

SEMI-AUTOMATIC RECOGNITION OF SEMANTIC RELATIONSHIPS IN ENGLISH TECHNICAL TEXTS

Ken Barker

Dissertation
submitted to the School of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

June, 1998

Ottawa-Carleton Institute for Computer Science
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada

© Ken Barker, 1998

ABSTRACT

When people read a text, they rely on *a priori* knowledge of language, common sense knowledge and knowledge of the domain. Many natural language processing systems implement this human model of language understanding, and therefore are heavily knowledge-dependent. Such systems assume the availability of large amounts of background knowledge coded in advance in a specialized formalism. The problem with such an assumption is that building a knowledge base with sufficient and relevant content is labour-intensive and very costly. And often, the resulting knowledge is either too specific to be used for more than one very narrow domain or too general to allow subtle analyses of texts.

In order to avoid the problems of manually encoding background knowledge, many researchers have abandoned symbolic language analysis in favour of statistical methods. The availability of large online corpora and improvements in computing resources have made it possible to make predictions about meaning based on observations of frequencies, contexts, correlation, and other phenomena in a corpus. Systems that use statistical methods have had some impressive successes, notably in part of speech tagging, word class clustering and word sense disambiguation. But these systems often require large amounts of analyzed language data to arrive at even shallow interpretations.

Both of these kinds of natural language processing systems seek models of a text—knowledge-intensive systems a deep semantic model, corpus-based systems a much shallower distributional one. And both kinds of system depend on outside sources of data. This dissertation describes the construction and evaluation of an interactive tool that also seeks a model of a text. The model takes the form of semantic relationships between syntactic elements in English sentences. The system also depends on an outside source of data: a cooperative user. Unlike knowledge-intensive and corpus-based systems, however, it does not require a large repository of semantic information and it does not require any previously analyzed data: it can start processing a text from scratch. The system inspects the surface syntax of a sentence to make informed decisions about its possible interpretations. It then suggests these interpretations to the user. As more text is analyzed, the system learns from previous analyses to make better decisions, reducing its reliance on the user. Evaluation confirms that the semi-automatic acquisition of the model of a text is relatively painless for the user.

The regular structure of the model identifies concepts that have different surface-syntactic forms. These concepts could be used as the knowledge base for expert systems or query answering systems. They could be used as a conceptual profile of a text, allowing, for example, text indexing on semantic concepts instead of just keywords. The concepts and semantic relationships between them could serve as base structures for text summarization. They could also be used as the domain-specific background knowledge core for natural language processing systems that attempt deeper understanding of a text.

ISTHMI

Every man is a peninsula.

I would like to acknowledge here all the people who contributed in some way to the development of this dissertation. But I can't. There isn't nearly enough space. Those who stopped to talk to me about my research when I needed inspiration, those who stopped talking to me about my research when I needed ventilation, those who shared sounds or words or ice time, those who with me supported my local brewery can be assured that I appreciate them, every one.

Then there are those whose contribution cannot be dismissed with cute coordination. Stan Szpakowicz gave me guidance in every facet of my academic life. His far-sighted direction is what led me here, and I'm grateful.

I would like to acknowledge my thesis committee: Jean-Pierre Corriveau, Stan Matwin, Ingrid Meyer and Fred Popowich each read and understood the dissertation thoroughly, attacking it on different levels from different angles. Their probing gave me a better understanding of my own thesis.

The "TANKA core" is my immediate research family. Sylvain Delisle's dedication helped make my work workable, and then he made it fun. Terry Copeck's talent and attention to a million things kept TANKA running, in spite of obstacles both man-made and natural.

Sylvia Boyd, Rob Holte and Stan Matwin deserve my gratitude for all the times I saw them go out of their way to open doors for me—and more for all the times I didn't.

My parents, Bob and Betty Jo, supported me in the ways that parents usually get acknowledgment for. That they supported me as an equal was unique. They have earned my thanks and admiration, and have shown me a friendship that rarely transcends generations. My brothers Thomas and Stephen and my sister Louise provided entertainment, sport, humour, warmth and really good food. If you could choose your family, you would choose mine.

Johanne Morin. She knows my algorithms, she knows my heart, she knows I don't like to shoot sides, she knows how to make me happy. She is beauty, intelligence, taste and mirth. If she could play nets, she would be perfect.

This research was supported by the Natural Science and Engineering Research Council of Canada.

CONTENTS—GENERAL

1	Overview of the Project	1
1.1	Automated Semantic Analysis.....	2
1.2	Goals of the Project.....	3
1.3	Semantic Analysis in TANKA.....	4
1.4	Evaluation.....	6
1.5	Applications.....	10
1.6	Organization of the Dissertation.....	10
2	Clause Level Relationship Analysis	15
2.1	Introduction.....	15
2.2	Input Structures.....	19
2.3	The Clause Level Relationships.....	20
2.4	The CLR Marker Dictionary.....	24
2.5	Assigning CLRs.....	25
2.6	Evaluation.....	39
2.7	An Example.....	43
2.8	Chapter Summary.....	46
3	Case Relationships	51
3.1	Introduction.....	52
3.2	Case Markers.....	58
3.3	Case System Design.....	60
3.4	The Cases.....	65
3.5	Assigning Cases.....	73
3.6	Evaluation.....	74
3.7	Chapter Summary.....	81
4	Noun Modifier Relationship Analysis	85
4.1	Introduction.....	85
4.2	Input Structures.....	92
4.3	Noun Modifier Bracketing.....	94
4.4	The Noun Modifier Relationships.....	105
4.5	The NMR Marker Dictionary.....	106
4.6	Assigning NMRs.....	108
4.7	Evaluation.....	116
4.8	An Example.....	126
4.9	Chapter Summary.....	131

5	Future Work	135
5.1	Clause Level Relationship Analysis	135
5.2	Case Analysis	137
5.3	Noun Modifier Relationship Analysis	139
5.4	A Unified Set of Semantic Relationships	141
5.5	Other Directions Altogether	142
6	Conclusions	147
6.1	Summary	147
6.2	Goals Revisited	148
6.3	Closing Words	152
7	References	153
Appendix I: CLR Marker Dictionary		161
Appendix II: Case Marker Dictionary		163
Appendix III: NMR Marker Dictionary		165

CONTENTS—DETAILED

1	Overview of the Project	1
<hr/>		
1.1	Automated Semantic Analysis	2
1.1.1	Oracles.....	2
1.1.2	Entities and Events.....	2
1.2	Goals of the Project	3
1.3	Semantic Analysis in TANKA	4
1.3.1	DIPETT Parse Trees.....	4
1.3.2	Three-Tiered Semantic Analysis.....	4
	Clause Level Relationships.....	4
	Case Relationships.....	5
	Noun Modifier Relationships.....	5
1.3.3	Background Knowledge.....	5
	Linguistic Knowledge.....	5
	Semantic Knowledge.....	6
	User Knowledge.....	6
1.4	Evaluation	6
1.4.1	Test Texts.....	8
1.4.2	Parser Evaluation.....	8
1.4.3	User Actions.....	9
1.4.4	User Burden.....	9
1.5	Applications	10
1.6	Organization of the Dissertation	10
1.6.1	Paper Map.....	12
2	Clause Level Relationship Analysis	15
<hr/>		
2.1	Introduction	15
2.1.1	Semantic Relationships.....	16
2.1.2	Tense and Modality.....	18
2.2	Input Structures	19
2.2.1	Verb Sequence Features.....	19
2.2.2	Clausal Organization.....	19
	Coordination or Subordination.....	19
	Correlative Coordination.....	20
	Subordinator/Conjunct Correlation.....	20
2.3	The Clause Level Relationships	20
2.3.1	CLR Glossary.....	22
	Causal CLRs.....	22
	Temporal CLRs.....	23
	Conjunctive CLRs.....	23
2.4	The CLR Marker Dictionary	24
2.5	Assigning CLRs	25
2.5.1	Verb Phrase Polarity and Connective Polarity.....	25

2.5.2	Verb Phrase Tense and Modality	27
	Absence of Modals.....	28
	Modals in Negative Verb Phrases	28
	Marginal Modals	29
	Should as a Past Tense of Shall.....	29
2.5.3	Using Argument Features to Choose CLR's	30
	CLR Preference Rules	31
2.5.4	CLR Competitions.....	35
2.5.5	Using User Assignments to Choose CLR's.....	36
	CLR Assignment Attributes	36
	When to Use User Assignments.....	37
	Least-General Generalization of Attribute Patterns	38
	Avoiding Overgeneralization	39
2.6	Evaluation	39
2.6.1	Diagnostic Evaluation	39
2.6.2	Performance Evaluation	40
	System Performance.....	40
	Coverage.....	42
2.6.3	Can CLRA Learn?.....	42
2.7	An Example.....	43
2.8	Chapter Summary.....	46
3	Case Relationships	51
<hr/>		
3.1	Introduction	52
3.1.1	Cases.....	52
3.1.2	Case Theory.....	53
3.1.3	Valency Theory	53
3.1.4	Other Case Systems.....	54
3.2	Case Markers.....	58
3.2.1	Positional Markers.....	58
3.2.2	Prepositional Markers	58
3.2.3	Adverbial Markers.....	59
3.2.4	Marker Order.....	59
3.3	Case System Design.....	60
3.3.1	The Case Marker Dictionary	61
3.3.2	Evaluation Criteria	62
	Generality	62
	Completeness	63
	Uniqueness	64
3.3.3	Using the Criteria to Guide Case Selection.....	64
3.4	The Cases	65
3.4.1	Case Glossary.....	65
	Participant.....	66
	Causality.....	69
	Space	70

	Time	71
	Quality	72
3.5	Assigning Cases	73
3.6	Evaluation	74
3.6.1	Case Analyzer Evaluation	74
3.6.2	Case System Evaluation	76
	Generality	76
	Completeness	79
	Uniqueness	79
3.7	Chapter Summary	81
4	Noun Modifier Relationship Analysis	85
4.1	Introduction	85
4.1.1	Noun Compounds.....	86
4.1.2	Semantic Relations in Noun Phrases.....	87
4.1.3	Recognizing Semantic Relations.....	89
4.2	Input Structures	92
4.2.1	Attributes and Head Noun Premodifiers	92
4.2.2	Noun Phrase Postmodifiers	93
	Relative Clauses	93
	Appositives.....	93
	Comparison	94
	Prepositional Phrases.....	94
4.3	Noun Modifier Bracketing	94
4.3.1	Subphrases and Reduced Subbracketings	95
4.3.2	Bracketing Models	96
4.3.3	A Bracketing Algorithm.....	96
4.3.4	Confidence in Branching Decisions.....	98
4.3.5	User Interaction	98
4.3.6	When Is “Left-Branching” “Not-Right-Branching”?	99
4.3.7	A Walk through Bracketing	100
4.3.8	When All Else Fails.....	102
	Redoing the Bracketing Interaction.....	102
	Ignoring Bracketing History.....	102
	Bracketing by Hand.....	103
4.4	The Noun Modifier Relationships.....	105
4.4.1	NMR Glossary.....	106
4.5	The NMR Marker Dictionary	106
4.6	Assigning NMRs	108
4.6.1	Reduced Modifiers and Heads	108
4.6.2	Modifier-Head-Marker Triples.....	109
4.6.3	Distance between Triples	110
4.6.4	The Best NMRs.....	111
	Absolute Frequency.....	112
	Relative Frequency.....	112

	Weighted Relative Frequency	112
4.6.5	User Interaction	113
4.6.6	Classifying Function of Premodifiers.....	115
4.7	Evaluation	116
4.7.1	Bracketer Evaluation	116
	System Performance.....	116
	The Effect of the Threshold.....	117
	Branching Frequencies	119
4.7.2	NMRA Evaluation.....	119
	System Performance.....	119
	Coverage.....	122
	User Burden.....	122
4.7.3	Some Difficulties.....	122
	Questionably Endocentric Compounds	122
	Postpositive Adjectives	124
	Adjectives in Paraphrases.....	126
4.8	An Example.....	126
4.9	Chapter Summary.....	131
5	Future Work	135
5.1	Clause Level Relationship Analysis.....	135
5.1.1	Evaluation.....	135
5.1.2	Generalizing User Assignments	136
5.1.3	Embedded CLR Structures	136
5.1.4	Are Attribute Patterns Empirical Preference Rules?.....	137
5.2	Case Analysis	137
5.2.1	Assigning Cases One by One	137
5.2.2	Aggressive Case Analysis	138
5.2.3	A General Purpose Case Pattern Dictionary	139
5.3	Noun Modifier Relationship Analysis	139
5.3.1	Noun Modifier Bracketing	139
	Storing Pairs from Prepositional Phrases	139
	Branching Frequencies	139
	Insensitivity to the Bracketing Threshold.....	140
5.3.2	Other Noun Modifier Relationships	140
5.3.3	Methods for Choosing the Best NMRs	140
5.3.4	Taxonomic Information from NMRs	141
5.4	A Unified Set of Semantic Relationships.....	141
5.5	Other Directions Altogether	142
5.5.1	Fully Automatic Recognition of Semantic Relationships	142
5.5.2	Extending HAIKU with WordNet	142
5.5.3	Other Sources of Syntactic Information	143

6	Conclusions	147
6.1	Summary	147
6.1.1	Clause Level Relationship Analysis	147
6.1.2	Case Relationships	148
6.1.3	Noun Modifier Relationship Analysis	148
6.2	Goals Revisited	148
6.2.1	Semantic Relationships	149
6.2.2	Semi-Automatic Recognition of Semantic Relationships	149
6.2.3	Learning	149
6.2.4	Coverage	151
6.2.5	User Burden	151
6.3	Closing Words	152
7	References	153
Appendix I: CLR Marker Dictionary		161
Appendix II: Case Marker Dictionary		163
Appendix III: NMR Marker Dictionary		165

FIGURES

Figure 1: CLRA interaction for (81)	44
Figure 2: CLRA interaction for (82)	45
Figure 3: CLRA interaction for (83)	46
Figure 4: CP assignments in the clouds experiment	75
Figure 5: CP assignments in the small engines experiment	75
Figure 6: Distribution of the cases in the clouds and small engines experiments, and among the markers.....	77
Figure 7: Branching for (187).....	99
Figure 8: User interaction for phrase (190)	100
Figure 9: An incorrect bracketing repaired by turning history off	103
Figure 10: Manual bracketing interaction for (196)	104
Figure 11: The <code>list</code> feature for example (220).....	114
Figure 12: The <code>create</code> feature for example (221).....	115
Figure 13: Bracketing decisions in the <code>sparc</code> experiment.....	117
Figure 14: Bracketing decisions in the small engines experiment	117
Figure 15: The effect of different threshold values on branching decisions for the <code>sparc</code> experiment	118
Figure 16: NMR assignments in the <code>sparc</code> experiment	120
Figure 17: NMR assignments in the small engines experiment.....	120
Figure 18: Bracketing and NMRA for a postpositive adjective	125
Figure 19: NMRA interaction for (236).....	127
Figure 20: NMRA interaction for (237)	129
Figure 21: NMRA interaction for (238).....	130
Figure 22: NMRA interaction for (239).....	131

TABLES

Table 1: Discourse relations from Schiffrin (1987)	17
Table 2: Coherence relations from Dahlgren (1988).....	17
Table 3: Explanatory connections from Gomez (1995)	18
Table 4: The clause level relationships (with abbreviations).....	21
Table 5: Strength of the modals	28
Table 6: Pairs of CLRs that share no markers in the CLR marker dictionary.....	30
Table 7: Pairs of CLRs that cannot be disambiguated using verb phrase features.....	31
Table 8: Outcomes of individual CLR competitions for (75)	36
Table 9: Points accumulated by each CLR in competitions for (75)	36
Table 10: CLRA user actions required in the building code experiment.....	40
Table 11: CLRA user actions required in the clouds experiment	41
Table 12: CLRA user actions required in the small engines experiment.....	42
Table 13: CLRA user actions required with partial matching on stored assignments	43
Table 14: List of cases from Fillmore (1968).....	54
Table 15: List of cases from Larson (1984)	54
Table 16: The case grid from Somers (1987).....	55
Table 17: Propositional and modal cases from Cook (1989).....	56
Table 18: Macroroles and thematic relations from van Valin (1993)	57
Table 19: The cases (with abbreviations).....	66
Table 20: Recoverably deletable predicates from Levi (1978)	88
Table 21: Noun-noun semantic classes from Warren (1978).....	89
Table 22: Adjective-noun semantic classes from Warren (1984)	89
Table 23: Noun-noun semantic relationships from Leonard (1984)	90
Table 24: Semantic features extracted from dictionaries in Vanderwende (1993)	90
Table 25: Semantic relations between nouns in Vanderwende (1993)	91
Table 26: The noun modifier relationships (with abbreviations).....	105
Table 27: NMRs with paraphrases and examples	107
Table 28: (M, H, Mk) triples for (211), (212) and (213).....	110
Table 29: Measures of distance between triples.....	111
Table 30: Best NMR for (217) using the absolute frequency method.....	112
Table 31: Best NMR for (217) using the relative frequency method.....	112
Table 32: Best NMR for (217) using weighted relative frequency scores	113
Table 33: Branching frequencies for the small engines text	119
Table 34: Branching frequencies for the sparc text.....	119
Table 35: Applying different “best NMR” methods to sparc.....	121
Table 36: Twelve case patterns for psubj-pobj.....	138
Table 37: Onus ratings for the clouds and small engines experiments	151