

Empirical Methods for Exploiting Parallel Texts

I. Dan Melamed

**The MIT Press
Cambridge, Massachusetts
London, England**

Contents

	Acknowledgments	xi
1	Introduction	1
I	TRANSLATIONAL EQUIVALENCE AMONG WORD TOKENS	5
2	A Geometric Approach to Mapping Bitext Correspondence	7
2.1	Introduction	7
2.2	Bitext Geometry	8
2.3	Previous Work	9
2.4	The Smooth Injective Map Recognizer (SIMR)	13
2.4.1	Overview	13
2.4.2	Point Generation	14
2.4.3	Noise Filter	17
2.4.4	Point Selection	18
2.4.5	Reduction of the Search Space	19
2.4.6	Enhancements	21
2.5	Parameter Optimization	23
2.6	Evaluation	24
2.7	Implementation of SIMR for New Language Pairs	30
2.7.1	Step 1: Construct Matching Predicate	30
2.7.2	Step 2: Construct Axis Generators	31
2.7.3	Step 3: Reoptimize Parameters	32
2.8	Conclusion	33
3	Application: Alignment	35
3.1	Introduction	35
3.2	Correspondence is Richer Than Alignment	35
3.3	The Geometric Segment Alignment (GSA) Algorithm	37
3.4	Evaluation	38
3.5	Conclusion	40
4	Application: Automatic Detection of Omissions in Translations	41
4.1	Introduction	41
4.2	The Basic Method	41
4.3	Noise-Free Bitext Maps	43
4.4	A Translator's Tool	44

4.5	Noisy Bitext Maps	45
4.6	ADOMIT	46
4.7	Simulation of Omissions	48
4.8	Evaluation	49
4.9	Conclusion	53
II	THE TYPE-TOKEN INTERFACE	55
5	Models of Co-occurrence	57
5.1	Introduction	57
5.2	Relevant Regions of the Bitext Space	58
5.3	Co-occurrence Counting Methods	59
5.4	Language-Specific Filters	62
5.5	Conclusion	63
6	Manual Annotation of Translational Equivalence	65
6.1	Introduction	65
6.2	The Gold-Standard Bitext	66
6.3	The Blinker Annotation Tool	68
6.4	Methods for Increasing Reliability	69
6.5	Inter-Annotator Agreement	72
6.6	Conclusion	77
III	TRANSLATIONAL EQUIVALENCE AMONG WORD TYPES	79
7	Word-to-Word Models of Translational Equivalence	81
7.1	Introduction	81
7.2	Translation Model Decomposition	82
7.3	The One-to-One Assumption	86
7.4	Previous Work	87
7.4.1	Non-Probabilistic Translation Lexicons	87
7.4.2	Re-estimated Sequence-to-Sequence Translation Models	89
7.4.3	Re-estimated Bag-to-Bag Translation Models	93
7.5	Parameter Estimation	94
7.5.1	Method A: The Competitive Linking Algorithm	97

7.5.2	Method B: Improved Estimation Using an Explicit Noise Model	99
7.5.3	Method C: Improved Estimation Using Pre-Existing Word Classes	103
7.6	Effects of Sparse Data	104
7.7	Evaluation	107
7.7.1	Evaluation at the Token Level	107
7.7.2	Evaluation at the Type Level	114
7.8	Application to MT Lexicon Development	119
7.9	Conclusion	121
8	Automatic Discovery of Non-Compositional Compounds	123
8.1	Introduction	123
8.2	Objective Functions	124
8.3	Search	126
8.4	Predictive Value Functions	127
8.5	Iteration	128
8.6	Credit Estimation	131
8.7	Single-Best Translation	134
8.8	Experiments	135
8.9	Related Work	143
8.10	Conclusion	144
9	Sense-to-Sense Models of Translational Equivalence	147
9.1	Introduction	147
9.2	Previous Work	148
9.3	Formulation of the Problem	150
9.4	Noise Filters	151
9.5	The SenseClusters Algorithm	152
9.6	An Application	154
9.7	Experiments	155
9.7.1	Quantitative Results	156
9.7.2	Qualitative Results	160
9.8	Conclusion	162

10	Summary and Outlook	165
A	Annotation Style Guide for the Blinker Project	169
A.1	General Guidelines	169
A.1.1	Omissions in Translation	170
A.1.2	Phrasal Correspondence	171
A.2	Detailed Guidelines	173
A.2.1	Idioms and Near Idioms	173
A.2.2	Referring Expressions	175
A.2.3	Verbs	177
A.2.4	Prepositions	178
A.2.5	Determiners	180
A.2.6	Punctuation	180
	Notes	183
	References	187
	Index	193