbs: A Study of Theta Role Assignment*. University of Connecticut dissertation.

HUDDLESTON, RODNEY, and GEOFFREY K. PULLUM. 2002. *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

HULL, RICHARD D., and FERNANDO GOMEZ. 1996. Semantic interpretation of nominalizations. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-1996)*, 1062–1068, Portland, Oregon.

HUMPHREYS, L., D. LINDBERG, H. SCHOOLMAND, and G.O. BARNETT. 1998. The unified medical language system: An informatics research collabration. *Journal of the American Medical informatics Assocation* 5.1–13.

IDE, NANCY, and JEAN VERONIS. 1998. Word sense disambiguation : The state of the art. *Computational Linguistics* 24.1–40.

ISABELLE, PIERRE. 1984. Another look at nominal compounds. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING-1984)*, 509–516, San Francisco, USA.

ISHIKAWA, K. 1999. Enlish verb-particle constructions and V-internal structure. *English Linguistics* 16.329–352.

JACKENDOFF, RAY. 1973. The base rules for prepositional phrases. In *A Festschrift for Morris Halles*, 345–356. New York: Halt: Rinehart and Winston.

——. 1997. *The Architecture of the Language Faculty*. Cambridge, USA: MIT Press.

——. 2002. *Foundation of Language*. Oxford, UK: Oxford University Press.

Jespersen, Otto. 1965. *A Modern English Grammar on Historical Principles, Part VI, Morphology*. London, UK: George Allen and Unwin Ltd.

Jiang, Jay, and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, 19–33, Taipai, Taiwan.

Johnston, Michael, and Frederica Busa. 1996. Qualia structure and the compositional interpretation of compounds. In *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, 77–88, Santa Cruz, USA.

Kan, Yee Fan Tan Min-Yen, and Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context (MWEmc)*, Trento, Italy.

Kastovsky, Dieter. 1982. *Wortbildung und Semantik*. Dusseldorf: Bagel/Francke.

Katz, Graham, and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 28–35, Sydney, Australia.

Katz, Jerrold J., and Paul M. Postal. 2004. Semantic interpretation of idioms and sentences containing them. In *Quarterly Progress Report (70), MIT Research Laboratory of Electronics*, 275–282. MIT Press.

Kayne, Richard S. 1985. Principles of particle constructions. In *Jacqueline Gueron*, 101–140. Dordrecht:Foris: Grammatical Representation.

Kearns, Kate, 2002. Light verbs in English.

Kim, Su Nam, and Timothy Baldwin. 2005. Automatic interpretation of compound nouns using WordNet similarity. In *Proceedings of 2nd International Joint Conference on Natual Language Processing (IJCNLP-2005)*, 945–956, Jeju, Korea.

——, and ——. 2006a. Automatic extraction of verb-particles using linguistic features. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, 65–72, Trento, Italy.

——, and ——. 2006b. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics (COLING/ACL-2006)*, 491–498, Sydney, Australia.

——, and ——. 2007a. Detecting compositionality of English verb-particle constructions using semantic similarity. In *Proceedings of Conference of the Pacific Association for Computational Linguistics*, 40–48, Melbourne, Australia.

——, and ——. 2007b. Disambiguating noun compounds. In *Proceedings of 22nd AAAI Conference on Artificial Intelligenc*, 901–906, Vancouver, Canada.

——, and ——. 2007c. Interpreting noun compounds using bootstrapping and sense collocation. In *Proceedings of Conference of the Pacific Association for Computational Linguistics*, 129–136, Melbourne, Australia.

——, and ——. 2007d. Melb-kb: Nominal classification as noun compound interpretation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, 231–236, Prague, Czech Republic.

——, and ——. 2008. Benchmarking noun compound interpretation. In *Proceedings of 3rd International Joint Conference on Natual Language Processing (IJCNLP-2008)*, 569–576, Hyderabad, India.

——, Meladel Mistica, and Timothy Baldwin. 2007. Extending sense collocation on interpreting noun compounds. In *Proceedings of Australasian Language Technology Workshop*, 49–56, Melbourne, Australia.

Landauer, Thomas K., Peter W. Faltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes* 25.259–284.

Lapata, Maria. 2002. The disambiguation of nominalizations. *Computational Linguistics* 28.357–388.

Lapata, Mirella, and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Langauge Techinology Conference and Conference on Empirical Methods in National Language Processing (HLT/NAACL-2004)*, 121–128, Boston, USA.

Lauer, Mark, 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Macquarie University dissertation.

Leacock, Claudia, and Nancy Chodorow. 1998. *Combining local context and WordNet similarity for word sense identification*. Cambridge, USA: MIT Press.

LEVI, JUDITH. 1978. *The Syntax and Semantics of Complex Nominals*. New York, New York, USA: Academic Press.

LI, WEI, XIUHONG ZHANG, CHENG NIU, YUANKAI JIANG, and ROHINI K. SRIHARI. 2003. An expert lexicon approach to identifying English phrasal verbs. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 513–520, Sapporo, Japan.

LIBERMAN, MARK, and RICHARD SPROAT. 1992. The stress and structure of modified noun phrases in English. In *Lexical Matters – CSLI Lecture Notes No. 24*, ed. by Ivan A. Sag and A. Szabolcsi. Stanford, USA: CSLI Publications.

LIDNER, SUE., 1983. *A lexico-semantic analysis of English verb particle constructions with OUT and UP*. University of Indiana at Bloomington dissertation.

LIN, DEKANG. 1993. Principle-based parsing without overgeneration. In *Proceedings of the 31th Association of Computational Linguistics (ACL-1993)*, 112–120, Columbus, Ohio, USA.

——. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL-1998)*, 768–774, Montreal, Canada.

——. 1998b. Extracting collocations from text corpora. In *Proceedings of the 1st Workshop on Computational Terminology*, Montreal, Canada.

——. 1998c. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, 296–304, Madison, Wisconsin, USA.

——. 1998d. Using collocation statistics in information extraction. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, USA.

——. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Association of Computational Linguistics (ACL-1999)*, 317–324, College Park, Maryland, USA.

LOHSE, BARBARA, JOHN A. HAWKINS, and THOMAS WASOW. 2004. Domain minimization in English verb-particle constructions. *Language* 80.238–261.

LYNOTT, DERMOT, and MARK T. KEANE. 2004. A model of novel compound production. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, Illinois, USA.

MacQueen, James B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistcs and Probability*, 281–297, Berkeley, USA. University of California at Berkeley Press.

Marcus, Mitchell. 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge, USA: MIT Press.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics* 19.313–330.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 73–80, Sapporo, Japan.

——, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*, 280–287, Barcelona, Spain.

——, Sriram Venkatapathy, and Aravind Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 200 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 369–379.

Mihalcea, Rada, and Ehsanul Faruque. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of the ACL/SIGLEX Senseval-3*, 155–158, Barcelona, Spain.

——, and Dan Moldovan. 1999. An automatic method for generating sense taggged corpora. In *Proceedings of the 16th Conference of the American Association of Aritificial Intelligence (AAAI-1999)*, 461–466, Orlando, USA.

Mimmelmann, Nikolaus P. 1998. Regularity in irregularity: Article use in adpositional phrases. *Linguistic Typology* 2.315–353.

Minnen, Guido, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering* 7.207–223.

Miyagawa, Shigeru. 1989. Light verbs and the ergative hypothesis. *Linguistic Inquiry* 20.659–668.

Moldovan, Dan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases.

In *Proceedings of HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, 60–67, Boston, USA.

Nakov, Preslav, and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketting. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, 17–24, Ann Arbor, USA.

——, and ——. 2006. Using verbs to characterize noun-noun relations. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA)*, 233–244, Bularia.

Nastase, Vivi, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 781–787, Boston, USA.

Ngai, Grace, and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL)*, 40–47, Pittsburgh, USA.

Nicholson, Jeremey, and Timothy Baldwin. 2005. Statistical interpretation of compound nominalisations. In *Proceedings of the Australian Language Technology Workshop*, 152–159, Sydney, Australia.

Nulty, Paul. 2007. Semantic classification of noun phrases using web counts and learning algorithms. In *Proceedings of the Association of Computational Linguistics 2007 Student Research Workshop*, 79–84, Prague, Czech Republic.

Nunberg, Geoffrey, Ivan A. Sag, and Tom Wasow. 1994. Idioms. *Language* 70.491–538.

O'Dowd, Elizabeth M. 1998. *Prepositions and Particles in English*. Oxford University Press.

O'Hara, Tom, and Janyce Wiebe. 2003. Preposition semantic classification via Treebank and framenet. In *Proceedings of the 7th Conference on Natural Language Learning*, 79–86, Edmonton, Canada.

Olsen, Susan. 2000. Against incorporation. In *Linguistische Arbeitsbertichte 74*, 149–172. University of Leipzig.

Ó Séaghdha, Diarmuid. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the 4th Corpus Linguistics Conference*.

——, and Ann Copestake. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, 57–64, Prague, Czech Republic.

Patrick, Jon, and Jeremy Fletcher. 2004. Differentiating types of verb particle constructions. In *Proceedings of Australian Language Technology Workshop*, 163–170, Sydney, Australia.

Patwardhan, Siddharth, 2003. Incorporating dictionary and corpus information into a context vector. Master's thesis, University of Minnesota, USA.

——, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*, 17–21, Mexico City, Mexico.

Pawley, Andrew, and Frances Hodgetts Syder, 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency.

Peabody, K.W., 1981. Constraints on the productivity of verb-particle combinations. Master's thesis, Ohio State University.

Pearce, Darren. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, 41–46, Pittsburgh, Pennsylvania, USA.

Pecina, Pavel. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, 13–18, Ann Arbor, USA. Association for Computational Linguistics.

Piao, Scott, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions wth a semantic tagger. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 49–56, Sapporo, Japan.

——, Paul Rayson, Olga Mudraya, Andrew Wilson, and Roger Garside. 2006. Measuring mwe compositionality using semantic annotation. In *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 2–11, Sydney, Australia.

Potter, Elizabeth, Jenny Watson, Michael Lax, and Miranda Timewell. 2000. *Collins Cobuild Dictionary of Idioms*. Cambridge, UK: Harper Collins Publishers.

Prager, J., and J. Chu-Carroll, 2001. Use of WordNet hypernyms for answering what-is questions.

PUSTEJOVSKY, JAMES. 1995. *The Generative Lexicon*. Cambridge, USA: MIT press.

QUIRK, RANDOLPH, SYDNEY GREENBAUM, GEOFFREY LEECH, and JAN SVARTVIK. 1985. *A Comprehensive Grammar of the English Language*. London, UK: Longman.

RAMCHAND, GILLIAN, and PEER SVENONIUS. 2002. The lexical syntax and lexical semantics of the verb-particle construction. In *Proceedings of WCCFL*, 387–400, Somerville, USA.

RESNIK, PHILIP. 1995. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the 3rd Workshop on Very Large Corpus*, 77–98, Cambridge, USA.

RIEHEMANN, SUSANNE, 2001. *A Constructional Approach to Idioms and Word Formation*. Stanford University dissertation.

ROSARIO, BARBARA, and HEARST MARTI. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, 82–90, Pittsburgh, Pennsylvania, USA.

ROSS, HAJ. 1995. Defective noun phrases. In *In Papers of the 31st Regional Meeting of the Chicago Linguistics Society*, 398–440, Chicago, Illinois, USA.

SAG, IVAN A., TIMOTHY BALDWIN, FRANCIS BOND, ANN COPESTAKE, and DAN FLICKINGER. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, 1–15, Mexico City, Mexico.

SALTON, GERARD, ALLAN WONG, and C.S. YANG. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18.613–620.

SANDERSON, MARK, 1996. *Word sense disambiguation and Information retrieval*. University of Glasgow dissertation.

SCHONE, PATRICK, and DANIEL JURAFSKY. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 6th conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, 100–108, Hong Kong.

SCHUTZE, HINRICH. 1998. Automatic word sense discrimination. *Computational Linguistics* 24.97–123.

SIDE, RICHARD. 1990. Phrasal verbs: sorting them out. *ELT Journal* 44.144–52.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19.143–77.

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17*, 1297–1304, Vancouver, Canada.

Sparck Jones, Karen. 1983. *Compound noun interpretation problems*. Englewood Cliffes, USA: Prentice-Hall.

Stevenson, Suzanne, Afsaneh Fazly, and Ryan North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, 1–8, Barcelona, Spain.

Stvan, Laurel Smith, 1998. *The Semantics and Pragmatics of Bare Singular Noun Phrases*. Northwestern University dissertation.

Svenonius, Peter, 1994. *Dependent nexus. Subordinate predication structures in English and the Scandinavian languages*. University of California at Santa Cruz dissertation.

Thanopoulos, Aristomenis, Nikos Fakotakis, and George Kokkinakis. 2003. Identification of multiwords as preprocessing for automatic extraction of lexical similarities. In *Proceedings of 6th International Conference on Text, Speech and Dialogue*, 8–11, Ceske Budejovice, Czech Republic.

Tschichold, Cornelia, 1998. *Multi-word Units in Natural Language Processing*. University of Basel dissertation.

Turney, Peter D. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of 9th International Joint Conference on Aritificial Intelligence (IJCAI-2005)*, 1136–1141, Edinburgh, Scotland.

Uchiyama, Kiyoko, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech and Language, Special Issue on Multiword Expressions* 19.497–512.

Utsuro, Takehito, Takao Shime, Masatoshi Tsuchiya, Suguru Matsuyoshi, and Satoshi Sato. 2007. Learning dependency relations of Japanese compound functional expressions. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, 65–72, Prague, Czech Republic.

Van de Curys, Tim, and Begona Villada Moiron. 2007. Semantics-based multiword expression extraction. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, 25–32, Prague, Czech Republic.

Van Der Beek, Leonoor, 2005. *Topics in Corpus-Based Dutch Syntax*. University of Rijksuniversiteit Groningen dissertation.

Vanderwende, Lucy. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational linguistics*, 782–788, Kyoto, Japan.

Venkatapathy, Sriram, and Aravind Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *In the Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, 899–906, Vancouver, Canada.

——, and ——. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the Coling/ACL : Multiword Expressions: Identifying and Exploiting Underlying Properties*, 53–60, Sydney, Australia.

Vickrey, David, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, 771–778, Vancouver, Canada.

Villada Moiron, Begona, 2005. *Data-driven Identification of Fixed Expressions and Their Modifiability*. University of Rijksuniversiteit Groningen dissertation.

Villavicencio, Aline. 2003a. Verb-particle constructions and lexical resources. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 57–64, Sapporo, Japan.

——. 2003b. Verb-particle constructions in world wide web. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in Computational Linguistics Formalisms and Applications*, Toulouse, France.

——. 2005. The availability of verb-particle constructions in lexical resources:how much is enough? *Computer Speech and Language, Special Issue on Multiword Expressions* 19.415–432.

——, Timothy Baldwin, and Benjamin Waldron. 2004. A multilingual database of idioms. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, 1127–1130, Lisbon, Portugal.

WACHOLDER, NINA, and PENG SONG. 2003. Toward a task-based gold standard for evaluation of NP chunks and technical terms. In *Proceedings of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT/NAACL-2003)*, 189–196, Edmonton, Canada.

WARREN, BEATRICE, 1978. *Semantic Patterns of Noun-Noun Compounds*. Actr Universitatis Gothoburgensis dissertation.

WIDDOWS, DOMINIC. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, 276–83, Edmonton, Canada.

——, and BEATE DOROW. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of ACL2005 Workshop on Deep Lexical Axquisition*, 48–56, Ann Arbor, USA.

WIERZBICKA, ANNA. 1982. Why can you have a drink when you can't *have an eat? *Language* 58.753–799.

WOOD, FREDERICK T. 1964. *English Verbal Idioms*. London, UK: Macmillan.

WU, ZHIBIAO, and MARTHA PALMER. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-1994)*, 133–138, Las Cruces, New Mexico, USA.

YAROWSKY, DAVID. 1993. One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop*, 266–271, Plainsboro, New Jerey, USA.

——. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics (ACL-1995)*, 189–196, Cambridge, USA.

YOON, JUNTAE, KEY-SUN CHOI, and MANSUK SONG. 2001. A corpus-based approach for Korean nominal compound analysis based on linguistic and statistical information. *Natural Language Engineering* 7.251–270.

ZHAO, JINGLEI, HUI LIU, and RUZHAN LU. 2007. Semantic labeling of compound nominalization in Chinese. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, 73–80, Prague, Czech Republic.

Appendix A

Abbreviations

CN compound noun

D-PP determinerless prepositional phrase

HMM hidden Markov model

IR information retrieval

LSA latent semantic analysis

LVC light-verb construction

MWE multiword expression

MT machine translation

NLP natural language processing

NC noun compound

POS part-of-speech

PP prepositional phrase

PV prepositional verb

QA question-answering

SR semantic relation

VPC verb-particle construction

Verb-PP or V-PP verb–prepositional phrase

WSD word sense disambiguation

Appendix B

Semantic Relations from SemEval-2007

Semantic relation	Definition & Examples
Cause-Effect (**CE**)	N_1 is the cause of N_2 *virus flu, hormone growth, inhalation death*
Instrument-Agency (**IA**)	N_1 is the instrument of N_2, N_2 uses N_1 *laser printer, ax murderer, sump pump drainage*
Product-Producer (**PP**)	N_1 is a product of N_2, N_2 produces N_1 *honey bee, music clock, supercomputer business*
Origin-Entity (**OE**)	N_1 is the origin of N_2 *bacon grease, desert storm, peanut butter*
Theme-Tool (**TT**)	N_2 is intended for N_1 *reorganization process, copyright law, work force*
Part-Whole (**PW**)	N_1 is part of N_2 *table leg, daisy flower, tree forest*
Content-Container (**CC**)	N_1 is store or carried inside N_2 *apple basket, wine bottle, plane carge*

Table B.1: The set of 7 semantic relations from SemEval-2007, where N_1 is the head noun and N_2 is a modifier

Appendix C

Results from SemEval-2007

Team	Precison	Recall	Fscore	Accuracy
Team759	66.1	66.7	64.8	66.0
Team281	60.5	69.5	63.8	63.5
Team633	62.7	63.0	62.7	65.4
Team161	56.1	57.1	55.9	58.8
Team538	48.2	40.3	43.1	49.9

Table C.1: Group A: WordNet = NO & Query = NO

Team	Precison	Recall	Fscore	Accuracy
Team901	79.7	69.8	72.4	76.3
Team777	70.9	73.4	71.8	72.9
Team281	72.8	70.6	71.5	73.2
Team129	69.9	64.6	66.8	71.4
Team333	62.0	71.7	65.4	67.0
Team538	66.7	62.8	64.3	67.2
Team571	55.7	66.7	60.4	59.1
Team759	66.4	58.1	60.3	63.6
Team220-A	61.7	56.8	58.7	62.5
Team220-B	61.5	55.7	57.8	62.7
Team371	56.8	56.3	56.1	57.7
Team495	55.9	57.8	51.4	53.7

Table C.2: Group B: WordNet = YES & Query = NO

Team	Precison	Recall	Fscore	Accuracy
Team633	64.2	66.5	65.1	67.0
Team759	66.1	66.7	64.8	66.0
Team538	49.4	43.9	45.3	50.1

Table C.3: Group C: WordNet = NO & Query = YES

Team	Precison	Recall	Fscore	Accuracy
Team401	67.3	65.3	62.6	67.2
Team759	66.4	58.1	60.3	63.6
Team538	60.9	57.8	58.8	62.3

Table C.4: Group D: WordNet = YES & Query = YES

Appendix D

Results of all Experiments to Model the Compositionality of English Verb-Particle Constructions

Feature	Semantics	Combination	Precision	Recall	F-score
base	–	yes	.731	–	–
line	–	no	.269	–	–
V	1_{st}	yes	.717	.585	.651
		no	.247	.370	.309
V	N_{th}	yes	.743	.728	.736
		no	.298	.315	.307
VP	1_{st}	yes	.743	.687	.715
		no	.292	.352	.322
	N_{th}	yes	**.737**	**.742**	**.739**
		no	.283	.278	.280
VPS	1_{st}	yes	.719	.626	.672
		no	.247	.333	.290
	N_{th}	yes	.721	.721	.721
		no	.241	.241	.241

Table D.1: Results of Experiment 1 (E1) with TiMBL

Feature	Semantics	Combination	Precision	Recall	F-score
base	–	yes	.731	–	–
line	–	no	.269	–	–
V	1_{st}	yes	.723	.816	.770
		no	.235	.148	.192
V	N_{th}	yes	.741	.837	.789
		no	.335	.204	.264
VP	1_{st}	yes	.726	.864	.795
		no	.240	.111	.176
	N_{th}	yes	.738	.864	.801
		no	.321	.167	.244
VPS	1_{st}	yes	.723	.905	.814
		no	.188	.056	.122
	N_{th}	yes	**.751**	**.905**	**.823**
		no	**.435**	**.185**	**.310**

Table D.2: Results of Experiment 1 (E1) with Maximum Entropy

Feature	Semantics	Combination	Precision	Recall	F-score
base	C,P	yes	.726	–	–
line		no	.274	–	–
base	1st,Nth	yes	.761	–	–
line		no	.239	–	–
C-C	–	yes	.756	.912	.834
		no	.423	.180	.302
C-S	1_{st}	yes	.773	.869	.821
		no	.310	.188	.249
	N_{th}	yes	.773	.869	.821
		no	.310	.188	.249
	P	yes	.756	.912	.834
		no	.423	.180	.302
S-C	1_{st}	yes	.791	.895	.843
		no	.429	.250	.339
	N_{th}	yes	.790	.888	.839
		no	.414	.250	.332
	P	yes	.761	.882	.822
		no	.412	.230	.321
S-S	1_{st}	yes	.720	.924	.822
		no	.214	.055	.134
	N_{th}	yes	.710	.897	.804
		no	.118	.036	.077
	P	yes	.761	.882	.822
		no	.412	.230	.321

Table D.3: Results of Experiment 2 (E2) with TiMBL

Feature	Semantics	Combination	Precision	Recall	F-score
base	C,P	yes	.726	–	–
line		no	.274	–	–
base	1st,Nth	yes	.761	–	–
line		no	.239	–	–
C-C	–	yes	.748	.977	.862
		no	.500	.067	.283
C-S	1_{st}	yes	**.759**	**.993**	**.876**
		no	.000	.000	.000
	N_{th}	yes	**.759**	**.993**	**.876**
		no	.000	.000	.000
S-C	1_{st}	yes	.766	.947	.857
		no	.333	.083	.208
	N_{th}	yes	.765	.941	.853
		no	.308	.083	.196
S-S	1_{st}	yes	.746	.931	.838
		no	.474	.164	.319
	N_{th}	yes	.743	.938	.841
		no	.471	.146	.308

Table D.4: Results of Experiment 2 (E2) with Maximum Entropy

Feature	Semantics	Combination	Precision	Recall	F-score
base	–	yes	.749	–	–
line	–	no	.251	–	–
V	1_{st}	yes	**.771**	**.864**	**.818**
		no	**.366**	**.234**	**.300**
	N_{th}	yes	.768	.796	.782
		no	.316	.281	.299
VP	1_{st}	yes	.770	.859	.814
		no	.357	.234	.296
	N_{th}	yes	.769	.801	.785
		no	.321	.281	.301
VPS	1_{st}	yes	.761	.817	.789
		no	.300	.234	.267
	N_{th}	yes	.768	.796	.782
		no	.316	.281	.299

Table D.5: Results of Experiment 3 (E3) with TiMBL

Feature	Semantics	Combination	Precision	Recall	F-score
base	–	yes	.749	–	–
line	–	no	.251	–	–
V	1_{st}	yes	.778	.843	.810
		no	.375	.281	.328
	N_{th}	yes	.765	.869	.817
		no	.342	.203	.273
VP	1_{st}	yes	.779	.869	.824
		no	.405	.266	.335
	N_{th}	yes	.755	.869	.811
		no	.286	.156	.221
VPS	1_{st}	yes	**.773**	**.895**	**.836**
		no	**.429**	**.234**	**.332**
	N_{th}	yes	.757	.880	.818
		no	.303	.156	.230

Table D.6: Results of Experiment 3 (E3) with Maximum Entropy

Feature	Semantics	Combination	Precision	Recall	F-score
base	–	yes	.598	–	–
line	–	no	.402	–	–
V	1_{st}	yes	**.632**	**.948**	**.759**
		no	.700	.180	.286
	N_{th}	yes	.644	.810	.718
		no	.542	.333	.413
VP	1_{st}	yes	.638	.879	.739
		no	**.588**	**.256**	**.357**
	N_{th}	yes	.632	.828	.716
		no	.524	.282	.367
VPS	1_{st}	yes	.613	.793	.692
		no	.455	.256	.328
	N_{th}	yes	.623	.845	.721
		no	.526	.257	.345

Table D.7: Results of Experiment 4 (E4) with TiMBL

Feature	Semantics	Combination	Precision	Recall	F-score
base	–	yes	.598	–	–
line	–	no	.402	–	–
V	1_{st}	yes	.613	.983	.798
		no	.750	.077	.414
	N_{th}	yes	.606	.983	.795
		no	.667	.051	.360
VP	1_{st}	yes	.596	.966	.781
		no	.333	.026	.180
	N_{th}	yes	**.604**	**1.00**	**.802**
		no	**1.00**	**.026**	**.513**
VPS	1_{st}	yes	**.604**	**1.00**	**.802**
		no	**1.00**	**.026**	**.513**
	N_{th}	yes	**.604**	**1.00**	**.802**
		no	**1.00**	**.026**	**.513**

Table D.8: Results of Experiment 4 (E4) with Maximum Entropy