

ABSTRACT

Title of Dissertation: WORD SENSE DISAMBIGUATION
 WITHIN A MULTILINGUAL FRAMEWORK

Mona Talat Diab, Doctor of Philosophy, 2003

Dissertation directed by: Professor Philip Resnik
 Department of Linguistics & UMIACS

Word Sense Disambiguation (WSD) is the process of resolving the meaning of a word unambiguously in a given natural language context. Within the scope of this thesis, it is the process of marking text with explicit sense labels.

What constitutes a sense is a subject of great debate. An appealing perspective, aims to define senses in terms of their multilingual correspondences, an idea explored by several researchers, Dyvik (1998), Ide (1999), Resnik & Yarowsky (1999), and Chugur, Gonzalo & Verdejo (2002) but to date it has not been given any practical demonstration. This thesis is an empirical validation of these ideas of characterizing word meaning using cross-linguistic correspondences. The idea is that word meaning or word sense is quantifiable as much as it is uniquely translated in some language or set of languages.

Consequently, we address the problem of WSD from a multilingual perspective; we expand the notion of context to encompass multilingual evidence. We devise a new approach to resolve word sense ambiguity in natural language, using a source of information that was never exploited on a large scale for WSD before.

The core of the work presented builds on exploiting word correspondences across languages for sense distinction. In essence, it is a practical and functional implementation of a basic idea common to research interest in defining word meanings in cross-linguistic terms.

We devise an algorithm, SALAAM for Sense Assignment Leveraging Alignment And Multilinguality, that empirically investigates the feasibility and the validity of utilizing translations for WSD. SALAAM is an unsupervised approach for word sense

tagging of large amounts of text given a parallel corpus — texts in translation — and a sense inventory for one of the languages in the corpus. Using SALAAM, we obtain large amounts of sense annotated data in both languages of the parallel corpus, simultaneously. The quality of the tagging is rigorously evaluated for both languages of the corpora.

The automatic unsupervised tagged data produced by SALAAM is further utilized to bootstrap a supervised learning WSD system, in essence, combining supervised and unsupervised approaches in an intelligent way to alleviate the resources acquisition bottleneck for supervised methods. Essentially, SALAAM is extended as an unsupervised approach for WSD within a learning framework; in many of the cases of the words disambiguated, SALAAM coupled with the machine learning system rivals the performance of a canonical supervised WSD system that relies on human tagged data for training.

Realizing the fundamental role of similarity for SALAAM, we investigate different dimensions of semantic similarity as it applies to verbs since they are relatively more complex than nouns, which are the focus of the previous evaluations. We design a human judgment experiment to obtain human ratings on verbs' semantic similarity. The obtained human ratings are cast as a reference point for comparing different automated similarity measures that crucially rely on various sources of information. Finally, a cognitively salient model integrating human judgments in SALAAM is proposed as a means of improving its performance on sense disambiguation for verbs in particular and other word types in general.

WORD SENSE DISAMBIGUATION
WITHIN A MULTILINGUAL FRAMEWORK

by

Mona Talat Diab

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland at College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2003

Advisory Committee:

Professor Philip Resnik, Chairperson/Advisor
Professor Bonnie Dorr
Professor Paul Pietroski
Professor Amy Weinberg
Professor David Yarowsky

© Copyright by
Mona Talat Diab
2003

DEDICATION

بسم الله الرحمن الرحيم
قل رب زدني علماً

لذكرى أبي الحبيب رحمه الله

In loving memory of my most beloved and most cherished father, Dr. Talat Diab (June 13, 1941 - January 27, 2003), who lived a life of dignity, courage, wisdom, patience and above all affection, and who will remain my personal hero and my inspiration forever. May God bless his soul,
Amen.

ACKNOWLEDGEMENTS

This thesis would not have materialized without the help and support of many people. I believe I am very blessed to be surrounded by such needed encouragement. So here it goes: I would like to start with acknowledging my parents and my brother Hatim's love, trust, guidance, support and confidence in me throughout the years of my studies and research; without them, I doubt that I would have achieved what I have today, thanks for believing in me. I would like to express my deepest gratitude to Philip Resnik, my research advisor, for his understanding and his support through all the good times and especially through the bad times. He was always there with great advice and a listening ear. I am grateful to the committee members on my thesis, Bonnie Dorr, Amy Weinberg, Paul Pietrosky and David Yarowsky, for all their insightful comments and remarks about this research.

I would also like to acknowledge the constant encouragement received from Doug Oard and Mari Olsen. I thank Peter Bock for grounding me in the scientific method. I would like to thank Franz Och for GIZA++, without which I am not sure this research would have been as feasible. I am very grateful to Julio Gonzalo and Irina Chugur for their help with Spanish data. I would like to acknowledge the support afforded me by Thierry Paquet from the University of Rouen during a very tough period of my life and this thesis. I would like to thank Ted Pedersen, Rada Mihalcea and Sid Patwardham for valuable discussions on similarity and WordNet issues.

Thanks a lot clippers for being there for me all the time, special thanks to Nizar Habash, Okan Kolak, David Zajic, Maria Katsova, Clara Cabezas,

Grazia Lassner, Michael Nossal, Rebecca Hwa, Gina Levow, Kareem Darwish, Fazil Ayan, Dina Demner, Adam Lopez and Laura Bright. I would also like to acknowledge the support of Louiqa Rachid, thanks for listening. I am eternally grateful to Mohamed Zahran, Kobi Snitz and Nizar Habash for help with thesis formatting and editing.

Last but not least, my dearest circle of friends who were always encouraging me and showering me with their love and support through all the good times and bad times, may God bless you all: Selda Kapan, Doaa Taha, Muna Yousef, Hannan Tamimi, Hanan Morsy, Khaled El Gindy, Tamer Soliman, Fadwa Attiga, Ayad Sleiman, Margaret Zaknoen, Mandy Chan, Amany El Anshasy, Dahliah Hawary, Mariam Anwarzai, Aslihan Yildiz, Mukul Ghandi, Aydan Kalyoncu, Sevim Kalyoncu, Leila Meshkat, Rania Al Mashat, Edward Balaban, Tamer Nadeem, Ingy Bakir, Heba Zaghloul, Hela Zouari, Tamer Sharnouby, Svend White, Shabana Mir, Jonathan Brown, Haifa Khalafallah, Nabilah Haque, Mike Sanford, Sean Glasheen, Sura AlSaaty, Mazen Bitar, Hazem Bitar, Mustafa Tikir, Burcu Ayan, Betul Attalay, Anuradha Shenoy, Moody Tamimi, and Taras Riopka.

Finally, this research was partly supported by National Science Foundation grant EIA0130422, DARPA/ITO Contract N66001-97-C-8540, DARPA/ITO Cooperative Agreement N660010028910, Department of Defense contract RD-02-5700, and ONR MURI Contract FCPO.810548265.

TABLE OF CONTENTS

List of Tables	ix
List of Figures	xi
1 Thesis Introduction	1
1.1 Introduction	1
1.2 Research Contributions	3
1.3 Thesis Layout & Brief Overview of Chapters	4
2 Related Work	6
2.1 Introduction	6
2.2 Pre-SENSEVAL WSD: Historical Perspective	6
2.3 The SENSEVAL Era	8
2.4 Multilingual Approaches to WSD	10
2.4.1 Word Sense Disambiguation Using Statistical Methods: <i>Brown et al.</i>	10
2.4.2 Using Bilingual Materials to Develop Word Sense Disambiguation Methods: <i>Gale et al.</i>	12
2.4.3 Word Sense Disambiguation Using a Second Language Monolingual Corpus: <i>Ido Dagan & Alon Itai</i>	14
2.4.4 Resolving Translation Ambiguity Using Non-Parallel Bilingual Corpora: <i>Kikui</i>	18
2.4.5 Summary	19
3 Word Sense Tagging Using Parallel Corpora: SALAAM	21
3.1 Introduction	21
3.2 Motivation	22
3.3 Problem Statement	22
3.4 Relevant Background	23
3.4.1 A Translational Basis for Semantics. <i>Helge Dyvik</i>	23
3.4.2 Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. <i>Philip Resnik & David Yarowsky</i>	24

3.4.3	Polysemy and Sense Proximity in the SENSEVAL-2 Test Suite. <i>Irina Chugur, Julio Gonzalo, Felisa Verdejo</i>	26
3.4.4	Cross-lingual Sense Discrimination: Can it work? <i>Nancy Ide</i> .	28
3.4.5	Discussion	29
3.5	Hypothesis	29
3.5.1	General Hypothesis Statement	31
3.6	Method	31
3.6.1	General Method Description	31
3.6.2	Required Resources	32
3.6.3	Detailed Method Description	33
3.6.4	Evaluation Metrics	39
3.7	Evaluation	40
3.7.1	Materials	40
3.7.2	Tools	44
3.7.3	Sense Selection and Similarity Measure	45
3.7.4	Development and Testing Environment	45
3.7.5	Evaluation Measure	46
3.7.6	Evaluation Parameters	46
3.7.7	Evaluation Factors	46
3.7.8	Evaluation Conditions	47
3.7.9	Experimental Hypotheses	49
3.8	Results	50
3.8.1	Testing Hypothesis 1	50
3.8.2	Testing Hypothesis 2	52
3.8.3	Testing Hypothesis 3	54
3.8.4	Testing Hypothesis 4	55
3.8.5	Testing Hypothesis 5	56
3.8.6	Testing Hypothesis 6	57
3.8.7	Overall results	58
3.9	Discussion	58
3.9.1	Summary of the Results	58
3.9.2	Analysis of Results	60
3.9.3	Precision	61
3.9.4	Recall	62
3.9.5	Coverage	63
3.9.6	Complementarity with Other WSD Systems	63
3.9.7	Evaluation of Target Language Tagging	63
3.10	Summary	64
4	Extensions to SALAAM	65
4.1	Introduction	65

4.2	Using Human Translations — Naturally-Occurring Parallel Corpora	65
4.2.1	Introduction	65
4.2.2	Motivation	66
4.2.3	Hypothesis	66
4.2.4	Evaluation	67
4.2.5	Discussion	72
4.2.6	Summary	73
4.3	Target Language Tagging Evaluation	73
4.3.1	Introduction	73
4.3.2	Motivation	73
4.3.3	General Hypothesis	74
4.3.4	Required Resources	74
4.3.5	Projected Sense Tagging on Arabic Data	75
4.3.6	Projected Sense Tagging on Spanish Data	79
4.3.7	General Discussion	89
4.3.8	Summary	89
4.4	Feasibility of bootstrapping a WordNet style ontology for Arabic	90
4.4.1	Introduction	90
4.4.2	Evaluation	91
4.4.3	Levels of representation	93
4.4.4	Summary	95
5	Exploration into Bootstrapping Supervised WSD	96
5.1	Introduction	96
5.2	Motivation	97
5.3	Related Work	98
5.4	Empirical Layout	99
5.5	University of Maryland Supervised Sense Tagging system (UMSST)	100
5.6	Bootstrapping Evaluation	101
5.6.1	Test data	102
5.6.2	Hand-Tagged Training Data	103
5.6.3	Gold Standard	104
5.6.4	SALAAM Training Data Corpora	105
5.6.5	SALAAM-tagged Training Data Creation	106
5.6.6	Experimental Conditions	107
5.6.7	Evaluation Metric	108
5.6.8	General Experimental Hypothesis	108
5.7	Results	109
5.8	Discussion	112
5.8.1	Analysis of factors affecting PR	115
5.9	Combining factors	131

5.10	Summary	132
6	Facets of Similarity	133
6.1	Introduction	133
6.2	Motivation	134
6.3	Models of Verb Similarity	135
6.3.1	Class 1: Taxonomic Models	135
6.3.2	Class 2: Distributional Co-occurrence Model	138
6.3.3	Class 3: Semantic Structure Model	138
6.4	Human Judgment Experiment	140
6.4.1	Participants	141
6.4.2	Materials	141
6.4.3	Conditions	142
6.4.4	Procedure	142
6.4.5	Results	144
6.4.6	Discussion	144
6.5	Application to SALAAM	146
6.5.1	Integrating Human Ratings in SALAAM: A Cognitive Based Feasibility Study	148
6.6	Summary	153
7	Conclusions & Future Directions	154
7.1	Conclusions	154
7.2	Thesis Problems & Limitations	158
7.3	Research Contributions	158
7.4	Future Directions	159
	Bibliography	168

LIST OF TABLES

2.1	Summary of Systems' Required Resources	19
2.2	Summary of Systems' Evaluation	20
3.1	Relative sizes of corpora used for evaluating SALAAM on SV2AW test set	43
3.2	SALAAM performance results on English source SV2AW test data in the default conditions	51
3.3	SALAAM performance with pre-alignment French target pseudo-translation merge	53
3.4	SALAAM performance in default condition vs. intralanguage post- alignment merge condition at MAX sense selection criterion	54
3.5	SALAAM performance in intralanguage post-alignment merge condi- tions with sense selection criterion MAX vs. THRESH	56
3.6	SALAAM performance for conditions 4 where evidence is obtained from monolingual intralanguage pseudo-translation merge vs. evidence obtained from interlanguage pseudo-translation intersection merge in condition 5	57
3.7	SALAAM performance for conditions 4 where evidence is obtained from monolingual intralanguage pseudo-translation merge vs. evidence obtained from interlanguage pseudo-translation union merge in Condi- tion 6	57
4.1	Relative sizes of the English side of corpora used in HT Evaluations .	68
4.2	SALAAM Results on SV2AW for MT & HT parallel corpora indepen- dently	70
4.3	SALAAM results using both HT and MT for augmenting the test corpus	71
4.4	Accuracy results of projected tagging onto Arabic SV2AW data mea- sured against English WN17pre sense definitions	78
4.5	Relative sizes of corpora used in projected Spanish tagging evaluation	81
4.6	Results in % for RBL, and the different evaluation conditions of test set SPSV2AW	85

5.1	Comparative results obtained by Mihalcea’s bootstrapping system when training an instance-based learning supervised WSD system using both human tagged data and GenCor tagged data as training examples . . .	99
5.2	Test items for SV2LS-test	102
5.3	Characteristics of hand-tagged training data for the SENSEVAL2 English Lexical Sample task	103
5.4	Current evaluation gold standard precision results obtained by UMSST-human	106
5.5	SALAAM-tagged training corpora sizes	106
5.6	Precision scores and PR of UMSST-SALAAM and UMSST-human on SV2LS-test	110
5.7	PRs of the best individual conditions using UMSST-SALAAM training data on the top 16 test nouns in Table 5.6: 1 is condition MT+HT+SV2LS-TR_THRESH_SP; 2 is condition HT+SV2LS-TR_THRESH_ML_I; 3 is condition HT+SV2LS-TR_THRESH_ML_U; 4 is condition MT+HT+SV2LS-TR_MAX_ML_I; and 5 is condition HT+SV2LS-TR_MAX_ML_I . .	111
5.8	Precision % scores obtained for SALAAM-SV2LS-TR and UMSST-SALAAM	114
5.9	List of test nouns with their corresponding number of senses and sense contexts in the SALAAM-tagged training data	117
5.10	Test nouns, the corresponding number of senses, perplexity values and PR	119
5.11	Test nouns with their corresponding Semantic Translation Entropy values and performance ratios, PR	122
5.12	Test nouns with their corresponding SDC values and PRs	124
5.13	Test noun items with the absolute difference between SALAAM-tagged perplexity and test data perplexity, PerpDiff, against the performance ratios, PR	126
5.14	Test nouns with SDC, manually grouped similar senses, and performance ratios, PR	129
5.15	Characteristics of the nouns <i>stress</i> and <i>church</i>	130
6.1	Aspectual features determining aspectual class for verbs	141
6.2	The final verb pairs used in the human judgments experiment	143
6.3	Comparing the different automated similarity measures to the two human conditions	144
6.4	Regression Coefficients for the automatic similarity measures	151
6.5	SALAAM performance results with different similarity measure conditions	152
7.1	Summary of Multilingual WSD Systems’ Required Resources	155
7.2	Summary of Multilingual WSD Systems’ Evaluation	156

LIST OF FIGURES

3.1	Common Senses Shared Between Polysemous Words	30
3.2	Flow chart demonstrating process flow in SALAAM method	33
3.3	A sample token alignment in a parallel corpus	34
3.4	Tokens aligned in a parallel corpus	34
3.5	Aligned token instances from target to source	35
3.6	Target word types and their corresponding source token sets	36
3.7	Source type sets for the target words RIVE and BANQUE	36
3.8	Sense Tagged Type Source Sets	38
3.9	Sense Tagged Token Source Sets	38
3.10	Projecting source inventory senses onto target language instances . . .	38
3.11	An excerpt from the noun database of WN17pre	41
3.12	SALAAM performance precision & recall results in the default conditions plotted against state-of-the-art WSD systems on the same test set SV2AW	51
3.13	SALAAM F-Measure results in the default condition measured against state-of-the-art WST systems on test set SV2AW	53
3.14	SALAAM F-Measure results in the default condition measured against MAX intralanguage condition for the three languages: AR, FR, SP . .	55
3.15	SALAAM F-Measure results on test set in SV2AW in the highest yielding conditions depicted against state-of-the-art WSD systems . .	59
4.1	An example of a transliterated Arabic sentence and its tokenization .	76
4.2	WN17pre entries for <i>evening</i>	77
4.3	English WN17pre entries for <i>care</i>	91
4.4	Metonymic sense of <i>tea</i> in WN17pre	92
4.5	English WN17pre senses for <i>ceiling</i>	92
4.6	Homonymic sense for <i>tower</i> in WN17pre	92
4.7	WN17pre senses for <i>experience</i>	93
5.1	Sense distribution correlations across different nouns in the test data and hand-tagged training examples	104
5.2	Trend lines of the perplexity measure for test data and hand-tagged training data	105

5.3	Comparison between Mihalcea’s results and SALAAM results on the same test set	112
5.4	Trend lines for the precision obtained by SALAAM-SV2LS-TR and UMSST-SALAAM	113
5.5	A plot of the distribution of the senses’ contexts of <i>bar</i> and <i>day</i>	120
5.6	A plot of the SDC and performance ratio on the 29 nouns	125
5.7	A comparative view of the different perplexity measures in SALAAM-tagged training data and the test data for the 29 nouns	127
6.1	Random sample of verb source type sets yielded by SALAAM	147
7.1	SALAAM F-Measure results depicted against state-of-art WSD systems	155
7.2	Comparison between Mihalcea’s results and SALAAM results on the same test set	157

Chapter 1

Thesis Introduction

1.1 Introduction

Ambiguity is an inherent characteristic of natural language, permeating its various levels of representation. From a human language processing perspective, ambiguity is not a severe problem. However, from a machine processing perspective, the story is quite different. Resolving ambiguity in natural language has been of central interest to researchers from the early 1950's. In particular, Word Sense Disambiguation (WSD) has occupied center stage in the early work on Natural Language Processing (NLP). Bar-Hillel (1960) claimed that word sense ambiguity is the main impediment facing the field of Machine Translation (MT); in his famous treatise on MT, he describes the problem of WSD as insurmountable which leads to the abandonment of MT all together in the late 60s [33].

Fortunately, we have come a long way from the 1960s. With the on-going surge in machinery allowing for the development of sophisticated techniques and algorithms, WSD is experiencing a revival of interest especially with the belief that it has the potential of improving several central tasks in NLP. Owing to WSD's acknowledged significance in the field of computational linguistics, the community organized the first SENSEVAL which took place 3 years ago. It was succeeded by SENSEVAL 2 in 2001. As the name indicates, SENSEVAL is a defined protocol for developing, testing and evaluating WSD systems. SENSEVAL provides the opportunity, for the first time, for researchers working in the area of WSD to investigate common material, share experiences and exchange ideas within a defined framework. An important contribution by SENSEVAL is the creation of standardized tests and tools for measuring systems' performance and comparing notes.

But what is WSD?

WSD is the process of resolving the meaning of a word unambiguously in a given natural language context. It is the process of marking text with explicit sense labels.

What constitutes a sense in natural language is a subject of vast debate, both in the areas of lexical semantics and computational linguistics. The study of word meaning is at the core of research in the field of lexical semantics. Researchers such as Cruse

(1986), Pustejovsky (1995) and Levin (1990), among others, investigate word meaning within the same language — monolingually — with the goal of quantifying meaning dimensions. An alternative approach is to use cross-linguistic correspondences for characterizing word meanings in natural language. This idea is explored by several researchers, Dyvik (1998), Ide (1999), Resnik & Yarowsky (1999), and Chugur, Gonzalo & Verdejo (2002) but to date, it has not been given any practical demonstration. This thesis is an empirical validation of this very notion of characterizing word meaning using cross-linguistic correspondences. The idea is that a word meaning or a word sense is quantifiable as much as it is uniquely translated in some language or set of languages.

To date, most large scale WSD methods have defined context for sense definition within a monolingual framework; the evidence for sense choice is typically from within the same language. In this thesis, we address the problem of WSD from a multilingual perspective; we expand the notion of context to encompass multilingual evidence. We devise a new approach to resolve word sense ambiguity in natural language in a novel way, using a source of information that was never exploited on a large scale for WSD before.

The core of the work presented in this thesis builds on exploiting word correspondences across languages for sense distinction. In essence, it is a practical and functional implementation of the basic idea common to the research interest which defines word meanings in cross-linguistic terms. We devise an algorithm that empirically investigates the feasibility and the validity of utilizing translations for WSD. The algorithm presented is an unsupervised approach for word sense tagging (WST) of large amounts of text given a parallel corpus and a sense inventory for one of the languages in the corpus.¹ We refer to the presented algorithm as SALAAM for Sense Assignment Leveraging Alignments and Multilinguality. A parallel corpus is defined as texts in translation comprising a source language and a target language. The approach is unsupervised inasmuch as it does not require sense-annotated data at the onset.

Availability of automated knowledge resources for different languages is a serious obstacle for the study of language within a computational framework. In a more globalized community, the need for knowledge resources in different languages is ever more pressing. Yet, the distribution of tools and resources is asymmetric with *rich* languages such as English possessing the lion's share. In this thesis, we address this issue with a new technique for leveraging off of the *rich* languages to help create resources for *poor* languages with minimal automated linguistic resources. Within the scope of this thesis, a language is defined as *rich* or *poor* depending on the amount of automated resources available for it.

Furthermore, we investigate the impact of bootstrapping supervised WSD systems with large amounts of noisy sense-annotated data produced by exploiting multilingual evidence by SALAAM. Typically in the area of bootstrapping supervised systems, re-

¹We use WST and WSD interchangeably throughout the thesis

searchers have relied on clean knowledge resources to create training examples for supervised systems. Given that cleanly tagged data is hard to come by for rich languages — let alone poor languages — this thesis explores to what extent bootstrapping off of noisy data is a feasible enterprise. In the process, we create a novel unsupervised learning technique for WSD that does not rely on the availability of manually tagged data, whose absence is a severe bottleneck for canonical supervised learning approaches addressing sense ambiguity.

Acknowledging the central role played by similarity in the field of WSD, we examine different facets of quantifiable semantic similarity. We are very interested in how such similarity measures compare against human similarity judgments. So far, in the thesis, the focus is on nouns due to their immediate relevance to several application areas such as Information Retrieval and Information Extraction, etc., but we realize the complexity of annotating verbs.² Owing to the endemic complex multi-dimensional nature of verbs, they present themselves as luring entities for exploring the various dimensions of semantic similarity; moreover, all WSD systems encounter problems when dealing with verbs. We posit that the crux of the problem lies primarily in the similarity measure at the core of the WSD system. Most approaches to similarity within WSD are monolithic in essence, relying on one source of information. In this thesis, we empirically establish the merit of combining evidence from different complementary sources of information based on a cognitively based functional study of verb similarity and how it relates to different automated similarity measures, thereby, serving as the motivation for enhancing similarity measurement within the scope of WSD in general and SALAAM in particular.

1.2 Research Contributions

This thesis contributes the following to the field of computational linguistics:

- The contribution of a novel robust unsupervised approach to WSD which constitutes a significant departure from the traditional monolingual approaches. The approach is a validation of a sound linguistic assumption that meaning characterizations can be captured cross-linguistically. We contribute a novel multilingual perspective on the notion of context for addressing the problem of WSD. The context scope is no longer confined monolingually. (See Chapter 3 for description of the basic method; See Section 4.2 Chapter 4 for a discussion on robustness of the approach)
- The provision of a detailed description of an end-to-end fully operational, modularly designed system for producing large amounts of good quality sense-annotated

²None of the algorithms presented in this thesis has an inherent restriction on part of speech.

data in both source and target languages for a parallel corpus. (See Chapter 3, Section 3.6, as well as Chapter 4, Section 4.3)

- The investigation of the quality of automatic sense-annotations for a language with few computerized linguistic resources such as Arabic. (See Chapter 4, Section 4.3.5)
- The provision of an operational end-to-end robust automatic framework for testing the quality of projected automatic sense-annotations for Spanish. (See Chapter 4, Section 4.3.6)
- The examination of the feasibility of automatically bootstrapping a WordNet style ontology for Arabic via projected sense tags from English. (See Chapter 4, Section 4.4)
- The investigation of the feasibility of bootstrapping WSD within a supervised paradigm using noisy data based on the results obtained using the novel unsupervised method for WSD described in Chapter 3. Simultaneously, presenting a novel unsupervised learning technique for WSD. (See Chapter 5)
- The provision of a novel experimental design for attaining human judgments on semantic similarity for verb pairs using contextual and non-contextual data. The thesis compares the results obtained by several automated semantic similarity measures against the human similarity ratings. (See Chapter 6)
- The thesis utilization of insights derived from the human similarity judgment experiment to motivate an operational cognitively based framework for utilizing similarity in a novel way for improving WSD results obtained for verbs. (See Chapter 6).

1.3 Thesis Layout & Brief Overview of Chapters

This thesis comprises 6 chapters briefly described as follows:

- In Chapter 2 briefly surveys earlier work in the field of WSD. We look at three different components: History of the field of WSD (Pre-SENSEVAL), SENSEVAL Era, and then we describe earlier related work of systems addressing the problem of WSD from within a multilingual framework, which are all, incidentally, Pre-SENSEVAL.
- Chapter 3 presents the underlying hypothesis driving the research theme of the thesis. In this chapter, we describe the relevant background which lends preliminary theoretical support to the developed approach. We then present an unsupervised method for word sense tagging (SALAAM) based on multilingual

evidence. We describe the method and system in detail and present rigorous evaluation of the approach against state-of-the-art WSD systems. SALAAM is evaluated using machine translated parallel corpora (pseudo-translated corpora). The evaluation is confined to the nouns in the source language only.

- Chapter 4 presents further evaluation of the robustness of SALAAM by extending the utilized corpora to naturally-occurring parallel corpora of non overlapping genres. In this chapter, projected sense tagging on the translation language of a parallel corpus is evaluated. We investigate the quality of the projected sense tagging on both Arabic and Spanish. Furthermore, we explore the potential of automatically bootstrapping an ontology for Arabic.
- Chapter 5 investigates the feasibility of bootstrapping a supervised WSD system using noisy data as training examples. In this chapter, data is obtained from the SALAAM sense tagging system for the source language, English; such data is used to train a machine learning algorithm for WSD. Furthermore, in this chapter, we explore different factors affecting the bootstrapping performance by comparing the performance of the learning system when trained on SALAAM tagged data against it when trained on manually tagged data.
- Chapter 6 presents a novel experimental design for obtaining verb semantic similarity judgments. We compare the results obtained by different automated similarity measures against human similarity judgments. We lay out a cognitively based framework for integrating different automated similarity measures in order to approximate human judgments in the similarity component of SALAAM.
- Chapter 7 concludes the thesis with overall observations and lessons learned. A close look at the limitations of the different proposals contributed in the thesis is rendered. We reiterate the contributions of this research to the field of computational linguistics. Finally, we conclude with a peek into the future with some suggested directions.

Chapter 2

Related Work

2.1 Introduction

In the literature, typically WSD/WST systems associate labels with discovered senses;¹ the labels may be words from a different language, sense codes or definitions from an ontology, or artificial codes. WSD, in this view, is a classification problem where the sense labels are the classes to which the WSD/WST process assigns the discovered senses.

2.2 Pre-SENSEVAL WSD: Historical Perspective

WSD within that framework has been a problem of central interest to computer scientists in general, and Artificial Intelligence practitioners in particular since the early fifties.² In fact, with the inception of the field of computer science, the question of addressing ambiguity in language assumed center stage, after-all the goal was to create machines that understand language the way humans do.

In the early fifties, the effort was consecrated to WSD within a Machine Translation, MT, framework. The earliest approach was by Weaver in 1949 [80]. He argued the need for WSD in MT, as described in his Memorandum. He investigated the size of context needed to resolve the ambiguity of a word. He concluded that there is no difference in disambiguation power between a context of two words versus the context of the entire sentence. His observations were further confirmed by several other researchers in the 70s and 80s. A very important notion was supported by researchers in the field, which was that possible senses of a polysemous word are bound by the domain of the document a polysemous word appears in. Interestingly, over forty years

¹WSD, when used to refer to discovering senses without labelling them, it is known as Word Sense Discrimination. Word Sense Discrimination is a completely unsupervised approach that is not subject to label granularity restrictions as WSD/WST is. For purposes of this thesis, we are not concerned with Word Sense Discrimination.

²For an excellent survey, see the paper by Ide & Véronis 1998 [33]

later, Gale et. al. [27], further emphasized by Yarowsky [85], use this same idea of a single sense per discourse in an axiomatic manner, guiding in their view, the sense distribution of polysemous words in documents.

Most of the methods that tackled the problem of WSD afterwards were AI — Artificial Intelligence — based. In the 70s and 80s, the majority of the approaches were grounded in language understanding theories where the systems tried to model deep knowledge of linguistic theory, especially syntax and semantics. A wave of word representation techniques was developed in an attempt to capture the relevant facets of meaning in order to solve word ambiguity issues. It was during this period of time that Semantic Networks by Quillian [65],³ Frames by Hayes [30],⁴ and Preference Semantics by Wilks [81], appeared on the scene of WSD. These three methods assumed the crux of symbolic approaches to WSD in that era. These techniques contrasted with more data driven solutions of the time. Examples of which are Small et al. [77] developed intricate representations of words referred to as word experts. Such experts were complex in nature but the approach in general constituted a departure from the rule dependent perspective to the more word oriented one. His methods were similar in spirit to those of Kelly & Stone [38] who also focused on word oriented approaches, but in Small's case his aim was broad natural language understanding in contrast to Kelly & Stone who had the specific intent of word sense disambiguation from the onset.

However, such methods lost their appeal by the late 80s due to the intensive labour involved in the creation of the required intricate knowledge representations, which bound the number of words and senses that could undergo analysis and disambiguation. This was coupled by a surge in machinery and the beginnings of the availability of Machine Readable Dictionaries (MRD) and Lexica. One of the first attempts to utilize such resources was Lesk [43]. He devised an algorithm that chooses the appropriate sense of a polysemous word by calculating the word overlap between the context sentence of the word in question and the word's definition in an MRD. Most of the algorithms that followed were in tune with that spirit from then onwards, more corpus based and knowledge based in nature where the role of the surrounding monolingual context is paramount in providing the needed evidence for a word sense. To date, most existing algorithms are a variant on the Lesk algorithm in their view of context and resolution of the WSD problem in general.

It was in the 90s that that perspective of WSD was shifted to being regarded as an enabling technology with a lot of prospect for NLP applications, if it were to be resolved once and for all. Yet, with the abundance of algorithms, it became extremely difficult to assess their quality or results. Every algorithm was evaluated on a different test set and with different evaluation criteria and metrics. It became ever more difficult to establish a *sense* for the state-of-the-art in WSD.

³Each word is represented as a node in an interconnected web

⁴Words are represented as entities with their roles and their connections to other words in the sentence explicitly defined

2.3 The SENSEVAL Era

Driven by the lack of common standards for evaluation and the need to assess different systems' performance, while simultaneously getting a feel for the myriad of approaches to WSD, the computational linguistics community decided to create a standardized test bed or yard stick through which it can facilitate communication and collaboration among researchers in the field as well as establish a rigorous means through which the community can evaluate the state-of-the-art in the field of WSD [41].

This was the marking of a new era in WSD, the SENSEVAL era. As the name indicates, SENSEVAL is an evaluation framework *à la* the different Information Retrieval type evaluation paradigms such as TREC and MUC, where the community decides and creates standardized data sets and test beds with well defined metrics for state-of-the-art assessment. The inception of SENSEVAL was triggered by a position paper by Resnik and Yarowsky presented in the Special Interest Group for Lexical Semantics (SIGLEX) workshop in DC, in 1997. They outlined a set of proposals and discussed different problems faced by practitioners in the field and how they believe these impediments may be overcome.[74] In their article, Resnik & Yarowsky make a set of observations about state-of-the-art WSD by comparing advancements in the field with progress levels achieved by other enabling technologies in NLP. They offer several proposals with the goal of improving the evaluation criteria for automatic WSD, improving the process of acquiring training and testing materials as well as defining sense inventories. They devise an iterative protocol for dealing with major issues that hinder the creation of a standardized benchmark for comparing the performance of WSD systems. The results of their proposals are adopted in both SENSEVAL 1 and SENSEVAL 2 exercises. The following is an outline of their proposals:

- **Evaluation criteria**

The authors criticize *exact match* as an evaluation metric that was the common practice until recently. The problem, they point out, is that *exact match* is a binary measure. It does not discriminate probabilistically between a tagger that claims ignorance for a sense assignment, for example, and a tagger that assigns an incorrect sense a lower probability; with *exact match* both systems are equally penalized. Therefore, they propose an evaluation metric based on a measure of cross-entropy that credits a tagger partially based on the probability assigned to the correct tag. (See Chapter 3 for a more detailed discussion of this measure).

- **Protocol for systematic evaluation of WSD systems and sense-tagged data acquisition**

They develop a protocol for acquiring large amounts of tagged data systematically and incrementally. Their rationale is based on the stipulation that it is better to tag large amounts of examples for a small number of words with clear guidelines and extensive in depth analysis of the tagging task rather than a small

number of examples for a large number of words. They justify their proposed iterative approach based on four main points:

1. The protocol combines an emphasis on broad coverage with advantages of evaluating a limited set of words by choosing the words that cover a wide range of frequencies, levels of ambiguity, etc;
2. A small, predefined set of words is more tractable for the manual annotator as s/he only needs to focus on one word at a time;
3. With a small number of words and large number of examples more attention can be dedicated to the specifications and guidelines of the manual annotation process, thereby reducing the number of possible problems;
4. This proposed protocol addresses needs of both supervised and unsupervised systems.

Based on this proposal and further discussions within the community, the first SENSEVAL was conceived. The organizers for SENSEVAL coordinate the creation of different tasks for different languages. The exercise takes place within a specified window of time with hard deadlines. The ontologies used for the tagging are determined before-hand for each language. So far, there are two types of task for any given language: a Lexical Sample task and an All Words task; not all languages have both tasks.

Typically a Lexical Sample task is where a number of polysemous words in a large corpus is chosen for tagging by the organizers of the task. The systems are required to tag instances of the chosen words in a specified corpus. The systems' performance is evaluated by the organizers. For instance, in the recent SENSEVAL 2 exercise, the number of English nouns in the Lexical Sample task is 29 nouns of different levels of polysemy. The organizers of the task provide trial, training and test data to the participating systems at predefined time intervals. The trial data is for calibrating the participating systems in terms of format issues. The training data provides large amounts of annotated data where the predetermined words are tagged in context. The contexts are typically two to four lines in length. Finally, the participants receive the testing data a number of days before the submission of results to the organizers for evaluation.

The All Words task follows the same time line as the Lexical Sample task but in this case the WSD systems are required to tag all words in a specified corpus, i.e., all content words in running text. Typically no training data is provided for this task.

SENSEVAL 3 is bound to take place in mid 2004 with even more languages.

2.4 Multilingual Approaches to WSD

In this section, we focus on previous WSD approaches that address the problem within a multilingual framework since it is directly related to the topic of this thesis. For a general survey of approaches to WSD, we recommend the article by Ide and Veronis [33]. Several systems have addressed the problem of WSD/WST within a multilingual framework. They are all pre-SENSEVAL. In this section, we will describe four representative approaches that addressed the issue. They all exploit the observation that when a polysemous word in a language (L1) is translated into another language (L2), often senses of the polysemous word in L1 are translated into distinct L2 words in different contexts. Accordingly, they use lexical translations as a source of sense distinction. This idea has been around since the early 1990s [7, 15, 27].

The four WSD/WST methods described below are statistical corpus based methods; they differ mainly in their algorithmic approaches and their resource requirements. The first two methods, Brown et al. (1991) and Gale et al. (1992), only require the availability of parallel corpora, that are sentence and token level aligned. They both use the translation target words as labels for the ambiguous words in the source language. Both the third and fourth presented approaches, Dagan & Itai (1994) and Kikui (1999), rely on bilingual comparable corpora and bilingual dictionaries. Moreover, the work by Dagan & Itai requires a parser for at least the source language. The first method and the last two methods explicitly aim at improving target word selection in a practical machine translation application environment. They all utilize monolingual context on the source language side and have some way of bridging cross linguistically to the target language side. In the first two methods, the authors use token level alignments in a parallel corpus as a bridge between the two languages; in the latter two studies, the authors use bilingual lexicons. All four systems present results on a handful of data and unfortunately the evaluation metrics are different in each paper, rendering it difficult to compare performance across systems.

2.4.1 Word Sense Disambiguation Using Statistical Methods: *Brown et al.*

One of the earliest WSD studies within a multilingual framework is research by Brown et al. [7]. They present a pilot study where they investigate the impact of adding a sense disambiguation component to a statistical machine translation system [6, 8]. The goal in this study is functional, namely, the improvement of lexical generation for an ambiguous word in a machine translation application. The notion of a sense in the context of translation combines pragmatic uses of words across languages that are not necessarily ambiguous in a source language together with genuine ambiguity such as is the case of a word like *bank* in English. An example of pragmatic disambiguation is demonstrated as follows: Choosing the correct translation for the word **il** in the French sentence **il y a une probleme.** is considered by the authors a sense

disambiguation problem with the choice between translating the French word **il** as *it* or *he* in English. Nonetheless, the ideas that are proposed are very interesting and may be extended to resolving paradigmatic genuine sense ambiguity problems. They obtain an improvement of 37% in the quality of translations.

Method Requirements

The method requires a parallel corpus and part of speech taggers for both languages of the parallel corpus.

Method Description

- The approach assumes the availability of a statistical machine translation system that creates alignments between words in the English-French parallel corpus.
- A set of most frequent words is extracted from both sides of the parallel corpus.
- Each of the words is described in terms of a number of contextual informants. The words in the target language, French, have seven informant features: *tense-of-current-word*, *word-to-left*, *word-to-right*, *first-noun-to-left*, *first-verb-to-left*, *first-verb-to-right*, and *first-noun-to-right*. For English words, only two informants are defined, *first-word-to-left* and *two-words-to-left*.
- Only two senses are allowed per word in either language. The WSD system makes a binary decision between the different informants and the translations of the word in question.
- The flip-flop algorithm is used [60] in conjunction with the splitting theorem [4] in a fashion similar to decision tree learning. The flip-flop algorithm asks binary questions of a set of English translations corresponding to a French word. It divides the translations into two classes. The splitting theorem's role helps in deciding the best informant (feature) based on mutual information in linear time. The best question about a potential informant is discovered; in turn, this question divides the French vocabulary set into two classes; the algorithm then uses the Splitting Theorem to divide the set of English translations into two sets that have maximum mutual information with the French sets. The process goes on, alternating between splitting the French vocabulary and the English translation sets. Since this is a binary process, only one bit of information is the bound; eventually, the algorithm converges on the English translation that has the maximal mutual information with the French word.

Evaluation

The method is evaluated on 100 randomly chosen English-French sentence pairs. The algorithm is incorporated in a statistical machine translation system; therefore, it is an indirect evaluation. The machine translation output is manually marked as *acceptable* or *unacceptable* by the authors. The use of sense disambiguation improves the results from 37% (without sense disambiguation) to 45% (with sense disambiguation).

Discussion

The authors acknowledge the limitations of the approach stating that it is a pilot study. They also acknowledge that the approach is binary, which is not a realistic scenario. They point out that if the number of classes is unbounded, they expect the results to improve even further; therefore, instead of having binary questions the system will have n questions.

In our view however, there is no guarantee if the number of senses is unbounded that the system will improve in performance; the system is faced with a problem of fan out with the level of noise increasing exponentially, especially if the alignments and part of speech taggers are not 100% accurate.

Moreover, this approach is limited by the possible sense-annotations, which are corpus specific; this creates problems when porting such an approach to different corpora even within the same domain. Furthermore, the informants are instance specific — tokens, i.e., they consider the actual token in the immediate context. One simple way of surmounting such a criticism would be using types instead of tokens as informants.

2.4.2 Using Bilingual Materials to Develop Word Sense Disambiguation Methods: *Gale et al.*

In addition to presenting a disambiguation approach, this study considers solving the training materials bottleneck for supervised systems. Gale et al.[27] make an explicit distinction between the sense disambiguation problem and the translation disambiguation problem. They provide a method within a multilingual framework for creating sense disambiguated materials using translations as labels to annotate a set of polysemous words in a source language of a parallel corpus. They report results of 90% correctness for a set of six polysemous words where each word has two senses.

Method Requirements

A parallel corpus with sentence and token level alignments.

Method Description

This is a supervised approach to WSD. There are two phases: a **Training Phase** and a **Testing Phase**. Given a parallel corpus that is sentence and token aligned, start by creating the training material and then test new items using the coefficients obtained from the training phase. The training and test phases are performed as follows:

- **Training Phase:** The sense of a polysemous word instance in context is identified based on its translation — its alignment — to a target language token.
- **Training Phase:** The context score of an instance of a polysemous word is calculated based on a variation on Information Retrieval (IR) techniques, where the contexts are considered in lieu of documents, according to equation (2.1). The score is obtained by calculating the probability of a token appearing within a window of 50 tokens on the right and left of a polysemous instance.

$$score(c) = \prod_{tokeninc} \frac{prob(token|sense_1)}{prob(token|sense_2)} \quad (2.1)$$

This model ignores word order and collocation information. Local token probabilities are too sparse in general; therefore, they opt for a weighting scheme as a smoothing approach. The weights are a ratio of the *local source token log likelihood probabilities* and the *global log likelihood probabilities*. Accordingly, a token that is very frequent in the entire corpus and is frequent in the local context is assigned a low weight value, while a token that is sparse in the entire corpus but frequent in the local context is assigned a high weight value.

- **Testing Phase:** Test instances of the polysemous words are identified
- **Testing Phase:** Test instances are scored using equation (2.1)
- **Testing Phase:** Test scores are compared with training scores and senses are selected based on context score proximity.

Evaluation

The method is trained and tested on six polysemous nouns with two distinct senses each. The nouns are $\{duty, drug, land, language, position, sentence\}$. The six nouns are selected because their senses correspond to distinct words in French. The training set has 60 training examples per sense. The test set comprises 90 instances per sense per word. The results obtained manually range from 82% to 100% accuracy.

In the process of their evaluation, the authors discuss and provide empirical justification for the context window size chosen. They establish that contextual clues are measurable up to a 10,000 words out from the word of interest. They relate this fact to

the nature of discourse structure. Moreover, they illustrate that, contrary to common belief that only ± 6 words are sufficiently useful.⁵ Up to ± 50 words is useful for a machine to make sense distinction decisions.

Furthermore, they explore the quality and quantity of the training data on sense disambiguation performance. For the impact of quantity, they, surprisingly, show that with only three training examples, their system is able to achieve an accuracy of 75% and accuracy asymptotes as the number of examples increases. As for the question of quality, the authors systematically show the degradation effect on the accuracy of their system's performance if the training data has errors. With 10% errors introduced in the training set, they obtain a 2% decrease in accuracy; with 30% errors in the training data, the precision decreases 14% only, which is still very robust in their view.

Discussion

Even though this evaluation is done on a very small scale, with six words of two senses each, the study tackles issues that are of central concern to us throughout this thesis.

This approach performs sense disambiguation using translation words as labels identifying the different senses of a word. The authors erred on the side of caution in this study by using only two homonymic senses per word; they did not look at other polysemic relations such as regular polysemy or metonymy. Homonymic senses are the most likely senses to translate to distinct words in other languages.

With the current availability of token alignment software and bilingual parallel corpora, it would be interesting to explore how this method scales up when given large amounts of data and polysemous words.

2.4.3 Word Sense Disambiguation Using a Second Language Monolingual Corpus: *Ido Dagan & Alon Itai*

In this paper [15], Dagan & Itai present a novel approach for resolving lexical ambiguity in one language using statistical information from a monolingual corpus in another language. The method aims to solve the problem of translation word selection for machine translation applications. The method uses a parser for both languages and statistical information from the target language corpus to decide on the most appropriate translation for an ambiguous word in a source language. The approach is evaluated on two different source languages, German and Hebrew, with English as the target language for both. The authors report performance scores of 91% accuracy on Hebrew-English translations and 78% on German-English translations.

⁵In contrast to the studies by Weaver (1949) [80]

Method Requirements

The approach requires the availability of a bilingual lexicon, a parser for the source language corpus and one for the target language. In principle, there is no restriction on the type of corpora, yet preferably they should be of the same genre.

Method Description

- **Parsing the source language into syntactic tuples**

In their implementation, the authors use Slot Grammars [53] which is a form of dependency parsing identifying *verb-object*, *verb-subject*, *word-adjunct*, etc. type syntactic relations. There is no commitment in the paper to a specific parsing paradigm as long as syntactic tuples may be extracted from the parsed corpus.

- **Locating ambiguous words in the source language**

An ambiguous source word is defined, in the context of this paper, as a word that has multiple translations in a bilingual lexicon and fits the syntactic frame of the specific source word instance in the source corpus. Given such a definition, many of the alternative source senses are pruned on syntactic grounds. The words are lemmatized before parsing to reduce sparseness.

- **Mapping source syntactic tuples to target language**

The method is straightforward: Using the bilingual lexicon, the words in the source syntactic tuple are translated to the target language. The authors identify syntactic divergences cross-linguistically, where there is a mismatch between syntactic frames in the source and target languages, as a source of controllable noise. For instance, for some verbs in German, their objects translate into subjects when translated into English. For example, the German sentence **Der Tisch geflaellt mir** is translated as *I like the table*; the subject **Tisch** in German becomes the direct object *table* in English; and the object **mir** in German becomes the subject *I* in English. Such divergences are dealt with by means of hand coded rules that target a class of verbs that exhibit such a phenomenon.

- **Choosing the most appropriate translation tuple from the target language corpus**

This phase involves several filtering steps.

- (1) The first step depends on the frequency of observing the translation tuple in the target corpus. This step weeds out some implausible tuples if they have not been seen in the target corpus.

- (2) The second step is addressed by a probabilistic model for the different possible target tuples denoted as T with frequencies n in the target corpus. T has a multinomial distribution and has the possible values T_1, \dots, T_k . p_i is the probability that T_i is the correct translation of T . The authors use the maximum likelihood estimator to estimate the probability p of any given tuple T . The counts n associated with different T values are sorted in a descending order. A threshold is set. The ratio of the estimated probability for a certain tuple and the estimated probabilities of all the other tuples has to exceed the set threshold. The threshold is small when the frequency counts are very distant and it is large when the counts are close. The ratio is referred to as the *odds ratio*.

This model entails three underlying assumptions: The events in T_i are mutually disjoint; a source language syntactic tuple can be translated into one of the tuples T_1, \dots, T_k ; and every occurrence of the tuple T_i can be the translation of only one source language syntactic tuple.

The authors then define a confidence interval for deciding on the quality of the data, i.e., whether the translation tuple is good enough to be chosen as a translation for the source tuple. The confidence interval depends on the counts of the target tuple in the target corpus and the *odds ratio threshold*. The threshold is higher when the values of n_1 and n_2 are smaller, thereby creating a dynamic threshold that has the desirable effect of pruning cases where the data is not supportive enough.

- (3) The third step is dealing with situations where there are multiple ambiguities from multiple syntactic tuples in the same sentence. The authors devise a constraint propagation algorithm that takes the list of all source tuples and their possible alternative translation target tuples and eliminates the tuples that do not satisfy the threshold set with a prespecified confidence level.

Evaluation

The method is evaluated on a random set of examples. The examples consist of source Hebrew and source German paragraphs. In both cases the target language is English. The Hebrew examples are randomly picked from Foreign News sections in the Israeli press. The German paragraphs are picked from the German press. The corresponding English target text is picked from American news articles as well as the Hansard corpus of the Canadian Parliament.

The choice of ambiguous words is simulated with a translator and a preliminary bilingual lexicon. For every source language word, the translator searches all possible translations in a bilingual dictionary; s/he eliminates those that do not fit the syntactic structure of the source instance. The translations in the bilingual dictionary are modified manually to be closer to what would be expected of a transfer translation lexicon.

Once the ambiguous source words are located, the syntactic tuples are determined and mapped into English. Since they do not have a parser for the source languages, they manually translate the source language paragraphs into English; the translation is a very literal translation; the resulting manual literal English translations are parsed using the ESG parser, which identifies the relevant syntactic tuples in the source language through a simple mapping routine. This process results in 103 ambiguous Hebrew words and 54 ambiguous German words.

The statistical English data is acquired from a 25 million word corpus that is filtered from a combination of The Washington Post, the Hansard corpus of the Canadian Parliament, and Associated Press news items. Only sentences of 25 words or less are used. This approach is referred to as Translation Word Selection (TWS). The baseline created is the most frequent translation target word. For convenience we will refer to it as (FB).

The authors report results using two evaluation metrics: *applicability* and *precision*. *Applicability* is a coverage measure, i.e. how many cases are attempted out of the possible cases; while *precision* is the typical metric of how many found items in those retrieved are correct instances. For Hebrew, TWS results are 91% precision and 68% applicability, while FB achieves 63% precision at the same applicability level as TWS. For German, the results are not as good with a precision of 78%, applicability of 50% for TWS, and precision of 56% for FB. The German results are lower due to the change in corpus genre from source test set to target language corpus genre.

Further results are reported on applying the TWS approach with the parser on the source side alone, approximating the parser on the target side with collocational information collected from the target corpus. The results yielded are lower at 85% precision and 64.3% applicability for Hebrew. These results are compared against an FB of 71.1% precision.

Discussion

This paper presents a very interesting approach to solving sense disambiguation for a specific target application, machine translation. The authors use linguistically motivated models — parsers — in conjunction with statistical information in a hybrid manner, combining different sources of information. They approximate the lack of tools such as parsers and bilingual lexicons using manual resources, yet they present very rigorous simulations and evaluation criteria, which is very inspiring for the current thesis. This paper dates back to the early 1990s when parallel corpora were still an extremely expensive resource to obtain, supporting their strong argument against using them. Yet, parsers for many languages do not exist, and their approximation with the manual translation is feasible due to the limited scale of the evaluation, with only 130 paragraphs in total for both source languages, which is a severe impediment for applying this method on a large scale using such simulations. The article seems to dismiss the complexity of acquiring a bilingual lexicon. Building such a resource with

an adequate level of coverage is not a trivial matter, especially if the lexicon is required to list syntactic subcategorization frames, which is a requirement for this approach to work.

2.4.4 Resolving Translation Ambiguity Using Non-Parallel Bilingual Corpora: *Kikui*

This is a more recent approach to WSD within a multilingual framework [39]. This study presents an unsupervised approach for choosing an appropriate translation for a source language word to a target language, given a specific context. The method incorporates two different unsupervised modules: a Distributional Sense Clustering algorithm applied to the source language; and a Translation Disambiguation algorithm applied to the target language by linking the source sense clusters to their translation equivalents in the target language. The method is tested on an English to Japanese machine translation system with promising results.

Method Requirements

- Large amounts of bilingual comparable corpora where the corpora are of the same domain and time frame.
- A bilingual dictionary

Method Description

- **Distributional Sense Clustering algorithm**

Both corpora are sense disambiguated using the distributional clustering approach introduced by Schütze [75]. The method encodes ambiguous words as vector profiles where the different dimensions are the words that fall within an n -sized window from the ambiguous word in question. The contents of the vectors are the co-occurrence frequencies. Similar to Schütze, Kikui uses Singular Value Decomposition to reduce the dimensionality of the data. An agglomerative clustering algorithm is applied to the vectors to create the sense clusters.

- **Translation Disambiguation algorithm**

The Distributional Sense Clustering algorithm is applied to the most frequent terms in a source language corpus where source sense clusters are created. Using IR techniques, the source words are pruned such that only words with high *tf-idf* values are kept, thereby creating a source term-list.⁶ The source term list is

⁶Term frequency (tf) divided by inverse document frequency (idf).

translated to the target language using a bilingual dictionary resulting in translation candidates.

The Distributional Sense Clustering algorithm is applied again to the target language. A cosine similarity measure is applied to the translation candidates and the resulting target language clusters, and those that have the highest similarity values are chosen as the target language translation.

Evaluation

The method is trained and tested on 1994 New York Times newspaper articles in English and 1994 Japanese Shinbon newspaper articles in Japanese. The gold standard is a set of manually corrected machine translation output. The method achieves an accuracy rate of 79.1% against the gold standard.

Discussion

The method as described aims at WSD using target language words as labels. Similar to the previous approaches, such a tagging technique is corpus specific; translation word instances are used as labels for the ambiguous source words. The method is language independent in the monolingual sense discrimination phase since the approach applies distributional clustering with no explicit language coding. Like all approaches that depend on bilingual dictionaries as a necessary bridging component, the method is limited by the coverage of the bilingual dictionary of the corpus terms. It would be interesting to explore how this method scales up to different genres corpora — or mixed genres corpora. The method is trained and tested on the same limited domain corpora. Accordingly, polysemous words in such corpora tend to have a very high bias toward specific senses.

2.4.5 Summary

System	Method	Corpus Type	Inventory	# Label	Linguistic Tools
Brown et al.	Sup.	Parallel		2 words	Tokenizers
Gale et al.	Sup.	Parallel		2 words	Tokenizers
Dagan & Itai	Unsup.	Comparable	Biling. Dict.	n words	Tokenizers & Parsers
Kikui	Unsup.	Comparable	Biling. Dict.	n words	Tokenizers

Table 2.1: Summary of Systems' Required Resources

In Table 2.1, we summarize the resources required by each of the systems described in this section. The first two methods require token aligned parallel corpora; the tag set — label — size is two senses where a sense is a target translation word. The second two methods utilize comparable corpora and require bilingual dictionaries. The tag set size is not limited; translation words are used as sense labels. The Dagan & Itai method requires parsers for both languages involved in the exercise.

System	Language	Metric	Size	Gold Standard	Performance
Brown et al.	En-Fr	improv.	100 inst.	No	8% improv.
Gale et al.	En-Fr	acc.	6 words, 140 inst./words	No	90% acc.
Dagan & Itai	Heb-En, Ger-En	prec., applic.	103 Heb., 54 Ger. inst.	No	91% prec. 63%applic.
Kikui	En-Jap	acc.	120 inst.	Yes	79.1 acc. %

Table 2.2: Summary of Systems' Evaluation

In Table 2.2, we give an overview of the four different evaluations we describe and discuss above. We characterize them in terms of the Language the approach is tested on, the Metric utilized, Size of data, presence of a Gold Standard, and Performance. We note that all systems use accuracy except for Dagan& Itai method. The first two systems do not define a Gold Standard. It is difficult to draw any conclusions on their respective performance since the different methods use different data sets and evaluation metrics.

Chapter 3

Word Sense Tagging Using Parallel Corpora: SALAAM

3.1 Introduction

Many researchers in the field of computational linguistics and specifically in the area of Word Sense Disambiguation (WSD) have exploited the observation that ambiguous words in one language translate into different words in a second language. In this chapter, we present a novel unsupervised method of Word Sense Tagging (WST) that builds on this very observation. It relies on texts in translation with the aim of resolving sense ambiguity in natural language; the approach described here is a multilingual unsupervised sense tagging approach that we will refer to as SALAAM which stands for Sense Assignment Leveraging Alignments and Multilinguality. SALAAM exploits the translators knowledge of language and the context of ambiguous words in language to sense-annotate large amounts of text for two languages simultaneously. As mentioned in the Introduction Chapter 1 to this thesis, WSD is believed to be an important enabling building block for potentially improving the performance in many NLP applications.

Within the area of data-driven WST, there are two main approaches with some hybrids: unsupervised methods [1, 49, 69, 83, 85], and supervised methods [10, 59, 84]. Supervised methods traditionally yield better performance results in WST [40]. The main difference between supervised and unsupervised methods lies in the need by the former for sense-annotated data for the training. Supervised methods are highly tuned to the training corpus type. This tuning helps in producing reliable results but it is a double edged sword since it significantly affects the portability of supervised systems to different corpora genres. Typically, supervised methods require large amounts of good quality data to produce good results. Unfortunately, large amounts of sense-annotated data do not exist for nearly all languages.

On the other hand, unsupervised methods have the advantage of making minimal assumptions about the data; they do not need sense-annotated data as a prerequisite. Comparatively speaking, unsupervised methods are less tuned to the corpus domain, which significantly impacts the quality of the tagging. If an unsupervised method achieves close to supervised methods' performance without relying on sense-annotated

data from the outset, then it is a significant contribution to the field.

The method we describe in this chapter, which is at the core of the following chapters as well, is an unsupervised method for word sense tagging a source and target language in a parallel corpus using the sense inventory of the source language. The method relies on the availability of large amounts of text in translation. It assumes the availability of asymmetric resources for two languages. Throughout this thesis, a source side of the parallel corpus is defined as the side that possesses the knowledge resources required.

This chapter is laid out as follows: The following section, Section 3.2 discusses the motivation for this study; Section 3.3 defines the problem; Section 3.4 describes the relevant background; in Section 3.5, we present the over-arching hypothesis and insight for the devised method; Section 3.6 describes the approach in detail; this is followed by a detailed evaluation of the source tagging quality in Section 3.7; results of the evaluation are presented in Section 3.8; results and shortcomings of the approach are discussed in detail in Section 3.9; finally, we conclude with a wrap up summary Section 3.10.

3.2 Motivation

What is the use in possessing large amounts of sense-annotated data?

sense-annotated data could potentially help improve many NLP applications. Moreover, possessing sense-annotated data in several languages provides an interesting test bed for exploring different cross-linguistic phenomena related to lexical semantics. It has been shown that several languages pattern in the same way with respect to certain metonymic relations such as container/contained sense transfers [82].

As mentioned earlier, supervised WSD methods yield better performance than unsupervised methods. But a severe bottleneck for supervised systems is the annotated data acquisition for training. Providing large amounts of sense-annotated data—albeit noisy—could potentially help alleviate this impediment. This idea is further investigated in Chapter 5.

Possessing sense-annotated data in a language with scarce knowledge resources potentially constitutes the initial step in bootstrapping sense inventories for such a language. We explore this issue further in Chapter 4.

3.3 Problem Statement

Manually sense-annotating texts is the guaranteed method of obtaining good quality tagged data. But alas, this is a very tedious and laborious job, let alone expensive [24]. Accordingly, automating the process is highly desirable. To our knowledge, all WSD systems aim at providing sense-tagged data in a single language at a time. Most methods naturally target languages with many automated resources and tools.

Languages with scarce resources—referred to as well as low density languages—are left behind in the process.

In this chapter, we introduce an unsupervised method, SALAAM, for word sense tagging that exploits texts in translation — parallel corpora. The method is unsupervised inasmuch as it does not rely on the existence of sense-annotated data as a prerequisite. The approach aims at resolving the ambiguity of polysemous words in two languages simultaneously: a language with rich resources and one with scarce resources. It aims at exploiting the asymmetry in resources for the benefit of both languages; SALAAM leverages off of rich source language resources to create a seed for acquiring automated resources for a low density language. The focus of this chapter is to lay out the methodology and present a detailed evaluation of the source language sense-annotation. Evaluation of the quality of the target language annotations is investigated in Chapter 4.

The approach presented explores the notion that senses of polysemous words in one language are often translated into different words in some set of other languages [74, 34, 32, 23]. Current approaches view the context of a polysemous word in question in terms of local monolingual features; the features could be in terms of the words, relations of words, or sentences surrounding the polysemous word. In contrast, we are defining a polysemous word’s context in cross-linguistic terms. Such a novel extension to the notion of context to cross language boundaries allows for the tagging of ambiguous words that are traditionally off the radar for contextually monolingual approaches.

3.4 Relevant Background

Using lexical translations as a source of sense distinction is an idea that has been around since the early 1990s [7, 15, 27] (see chapter 2). The key observation is that when a polysemous word in one language (L1) is translated into another language (L2), the polysemous word in L1 is translated into several distinct L2 words in different contexts corresponding to the L1 word’s various senses. The following sub-sections discuss different attempts at using that same idea for the purposes of WSD.

3.4.1 A Translational Basis for Semantics. *Helge Dyvik*

In this study [23], Dyvik examines how translational phenomena may be used as data for the development of linguistic semantics. He treats the translational relation between two languages as a primitive; a phenomenon that is accessible via bilingual informants. He distinguishes a translational relation from an abstract linguistic expression such as synonymy. Dyvik develops a theoretical framework for testing the validity of translational capacity as a discriminating basis for senses of ambiguous words.

Dyvik conducts a qualitative study on a set of Norwegian polysemous words and

their translations into English. He proposes an unsupervised method using texts in translation that does not rely on any external resources for sense distinction. He discovers word senses in a corpus by using translations and their reverse translations, i.e., manually locating the translations of a Norwegian polysemous word in the English text then searching for the translation of the English words that correspond to the original Norwegian word in the Norwegian text and so on, back and forth. He concludes that translation could indeed be used reliably for sense distinction since it is a linguistic primitive. Exploiting translations enabled him to discover appropriate senses for the majority of the Norwegian polysemous words investigated.

3.4.2 Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Philip Resnik & David Yarowsky*

In this article [74], Resnik & Yarowsky make a set of observations about state-of-the-art WSD by comparing advancements in the field with progress levels achieved by other enabling technologies in the NLP. They offer several proposals with the goal of improving the evaluation criteria for automatic WSD, improving the process of acquiring training and testing materials as well as defining sense inventories. They devise an iterative protocol for dealing with all the issues that hinder the creation of a standardized benchmark for comparing the performance of WSD systems. Moreover, such a protocol addresses the obstacles faced by researchers in the field. The results of their proposals are adopted in both SENSEVAL 1 and SENSEVAL 2 exercises.

- **Evaluation Criteria**

The authors criticize *Exact Match* as an evaluation metric that was the common practice until recently. The problem, they point out, is that *Exact Match* is a binary measure. It does not discriminate probabilistically between a tagger that claims ignorance for a sense assignment, for example, and a tagger that assigns an incorrect sense a lower probability; with *Exact Match* both systems are equally penalized. Therefore, they propose an evaluation measure based on a measure of cross-entropy that credits a tagger partially based on the probability assigned to the correct tag. The measure they propose is computed as

$$-\frac{1}{N} \sum_{i=1}^N \log_2 Pr_A(cs_i | w_i, context_i) \quad (3.1)$$

where N is the number of test instances and Pr_A is the probability assigned by the algorithm A to the correct sense, cs_i for the word w_i in $context_i$.

Given a hierarchical sense inventory, they further propose the evaluation measure be sensitive to the semantic distance between the sense labels. Therefore, if a tagger assigns a sibling of the correct sense to the word in question, the tagger should be penalized less than if it assigns the label for a homonymous sense of the word. Accordingly, they devise penalty distance matrices that capture taxonomic semantic distance in hierarchical Ontologies. Entries in the matrix are based on a pairwise calculation of semantic distance for all the senses of a given word. Taggers’ sense assignment is to be weighted by the communicative distance per sense pair in the ontology. They give the calculation as

$$\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{S_i} distance(cs_i, s_j) Pr_A(s_j|w_i, context_i) \quad (3.2)$$

where, for any test example i , all S_i senses (s_j) of word w_i are considered, weighting the probability mass assigned by the tagger A to incorrect senses ($Pr_A(s_j|w_i, context_i)$) by the cost or distance of the mistagging.

- **Multilingual Inventory**

The authors put forward a proposal to restrict a sense inventory to the distinctions that are attested for—lexicalized—cross-linguistically in some minimum number of languages. They do not specify how many languages constitute a reasonable *many*, however. In their view, this is a mid-point between very coarse-grained listings and the very fine distinctions permeating Ontologies such as WordNet.

- **Translation as a source of sense distinction**

In order to validate their proposal for sense inventories based on multilingual evidence, Resnik & Yarowsky explore the relationship between monolingual sense inventories and translation distinctions cross linguistically. They measure the probability of an English sense distinction being lexicalized differently in 12 diversely different languages, at various levels of granularity. They analyze native speakers annotations of 222 polysemous contexts across the 12 languages. They show that monolingual sense distinctions could be discriminated in some set of second languages. Moreover, their findings suggest a correlation between language family distance and the extent to which polysemous words express their various senses as distinct words, i.e., the farther the family distance of L1 from L2, the better the sense distinction. They cluster the resulting manual sense annotations of the English words and their corresponding translations. They obtain results that correlate well with monolingual sense distances in the hierarchical Hector sense inventory [41], thus lending support to the plausibility of hierarchical sense inventories. The clustering is performed based on a measure of

Sense Proximity. This is a cross-linguistic measure for calculating the extent to which two senses of a word lexicalize differently in a given language. The measure is defined as follows:

$$P_L(\text{different_lexicalization} | s_i, s_j) = \frac{1}{|s_i||s_j|} \sum_{\substack{x \in \{s_i \text{ examples}\} \\ y \in \{s_j \text{ examples}\}}} \text{translation}[x, L] == \text{translation}[y, L] \quad (3.3)$$

where s_i and s_j are senses of the same word in a given language.

Based on the probability of distinct lexicalization P_L , the levels of granularity for sense lexicalizations cross-linguistically are quantified; Resnik & Yarowsky conclude that all languages make robust distinctions on the homograph level 95% of the time; on the major sense level 78% of the time; and fine-grained level distinctions 52% of the time.

3.4.3 Polysemy and Sense Proximity in the SENSEVAL-2 Test Suite.

Irina Chugur, Julio Gonzalo, Felisa Verdejo

In an extension of the study by Resnik & Yarowsky, Chugur, Gonzalo & Verdejo investigate the possibility of characterizing sense inventories both quantitatively and qualitatively.[34] They address specific issues:

- What are the ways in which senses of a given word relate and what is the type of that relationship;
- How well are individual senses defined? Is it fine enough, coarse enough, etc.;
- How do such issues affect the evaluation of WSD systems?

Bearing these questions in mind, the authors describe the SENSEVAL 2 WordNet 1.7 subset. They characterize the ontology based on two parameters related to granularity. The first is fine-grainedness, namely, how specific are the sense distinctions, and is it possible for WSD systems to discriminate between senses. The second parameter examines the flip side of the previous parameter: are the sense definitions too coarse-grained?

To that end, the authors devise a complementary measure to the *Sense Proximity* measure defined by Resnik & Yarowsky. The measure is referred to as *Sense Stability*. Based on cross-lingual evidence, *Sense Stability* measures the likelihood that a pair

of occurrences for a word sense w_i receives the same translation for a language L , averaged over as many languages as possible. Quantitatively, *Sense Stability* is defined as

$$stability(w_i) \equiv \frac{1}{|languages||w_i|^2} \sum_{\substack{L \in \{languages\}, \\ x, y \in \{w_i examples\}}} tr_L(x), tr_L(y) \quad (3.4)$$

Based on Equation (3.4), coarse-grained senses have low stability as their contexts may lead to them lexicalizing differently across different languages.

Chugur et al. design an experiment to test the *Sense Stability* and *Proximity* of the words in WordNet 1.7 which are used in the SENSEVAL 2 exercise. They adopt the same experimental design as that of the Resnik & Yarowsky's human experiment described in the previous section. They have 11 native/bilingual speakers of four different languages who are asked to translate words marked in context to their native language. There are 508 short contexts for 182 senses of 44 words in the SENSEVAL test suite. In analyzing their results, they consider four different factors.

- **Language Family Distance**

Chugur et al. conclude that there is no significant sense difference across four languages utilized; they do not believe that adding more languages is critical for observing a stronger indicator of language distance.

- **Proximity and Stability**

The *Stability* and *Proximity* measures are integrated in a single matrix where the diagonal of the matrix is the *Stability* while the rest of the matrix cells are the pairwise *Proximity* measure values. They propose the integrated matrix of both measures as an evaluation criterion for SENSEVAL 2 systems.

- **Similarity and semantic relations between senses**

Four different semantic relations are examined: Homonymy, Metaphor, Specialization/Generalization, and Metonymy. Homonymy is a **no relation** case such as *bar - the law sense* and *bar - the unit sense*. Metaphor is a **similarity** case for instance, *child - the kid sense*. Specialization/generalization is a case of **extending or reducing the scope** of the original sense, for example, *fine- the greeting sense* and *fine - the ok sense*. Metonymy is a case of **semantic contiguity**, for instance, *yew - the tree sense* and *yew - the wood sense*.

Chugur et al. conclude that multilingual sense distinctions are reliable for homonyms. 27% of metaphors have a proximity of over 0.5, but multilingual distinctions are not sufficient indicators for them. Specialization/generalization behaved as expected with medium to high proximity. For metonymy, they conclude that multilingual evidence is a good first approximation but it is not sufficient as a sole criterion for metonymic distinctions.

- **Consistency of the data**

The Chugur et al. experiment suffers from very low inter annotator agreement, at 54% due to inconsistencies in the tagging by the participating subjects. Some annotators tag the same sense with different translation words due to variability in synonym sets in a particular language. Variable syntactic realizations cross-linguistically cause problems with the data; for instance, a noun modifying a noun in English becomes both an adjective and a noun in Russian. Such situations result in slightly variable forms of a unique root, however, the counting algorithm counts them as different translation collocations when words are part of complex expressions. Finally, they examine problems with the human annotations of the SENSEVAL-2 data.

All in all, the authors conclude that WordNet 1.7 is a good test bed for WSD systems. They confirm the conclusion drawn by Resnik and Yarowsky: that multilingual evidence is a good basis for sense disambiguation.

3.4.4 Cross-lingual Sense Discrimination: Can it work? *Nancy Ide*

In her study [32], Ide attempts to explore questions that arise from the proposal made by Resnik & Yarowsky with respect to sense definitions. She poses the questions of how many languages are sufficient to produce reliable sense distinctions? when do we know we have sufficient sense distinctions? how can we generate such sense distinctions from currently available resources? She acknowledges the limited usefulness of bilingual dictionaries owing to the lack of standards and the pervasive inconsistencies among them. She concludes by stating that parallel corpora are the optimal test bed for these ideas.

Ide conducts a manual experiment to investigate the feasibility of using parallel corpora for identifying distinct senses of polysemous words inasmuch as they lexicalize differently in five different languages. The parallel corpus is **George Orwell's Nineteen Eighty Four** translated from English into Slovene, Estonian, Romanian, and Czech. The languages pertain to four different language families: Germanic, Slavic, Finno-Ugrec, and Romance. The text comprises 100,000 words translated directly from the original English text. The corpus is sentence aligned.

For purposes of the experiment, she picks four words: *hard*, *line*, *country* and *head*. Parallel sentences with the words in question are extracted and sent to a linguist who is a native speaker of the language of translation. The task of the linguist is to identify the translation of the ambiguous English word in the translated sentence.

More than 85% of the English word occurrences have a corresponding lexical unit in any of the four translation language corpora. A manual association link is created for the English word and its translation with the WordNet a sense number.

A Coherence Index (CI) is devised to measure the extent to which a word in English is lexicalized differently in a translated text. Given a pair of senses for a word, the CI

is measured as follows:

$$CI(S_q S_r) = \frac{\sum_{i=1}^n S_{<q,r>^{(i)}}}{m_{S_q} m_{S_r} n} \quad (3.5)$$

where n is the number of comparison languages under investigation; m_{S_q} and m_{S_r} are the number of occurrences of sense S_q and sense S_r in the English corpus; $S_{<q,r>^{(i)}}$ is the number of times senses q and r are translated as the same lexical unit in a language i .

The value of CI is between 0 and 1, similar to the *Sense Stability* measure proposed by Chugur et al. described above, the higher the CI, the more coherent the senses, the lower the CI the more the senses lexicalize differently.

Ide considers language relatedness and distance among the different languages in her study. No significant impact is detected based on family relatedness.

Based on the CI values, Ide applies agglomerative clustering to the data to test if structures resembling dictionary entries emerge. She finds a strong correlation between the cluster-induced hierarchies and some dictionary entries for the words *hard* and *head* on a coarse-grain level. Accordingly, Ide concludes that translation can successfully be used as a filter for sense distinction.

3.4.5 Discussion

The crux of the current chapter builds on the ideas presented in the papers described above. All four studies exploit the different cross-linguistic sense lexicalizations for sense discrimination. Each of the studies comprises a manual exploration of the feasibility of using parallel corpora for sense discrimination. In this chapter, we devise a method that takes this core idea, expounds on it, and creates a practical demonstration, on a large scale, of its empirical feasibility and validity.

3.5 Hypothesis

Inspired by previous research described in chapter 2 and in the previous section, this investigation explores the relationship between translations of multiple instances of a polysemous word in a corpus. We emphasize two key observations:

- **Translation Distinction Observation (TDO)**

Senses of ambiguous words in one language are often translated into distinct words in a second language.

To exemplify TDO, we consider a sentence such as *I walked by the bank*. where the word *bank* is ambiguous with n senses. A translator may translate *bank* into

rive corresponding to the *geological formation* sense or to **banque** corresponding to the *financial institution* sense depending on the surrounding context of the given sentence. Essentially, translation has distinctly differentiated two of the possible senses of *bank*.

- **Foregrounding Observation (FGO)**

If two or more words are translated into the same word in a second language, then they often share some element of meaning.

FGO may be expressed in quantifiable terms as follows: if several words (w_1, w_2, \dots, w_x) in *L1* are translated into the same word form in *L2*, then (w_1, w_2, \dots, w_x) share some element of meaning which brings the corresponding relevant senses for each of these words to the foreground. For example, if the word **rive**, in French, translates in some instances in a corpus to *shore* and other instances to *bank*, then *shore* and *bank* share some meaning component that is highlighted by the fact that the translator chooses the same French word for their translation. The word **rive**, in this case, is referring to the concept of *land by a water side*, thereby making the corresponding senses in the English words more salient. It is important to note that the foregrounded senses of *bank* and *shore* are not necessarily identical, but they are the closest senses to one another among the various senses of both words.¹ Figure 3.1 below illustrates FGO.

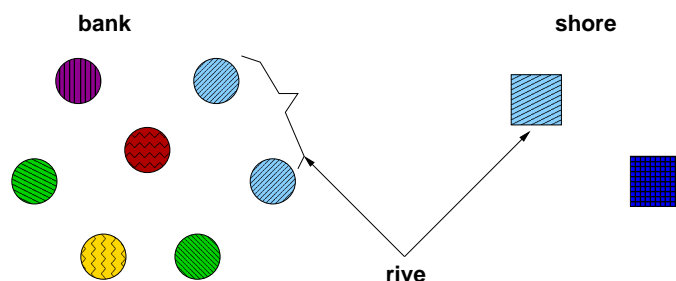


Figure 3.1: Common Senses Shared Between Polysemous Words

In Figure 3.1, the direction of the line fillings in the geometrical shapes is an indication of shared meaning characteristics between the senses of the words *bank* and *shore*. The difference in geometrical shape illustrates the fact that the close senses for the two words are not necessarily identical. As demonstrated in the diagram, the French word **rive** is the translational choice for both polysemous English words. In this diagram, **rive** highlights the shared meaning component for the two words in English; the shared semantic attribute is *water edge/geological formation*.

¹FGO as currently stated makes the implicit assumption that the *L2* word is not ambiguous. This assumption is fully explored in the discussion section.

Given observations TDO and FGO, the crux of the SALAAM approach aims to quantifiably exploit the translator’s implicit knowledge of sense representation cross-linguistically, in effect, reverse engineering a relevant part of the translation process.

3.5.1 General Hypothesis Statement

Given texts in translation with a source and target language, where a language is defined as a source language based on the fact that it has a sense inventory, we hypothesize that a target language word that is translated into distinct source language words serves as a good source of evidence for grouping the source language words.

Accordingly, in the current example, **rive** is a good source of evidence — anchor — for grouping the words *bank* and *shore*.

3.6 Method

3.6.1 General Method Description

Hypothetically, if the task of sense-annotating a parallel corpus (comprising a source and a target language) is manually attempted, where the annotator is tagging a polysemous source word with its corresponding target translation word, then the study requires the annotator’s knowledge of both source and target languages. S/he will create a mapping of words in L1 to words in L2. Accordingly, the words *bank* and *shore* are tagged with the word **rive**. Yet, tagging a source language with target words renders the annotations extremely corpus-specific. To achieve generality with the sense-tagging, we tag the corpus with an independent tag set from a source sense inventory.

For illustration purposes, we assume the source language is English and the target language is French. Furthermore, the existing sense inventory is in English, corresponding to the source language.²

Given a parallel corpus, a high-level view of the method is summarized in the following five steps:

1. Locate words in the English source corpus and their corresponding French target translations
2. Group source words that translate to the same target word orthographic form, thereby creating source groups

²In Diab and Resnik (2002), the naming of the corpora is reversed in accordance with the noisy channel naming convention. We decided to make it less confusing for the reader by following the more intuitive reading since the source is also linked to resource availability for SALAAM.

3. Measure the similarity among the different senses of the words in the source group based on their distance in a source language sense inventory
4. Assign the selected sense tags to the respective words in the corpus
5. Project the assigned sense tags from the source language words to the corresponding target language words in the parallel corpus

The first step in the preceding generic description would be a labor-intensive exercise if attempted manually. In order to automate the process on a large scale for parallel corpora, the need arises for a method that automatically discovers source-target word mappings (alignments).

Once the translational correspondences are discovered, grouping source words based on their translation to the same target word is directly applied.

Step 3 assumes the existence of a large independent sense inventory that is amenable to computational systems; moreover, it is assumed to have an associated quantified similarity measure between the words' senses. The similarity measure is used to calculate the similarity between the different source words' senses.

The closest senses resulting from the previous step are chosen for tagging the source words in the source groups.

Once the words in the source language are annotated with their appropriate sense tags, the tags are projected to their corresponding translations in the target corpus. Effectively, the sense tag assigned to the source word is the same sense tag projected onto the target word, thus creating a link for the target word from the translation language in the source inventory.

3.6.2 Required Resources

Our goal is to realize the described method automatically on a large scale. Therefore, two knowledge resources are required:

- Large amounts of text in translation are required, hence the need for a parallel corpus. Parallel corpora exist in myriad languages, for example, in religious books such as the Quran and the Bible [73], as well as in the UN Proceedings, and the Canadian Parliamentary Proceedings. Moreover, researchers have devised methods to mine the internet for large amounts of parallel corpora automatically with relatively minimal manual labor at high accuracy levels [71].
- A sense inventory is needed for the source language, where each word is represented with its/as its corresponding senses. This inventory is required for only one of the languages of the parallel corpus.³ The sense inventory should be rich enough to provide maximum coverage for the parallel corpus above.

³The translator in the context of a parallel corpus is trusted to have chosen the most faithful target lexical translation that conveys the sense of the source word by preserving the salient meaning element.

3.6.3 Detailed Method Description

As mentioned earlier, this approach, SALAAM, is unsupervised in that it does not rely on the availability of sense-annotated data for either language of the parallel corpus. Figure 3.2 provides a schematic view of the method followed by a detailed description of the individual processes. In the figure, each process is presented with an example on the right hand side.

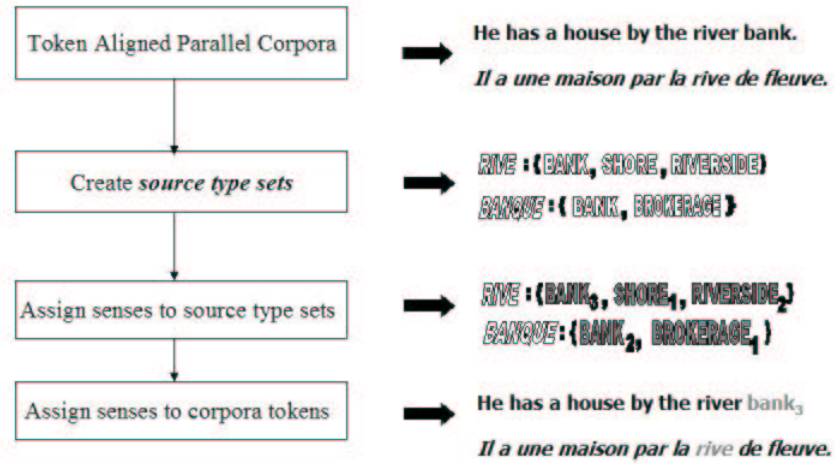


Figure 3.2: Flow chart demonstrating process flow in SALAAM method

The details of the schematic figure are as follows:

Word Align Parallel Corpus

SALAAM assumes the availability of token-aligned parallel corpora. A *token* is defined as a space delimited unit in a tokenized text. A token could be a number, a punctuation mark, a symbol or a word instance. Alignment is the process of discovering the translational mappings of token instances between source and target languages in a parallel corpus. A token instance is a unique occurrence of a token in a corpus. Figure 3.3 illustrates an example of token-aligned text expected as input by the algorithm.

Every token in the source corpus is aligned to some token or set of tokens in the target corpus. One-to-many alignments, in most cases, arise due to lexicalization divergences. A source token may align with the NULL token, which is the empty token indicating the non-existence of an appropriate alignment token on the target side.

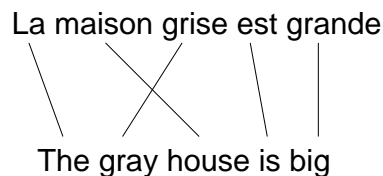


Figure 3.3: A sample token alignment in a parallel corpus

Create Source Type Sets

The process of creating source type sets involves the following steps:

- **Identify Aligned Tokens**

Figure 3.4 illustrates the alignment of target French token instances to English source token instances. An instance of the French token **rive** aligns with the source token instance *bank*; the second French token instance **rive**, in the figure, aligns with the source token instance *shore*; similarly, target token instances of **banque** align with source tokens *bank* and *repository*; the dots indicate running text.

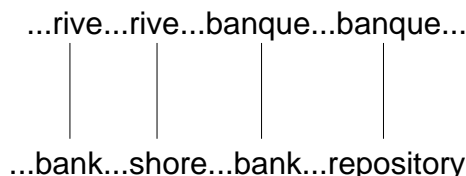


Figure 3.4: Tokens aligned in a parallel corpus

Figure 3.5 shows different source token instances of *bank* and *shore* which align with instances of the French token **rive**; likewise, the figure illustrates source token instances of *bank* and *repository* aligning with instances of **banque**. The numbers in the figure demonstrate the process of bookkeeping the information for corpus location and occurrence. For example, **rive#7#1#27** is the token instance of the target word **rive**, where 7 is a line identification number in the corpus, 1 is **rive**'s location in the line — all token instances in a line are numbered from 0 to n — and 27 is the frequency of occurrence of the token **rive** in the target corpus. It is worth noting that token alignment is between source and target lines with the same identification number, therefore, in Figure 3.5, the line identification number is the same for target and source token instances for all the listed pairs.

Given a parallel corpus, it is not always the case that a line or a sentence on the source side will correspond to a line or sentence on the target side. In many cases, we find a sentence on the source side corresponding to multiple sentences

on the target side or vice versa. Several researchers devise automatic approaches for automatically discovering sentence alignment in parallel corpora [55]. Sentence or line alignment is an interesting problem but it is outside the scope of the current research. SALAAM assumes that the source and target corpora are line aligned (sentence aligned).

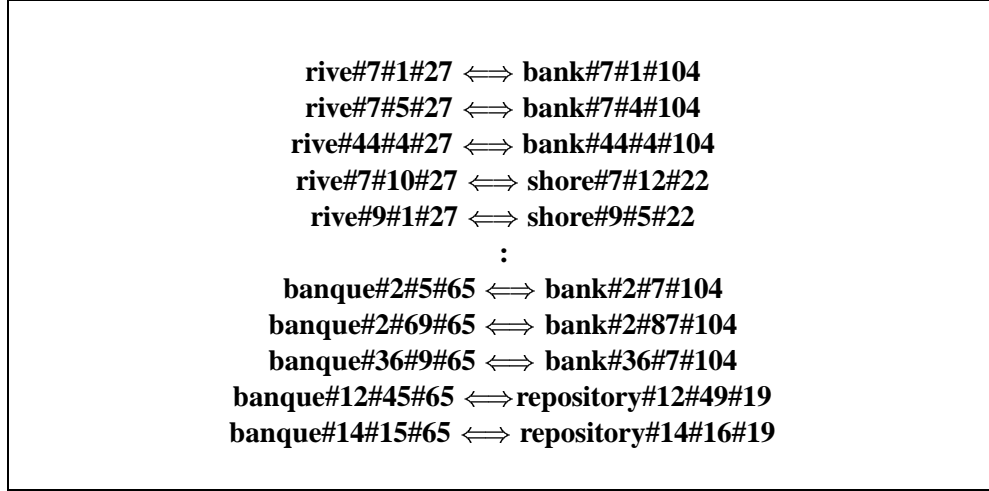


Figure 3.5: Aligned token instances from target to source

- **Conflating Alignments**

Once the parallel corpus is token-aligned as shown in Figure 3.5, the target token instances are conflated into target word types. Accordingly, {**rive#7#1#27**, **rive#7#5#27**, **rive#44#4#27**, **rive#7#10#27**, **rive#9#1#27**} are conflated into the target word type RIVE and, similarly, {**banque#2#5#65**, **banque#2#69#65**, **banque#36#9#65**, **banque#12#45#65**, **banque#14#15#65**} are conflated into the target word type BANQUE. All instances of source tokens that align with the same target word type are grouped in a source token set as illustrated in Figure 3.6. All the source token instances aligned with **rive** target token instances are grouped to form a source token set corresponding to the target word type RIVE and those aligned with **banque** are grouped to form the source token set corresponding to the target word type BANQUE.

In order to create **source type sets**, the source token instances are conflated into source type words, in a similar manner to the conflation of target tokens into target types, source tokens undergo the same process. For example, Figure 3.7 illustrates the source type sets for RIVE and BANQUE.

In the Figure 3.7, the word tokens *bank* are conflated to form the source word type BANK and likewise for SHORE and REPOSITORY.

- **Sense Assignment to Source Type Sets**

RIVE:{bank#7#1#104, bank#7#4#104, bank#44#4#104, shore#7#12#22,
shore#9#5#22}

BANQUE: {bank#2#7#104, bank#2#87#104, bank#36#7#104,
repository#12#49#19, repository#14#16#19}

Figure 3.6: Target word types and their corresponding source token sets

RIVE: {**BANK, SHORE**}
BANQUE: {**BANK, REPOSITORY**}

Figure 3.7: Source type sets for the target words RIVE and BANQUE

A distance metric is defined to measure the similarity between the senses of the source word types in the source type sets. A similarity function $sim(w_x, w_y)$, where sim calculates the distance between all the senses of word (w_x) and word (w_y) , for all word pairs in the source set. The goal is to maximize the overall similarity among the word senses across the source word types in the source type set. The resulting similarity measure is $max(sim(w_x, w_y))$ which is an optimization function; it chooses the senses that are most similar among all the senses of all the words in a given source set. Given the source set for RIVE as {BANK, SHORE}, all the senses corresponding to the two words in a sense inventory are compared and the ones that are the most similar according to the defined similarity measure are chosen as the appropriate tags for the respective word types.

For illustration, if we look up the words BANK and SHORE in the Collins Cobuild Dictionary [76], we find five nominal senses listed for BANK and two for SHORE. Accordingly, the sim function computes 2×5 comparisons, each comparison resulting in a similarity value. The sense tags that yield the highest similarity value are assigned to their corresponding word types. In fact, more than one tag may score the highest sim value.

For illustration, the five senses listed for BANK are:

1. *a bank is an institution where people or businesses can keep their money*
2. *the bank in a gambling game is the money that belongs to the dealer to the casino management*
3. *a bank is the raised ground along the edge of a river or a lake*

4. *a bank of something such as computer data or blood is a store of it that is kept ready for use when needed*
5. *a bank of switches, keys, etc., on a machine*

The two senses listed for SHORE are:

1. *the shore of a sea, lake or wide river is the land along the edge of it*
2. *a particular country with a coastline is sometimes referred to in literary English as the shores of the country*

The senses listed for REPOSITORY are:

1. *a person or a group of people who you can rely on to look after something important*
2. *a place you can keep objects of a particular kind*

By inspecting the definitions of the different sense entries for the source word types BANK and SHORE, we see that sense #3 of BANK and sense #1 of SHORE are the most similar among the different possible pairings of senses. We judge them to be similar based on the proximity in the meanings of the definitions rendered. Therefore, the source word types BANK and SHORE are assigned those senses, respectively. In this phase, the role of the similarity measure is to produce quantitative values for the distance between the different senses of the different words. In quantified terms, the similarity function just utilized is nothing but a computation of the overlap between the content words that make up the sense definitions [43]. The choice of senses for tagging is based on setting a sense selection criterion. For the given example, the selection criterion is set to the senses that have a maximum overlap in the words of the sense definitions. Based on the chosen sense definitions, the salient meaning element shared is *land by the water edge*. Similarly, for the source words BANK and REPOSITORY in correspondence with the target word BANQUE; BANK is assigned its sense #4 and REPOSITORY is assigned its sense #2. In this case, the salient meaning component is *a place to keep objects of a kind*. The resulting source type tag set is illustrated in Figure 3.8, as well as the senses propagated to the token instances corresponding to the word types as illustrated in Figure 3.9. In the figure, the subscripts indicate sense numbers.

It is extremely important to note that source type sets have to have at least two members in the set in order to apply a similarity function among word senses, i.e. by definition, a similarity function applies to a minimum of two items. Therefore, it is crucial to highlight the significance of variability in alignment. To illustrate, if throughout the parallel corpus, all instances of the French target

RIVE: ($BANK_3, SHORE_1$)
 BANQUE: ($BANK_4, REPOSITORY_2$)

Figure 3.8: Sense Tagged Type Source Sets

RIVE: { $bank_3\#7\#1\#43, bank_3\#7\#4\#343, bank_3\#44\#4\#173$
 $shore_1\#7\#10\#61, shore_1\#9\#1\#121$ }
BANQUE: { $bank_4\#2\#7\#44, bank_4\#2\#87\#64,$
 $bank_4\#36\#7\#194, repository_2\#12\#45\#92 repository_2\#14\#15\#342$ }

Figure 3.9: Sense Tagged Token Source Sets

word instances **rive** align with source instances *shore*, the resulting source type set, after conflation, will have a single word type SHORE which cannot be submitted to the similarity function, consequently, neither instances of *shore* nor the corresponding target instances of **rive** are assigned sense tags.

- **Project Source Sense Tags to Target Tokens**

Finally, source sense tags assigned to source tokens from the source sense inventory are projected onto target language corpus tokens, which is a direct mapping step.

$rive_{BANK_3}\#7\#1\#6 \iff bank_3\#7\#1\#43$
 $rive_{BANK_3}\#7\#4\#38 \iff bank_3\#7\#4\#343$
 $rive_{BANK_3}\#44\#4\#18 \iff bank_3\#44\#4\#173$
 $rives_{SHORE_1}\#7\#10\#6 \iff shore_1\#7\#10\#61$
 $rives_{SHORE_1}\#9\#1\#13 \iff shore_1\#9\#1\#121$
 \vdots
 $banque_{BANK_4}\#2\#7\#5 \iff bank_4\#2\#7\#44$
 $banque_{BANK_4}\#2\#87\#8 \iff bank_4\#2\#87\#64$
 $banque_{BANK_4}\#36\#7\#21 \iff bank_4\#36\#7\#194$
 $banque_{REPOSITORY_2}\#12\#45\#7 \iff repository_2\#12\#45\#92$
 $banque_{REPOSITORY_2}\#14\#15\#6 \iff repository_2\#14\#15\#42$

Figure 3.10: Projecting source inventory senses onto target language instances

In Figure 3.10, instances of **rive** and **banque** are assigned the senses corresponding to the source language sense inventory entries indicated by the subscripts,

thereby creating links for the French words in the Collins Cobuild Dictionary.

3.6.4 Evaluation Metrics

1. Precision (P)

A measure of accuracy for sense tagging where the tags resulting from SALAAM are evaluated against a predefined gold standard set. Quantitatively, *Precision* is measured as follows:

$$\text{Precision (P)} = \frac{|\text{correct tags}|}{|\text{items tagged}|} \quad (3.6)$$

2. Recall (R)

A measure of the retrieval capacity of a system where the tags resulting from SALAAM are evaluated against a predefined gold standard set. Quantitatively, *Recall* is measured as follows:

$$\text{Recall (R)} = \frac{|\text{correct tags}|}{|\text{items in gold standard}|} \quad (3.7)$$

3. F-Measure (FM)

This is an Information Retrieval measure which is a summary measure of *Precision* and *Recall*. Quantitatively, *F-Measure* is measured as follows:

$$\text{F-Measure (FM)} = \frac{2PR}{R + P} \quad (3.8)$$

4. Coverage (COV)

This is a measure of the number of items attempted by SALAAM out of the possible items in the gold standard. *Coverage* is measured as follows:

$$\text{Coverage (COV)} = \frac{|\text{gold standard set items tagged}|}{|\text{items in gold standard}|} \quad (3.9)$$

5. Zscore Significance Test (Z)

This is a two-tail significance test of difference between two proportions.⁴ The significance level is set to 95%, i.e., a test is significant if $Z < -1.96$ or $Z > 1.96$

⁴<http://franz.stat.wisc.edu/rossini/courses/introbiomed>

3.7 Evaluation

In order to formally evaluate SALAAM for English word sense tagging, we need four different components:

1. A parallel corpus with English on one side as the source language. The corpus needs to be large enough for training stochastic translation models for the automatic discovery of token mappings — translation alignments. Moreover, the corpus has to exhibit enough variability in order to render the similarity measure operational. Accordingly, the need arises for a balanced corpus. A balanced corpus is defined as a corpus that has equivalent amounts of data pertaining to diverse topics.
2. A broad coverage sense inventory for the English source language
3. A hand-annotated subset of the corpus to provide a gold standard for evaluation
4. Performance figures for other systems on the same task, evaluated against the same gold standard

Acquiring all four components simultaneously proves to be a challenge. To our knowledge, there are no balanced parallel corpora with a hand-annotated gold standard. On the other hand, the few hand-annotated sets available do not exist for parallel corpora. Then, we pose the question: Which is more feasible, translating a corpus that has an associated gold standard or hand-annotating a portion of the English side of a parallel corpus? Given how involved the process of hand-annotating a corpus is, we opt for the former solution of translating a parallel corpus that has an associated gold standard set.

SENSEVAL

The requirement for a hand-annotated set as a gold standard which also is used for evaluating other WSD systems is met through the SENSEVAL 2 exercise English All Words task.⁵ (See SENSEVAL era Section in Chapter 2)

In SENSEVAL 2, the English ontology is WordNet 1.7 pre (WN17pre).

3.7.1 Materials

Ontology

Like previous WordNet editions [24], WN17pre is a computational semantic lexicon for English. It is rapidly becoming the community standard lexical resource for English since it is freely available for academic research. It is an enumerative lexicon that

⁵<http://www.senseval.org>

combines the knowledge found in traditional dictionaries in a Quillian (1968) style semantic network [65]. Words are represented as concepts, referred to as synsets, that are connected via different types of relations such as hyponymy, hypernymy, synonymy, meronymy, antonymy, etc. Words are represented as their synsets in the lexicon. For example, the word *bank* has 10 synsets in WN17pre corresponding to 10 different senses. The concepts are organized taxonomically in a hierarchical structure with the more abstract or broader concepts at the top of the tree and the specific concepts toward the bottom of the tree. Accordingly, the concept FOOD is the hypernym of the concept FRUIT, for instance.

Similar to previous WordNet taxonomies, WN17pre comprises four databases for the four major parts of speech in language: nouns, verbs, adjectives, and adverbs. The nouns database consists of 69K concepts and has a depth of 15 nodes. The nouns database is the richest of the 4 databases. Majority of concepts are connected via the IS-A identity relation. In this chapter, we focus on nouns only.⁶ An excerpt of the noun database for WN17pre is shown in Figure 3.11 below.

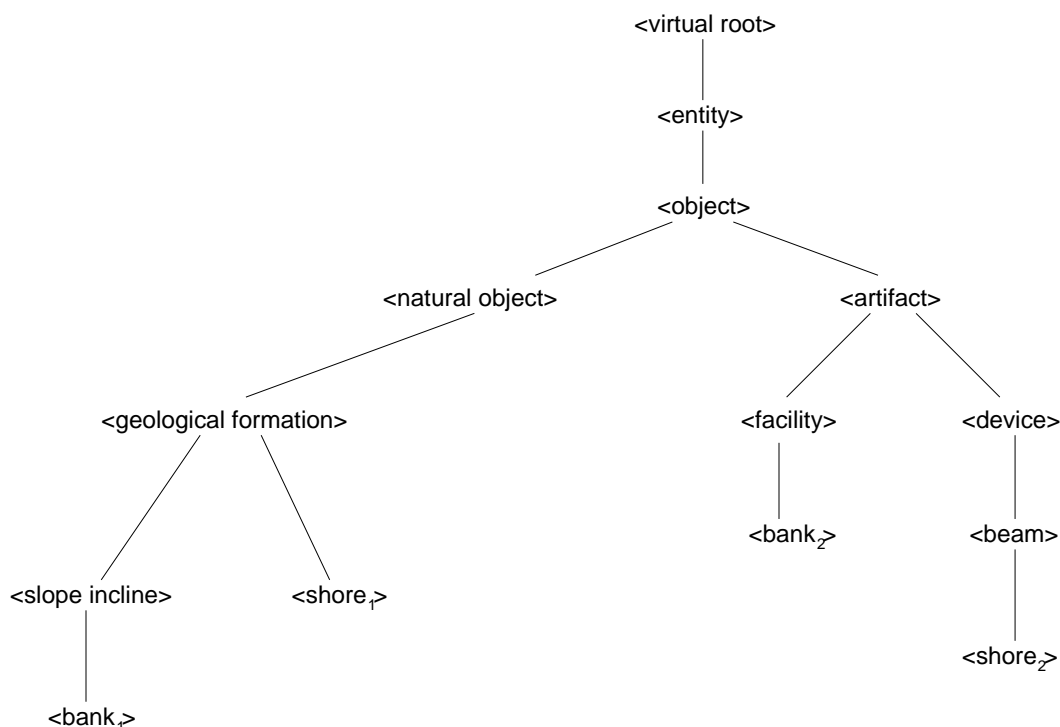


Figure 3.11: An excerpt from the noun database of WN17pre

In the figure, the dotted lines indicate several nodes omitted for space consideration and subscripts indicate sense numbers.

⁶Nothing inherent to SALAAM restricts it to a specific part of speech.

Gold Standard

To evaluate the performance of SALAAM, we use the SENSEVAL 2 English All Words tag set as a gold standard. The gold standard is manually annotated with WN17pre by the organizers of the SENSEVAL 2 exercise. In this chapter, we are only interested in the nouns in the set. The nouns in the gold standard are annotated with one or more senses from WN17pre. Some nouns are tagged with P and others with U, where the P tag indicates proper nouns, and the U tag indicates unassignables, where the annotator could not find the appropriate sense in the list of WordNet senses. The gold standard for this evaluation comprises 1071 nouns after excluding instances annotated with U or/and P tags.

Test Set

The nouns in SENSEVAL 2 English All Words (SV2AW) test corpus constitute the test set for this evaluation. SV2AW comprises 3 articles from The Wall Street Journal amounting to a total of 242 lines and 5815 tokens. The articles discuss three topics: culture, medicine and education.

Corpora

SV2AW is a very small corpus for SALAAM to be applied. First off, for stochastic token alignment, the need arises for a large corpus to ensure reliable alignment results. Moreover, variability in contexts is essential to produce source type sets that have several members. Therefore, the test corpus needs to be augmented with a large enough parallel corpus in order to ensure two factors: Reliable alignment quality and variability in contexts. Accordingly, SV2AW is augmented with four corpora that are deemed balanced. The corpora are described as follows:

1. The Brown Corpus of American English (BC):[26]

BC comprises articles from specialized scientific journals, novel excerpts and news articles as well as non-fiction work. It is a balanced corpus of approximately one million words.

2. The SENSEVAL 1 Trial, Training and Test corpus (SV1) [40]

SV1 comprises excerpts from the following different corpora: The Wall Street Journal, which is mainly news articles; The British National Corpus, a balanced corpus of roughly 100 million words of different genres, similar to BC but in British English; and IBM technical manuals. All in all, SV1 amounts to 1.5 million tokens.

Corpora	Lines	Tokens
BC-SV1	101841	2498405
SV2-LS	74552	1760522
WSJ	49679	1290297
SV2AW	242	5815
<i>Total</i>	<i>226314</i>	<i>5555039</i>

Table 3.1: Relative sizes of corpora used for evaluating SALAAM on SV2AW test set

3. The SENSEVAL 2 Lexical Sample trial, training and test corpus (SV2-LS) ⁷

This corpus is similar to the SV1 corpus in constitution. It comprises 1.76 million tokens.

4. The Wall Street Journal (WSJ) corpus

WSJ comprises sections 18-24 of the Penn Tree Bank. This corpus has 1.29 million tokens. The WSJ mainly contains news articles.

The relative sizes of the four corpora used for augmentation listed above as well as the test corpus are illustrated in Table 3.1.

None of the corpora exists in translation. Resorting to human translators would have been an ideal solution but considering the expense and time factors, we opt for off-the-shelf machine translation (MT) systems to do the job. We use commercially available MT systems as an approximation (pseudo-translation) [18]. The process of pseudo translation is appealing from several angles: It is cheap and fast to produce large amounts of translated data in a reasonable amount of time; one could use several MT systems for several languages. Accordingly, we pseudo-translate the four augmenting corpora as well as the test corpus into 3 different languages: Arabic, French and Spanish. We use two machine translation systems per language. For Arabic, we use two machine translation systems available on the Web, Al-Misbar (AM⁸), and Tarjim (TR⁹). For French and Spanish, we use two MT systems: Global Link 6.4. Pro (GL) and Systran Professional Premium 2.0 (SYS). The process of pseudo translation results in six parallel corpora, two for each language. The choice of languages is mainly influenced by the claimed quality of translations for both GL and SYS in French and Spanish. Moreover, EuroWordNet exists for both French and Spanish, and could later serve as a test bed for the projected tagging on the target language side of the parallel corpus. As for Arabic, the choice is mainly

⁷<http://www.senseval.org/>

⁸URL is <http://www.almisbar.com>

⁹URL is <http://www.tarjim.com>

because of its distance from English. As a Semitic language, Arabic is farther from English than Spanish and French which are both Latin based. Moreover, Arabic, by many standards, is considered a low density language which creates a realistic test case for SALAAM.

3.7.2 Tools

- **Part Of Speech (POS) Tagger**

There are no inherent constraints within SALAAM for a specific POS. But in order to constrain the search space in the sense inventory for this evaluation, we restrict the POS to nouns. Both the BC corpus and the test corpus SV2AW are manually POS tagged. The rest of the corpora are tagged using the Brill POS Tagger [5]. The Brill POS Tagger is trained on the manually POS tagged BC.

- **Tokenization**

In this evaluation of SALAAM, we process four languages: English, Arabic, French and Spanish. For the English and French corpora, we use the tokenizer provided by Dan Melamed (personal communication) with some modifications. For Spanish, we use a tokenizer developed by Nizar Habash and Bonnie Dorr (personal communication), with some modifications. As for the Arabic corpora, we created a simple stemmer/tokenizer which based on standard regular pattern matching. The Arabic text is first transliterated into Latin script, then the tokenization separates out suffixes and prefixes. In Arabic, suffixes are typically pronouns and prefixes are usually prepositions or articles ¹⁰

- **Stochastic Alignment Tool: GIZA++**

SALAAM assumes token aligned corpora as input. However, since the field of alignment is still in its early stages, token aligned parallel corpora that meet the specific requirements for SALAAM do not exist. Therefore, assuming the parallel corpora are sentence aligned, we use an automated token alignment system, the GIZA++ package [62]. GIZA++ is part of the EGYPT statistical machine translation package [2]. GIZA++ is an implementation of IBM models 1-5 [8]. The models are trained in succession where each of these models produces a Viterbi alignment. The final parameter values from one model are used as the starting parameters for the next model. Given a source and target pair of aligned sentences, GIZA++ produces the most probable token-level alignments. Multiple token alignments are allowed on the source language side, i.e., a token in

¹⁰The tokenization is intentionally kept at a minimum in order to maintain a comparative base among the three target languages while simultaneously making minimum assumptions as far as the target language requirements are concerned.

English may align with multiple tokens on the French side. Tokens on either side of the parallel corpus may align with an empty token indicated by the NULL token.

3.7.3 Sense Selection and Similarity Measure

As described earlier in Section 3.6.3, a similarity measure is needed to determine the quantitative distance between the senses of the words in question. For the purposes of this evaluation, we use *Noun_Groupings* (NG) distance measure for calculating the similarity values between the different senses in a source type set. NG is an algorithm proposed and implemented by Resnik [70]. The algorithm is an optimization function.

Given a source type set of words in English, NG calculates the pairwise similarity across all senses of the words in the source set and assigns the highest confidence scores to those senses that are the closest in the set. The confidence scores range from 0 to 1. At the core of NG is an information theoretic similarity measure devised by Resnik [72]. Given a taxonomy of concepts and frequencies of words in a large corpus, Resnik's similarity measure calculates the distance between two concepts as:

$$sim_{info}(c_1, c_2) = max_{c \in S(c_1, c_2)} [-logp(c)] \quad (3.10)$$

where $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2 , and where $-logp(c)$ is the information content of node c . $p(c)$ is estimated by observing frequencies in a corpus.

Therefore, the quantity defined in this similarity measure, calculates the information content of all the nodes that subsume two synsets and returns the one with the maximum information content as well as the two senses that are the closest. Intuitively, if two senses are not similar, then the information content returned will be very small indicating that the most informative subsumer is very high up in the taxonomy. In the worst case, where there is no similarity at all, the information content is 0 pertaining to the top node in the hierarchy. NG provides the quantitative values of the distance between the different senses of the words in the source type sets but it does not choose the appropriate senses.

Consequently, in SALAAM, we devise a sense selection criterion threshold to choose the appropriate senses to assign to the words in the source type sets.

3.7.4 Development and Testing Environment

The SALAAM system is developed and tested on a Sun Blade 1000 with 1GB of RAM running OS Solaris 2.8. The system is written in C and Perl.

3.7.5 Evaluation Measure

In this chapter, we use the SENSEVAL 2 scoring program `scorer2`.¹¹ `scorer2` is an implementation by Cotton [13] of the Melamed and Resnik [56] metric for sense disambiguation evaluation. The measure is a principled metric for tagger evaluation given hierarchical tag sets. This measure rewards a system by a 1 if the sense assigned is correct, a 0 if it is completely incorrect. This measure is different from traditional measures as it allows room for partial credit. Therefore, if a WSD system assigns an incorrect sense to a polysemous word but this sense is a sibling and a direct descendent of the same hypernym of the correct sense, the WSD system is rewarded with partial credit. `scorer2` can report results in three different modes: Fine grain mode, coarse grain mode; and mixed grained mode. The fine-grain mode is the strictest evaluation metric.

All the results reported here are evaluated using `scorer2` in the fine-grain mode. We are not reporting an evaluation against a baseline since none was used for the official evaluation in **SENSEVAL2** exercise.

3.7.6 Evaluation Parameters

- **Part of Speech**

As mentioned earlier, SALAAM has no inherent constraints on POS tag. However, for the purposes of this evaluation, we set the POS to nouns.

- **Stop Word List**

Another parameter is the removal of closed word items from the alignments since they are a source of noise. We use a stop word list that contains mainly punctuation, prepositions and articles in the source language English.

- **GIZA++ Parameters**

The parameters are set as follows: 20 iterations for model 1, 10 iterations for HMM and 20 iterations for model 4.¹² The maximum sentence length on either or both sides of a parallel corpus is set to 70 tokens.

3.7.7 Evaluation Factors

- **Different Target Languages**

We have three different languages for this evaluation, Arabic, French, and Spanish.

¹¹We use the version fixed by Rada Mihalcea <http://www.senseval.org>

¹²Models 2 and 3 are eliminated from the alignment based on the discussion in [62]; HMM and model 4 essentially replace models 2 and 3. Model 5 is excluded owing to the excessive time requirements.

- **Different MT systems**

Each language is pseudo-translated using two different MT systems.

- **Sense Selection Criterion**

NG assigns a confidence score to each word sense when calculating the similarity between the different words in the source type set. If NG is not confident of the scores it typically divides the 1.0 confidence score among all the senses of a given word, yielding a uniform confidence distribution. Consequently, we have two sense selection thresholds

1. **MAX**

The sense tag(s) with the highest confidence score is(are) selected. There is no minimum confidence score threshold.

2. **THRESH**

The sense tag(s) with the highest confidence score is(are) selected. A minimum threshold of (> 0.5) is set.

3.7.8 Evaluation Conditions

Based on the three factors, we devise several experimental conditions. In all the conditions, in accordance with the method description in Section 3.6.3, the target language is used as the source of evidence to create the source type sets.

The first condition describes the default set of conditions. This is followed by a set of conditions where the output of MT systems for the same target language is merged pre-alignment or post-alignment. The idea behind merging the output of two pseudo-translations is to maximize the translation variability assuming that two different MT systems most likely use different knowledge bases for the translation process. We conclude with a set of conditions where the output of applying SALAAM to the parallel corpora pertaining to different languages is merged in different modes. The impetus for such a merge is to test to what extent evidence from different languages aids the performance of the SALAAM tagging system.

1. **Default conditions: AR-TR, AR-AM, FR-GL, FR-SYS, SP-GL, SP-SYS**

We have six default conditions which result from the intersection of factors 3.7.7 and 3.7.7 at sense selection criterion MAX as described in Factor 3.7.7.1. The six conditions are FR-GL, FR-SYS, for evaluating the English test set resulting from applying SALAAM to the English-French parallel corpus yielded from pseudo-translating the test corpus and the augmenting corpus using the GL MT system and the SYS MT system, respectively; SP-GL and SP-SYS for evaluating the results obtained applying SALAAM given the English-Spanish parallel corpus when pseudo-translated using GL and SYS, respectively; and similarly,

AR-AM and AR-TR for Arabic where the English source corpora are pseudo-translated using AM and TR, respectively.

2. Intralanguage pre-alignment merge with MAX sense selection criterion: FR-GLSYS

Translation resulting from the two MT systems are interleaved where translations of the English odd lines are translated using the GL MT system and the even lines are translated using the SYS MT system. This condition is only applied to the French pseudo-translated corpora. The sense selection criterion is set to MAX

3. Intralanguage post-alignment merge with MAX sense selection criterion: GLSYS-FR_M, GLSYS-SP_M, AMTR-AR_M

Source and target token alignments resulting from GIZA++ are merged prior to submission to the NG algorithm for calculating the similarities. This condition, as mentioned before, aims at maximizing variability in the source type sets. The sense selection criterion is set to MAX.

4. Intralanguage post-alignment merge with THRESH sense selection criterion: GLSYS-FR_T, GLSYS-SP_T, AMTR-AR_T

Similar to Condition 3 but the sense selection criterion is set to THRESH as described in Factor 3.7.7.2.

5. Pairwise and three-way interlanguage intersection at THRESH sense selection criterion: AR-FR_I_T, AR-SP_I_T, FR-SP_I_T, AR-FR-SP_I_T

In this set of conditions, we evaluate the results of intersecting the SALAAM tag sets resulting from evidence obtained from two and three target languages, respectively, with the MAX sense selection criterion. This merge occurs after the application of the NG algorithm and the sense assignment to the nouns in the test set. The intersection mode is where only the senses that overlap for the commonly tagged noun instances are kept in the final tag set. Unique noun instances for each language are also kept in the final tag set. In the pairwise conditions, we intersect the tag set that results from Condition 4 for two languages at a time, therefore, AR-FR_I_T is the intersection of the tag sets resulting from conditions AMTR-AR_T and GLSYS-FR_T. In the three-way condition, AR-FR-SP_I_T, the same process is applied but with the tag sets resulting from all three languages in Condition 4.

6. Pairwise and three-way interlanguage union at THRESH sense selection criterion: AR-FR_U_T, AR-SP_U_T, FR-SP_U_T, AR-FR-SP_U_T

In this set of conditions, we evaluate the results of union merging the SALAAM tag sets resulting from evidence obtained from two and three target languages,

respectively, with the THRESH sense selection criterion. This merge occurs after the application of the NG algorithm and the sense assignment to the nouns in the test set. The union mode is where all the senses that are assigned to the commonly tagged noun instances are kept in the final tag set. Unique noun instances for each language are also kept in the final tag set. In the pairwise conditions, we union merge the tag set that results from condition 4 for two languages at a time, therefore, AR-FR_U_M is the union of the tag sets resulting from conditions AMTR-AR_T and GLSYS-FR_T. In the three-way condition, AR-FR-SP_U_M, the same process is applied but with the tag sets resulting from all three languages in Condition 4.

3.7.9 Experimental Hypotheses

We have the following experimental hypotheses corresponding to the eight experimental conditions:

1. **Hypothesis 1**

SALAAM exploits translation evidence in the default condition 1 yielding comparable FM to state-of-the-art unsupervised WSD systems.

2. **Hypothesis 2**

SALAAM applied in Condition 2 yields improved precision when compared to default Condition 1 since it increases the variability in the source type sets, however recall is comparable to recall results obtained using the two pseudo-translations, independently, since this condition, FR-GLSYS, has only half of each pseudo-translation.

3. **Hypothesis 3**

SALAAM applied in Condition 3 improves FM over SALAAM in the default Condition 1 because of the increase in variability in the source type sets.

4. **Hypothesis 4**

SALAAM applied in Condition 4 improves precision, P, over the performance of SALAAM applied in Condition 3 as the higher sense selection threshold, THRESH, weeds out senses where the NG yields a uniform confidence score distribution indicating the lack of a bias in the similarity measure toward any of the senses involved in the sense similarity calculation. In short, removing the noise from the final tag set.

5. **Hypothesis 5**

SALAAM applied in Condition 5 significantly improves precision results over results obtained by SALAAM when applied in condition 4, as Condition 5 is exclusive merging of evidence from two languages which are themselves merges

of two pseudo translations at the THRESH sense selection criterion. When the output tag sets are intersected, the precision increases as the tag set is further refined by the intersection process but the recall decreases as some valid senses might be weeded out if they are not shared across the tag sets of the two or more languages merged.

6. Hypothesis 6

Recall values are improved when SALAAM is applied in condition 6 over recall yielded by Condition 4 since it is the union of multiple tag sets pertaining to several languages, thereby including evidence from several languages which most likely cover different portions of the data. Accordingly, we expect evidence from Arabic and any other language to yield the better results, therefore, both the coverage and the performance of SALAAM as measured by FM improves when applied in Condition 6 over performance of SALAAM in Condition 5 as the THRESH selection criterion removes the noise from the source type sets.

3.8 Results

In this section, we present the results of applying SALAAM in the different experimental conditions presented in Section 3.7.8 to evaluate English source language tagging of nouns in the test corpus SV2AW when evaluated against a hand tagged gold standard. The following subsections correspond to the different hypotheses described in Section 3.5.

3.8.1 Testing Hypothesis 1

We have six experimental conditions in the default condition 1 corresponding to six parallel corpora. Table 3.2 illustrates the results obtained by applying SALAAM in these default conditions where the source of tagging evidence is from a single pseudo-translated target language using a single MT system.

The performance scores depicted in Table 3.2 are evaluated using the `scorer2` software in the fine-grain mode.

Figure 3.12 illustrates the relative performance of SALAAM against state-of-the-art WSD systems on the same task of sense tagging nouns.¹³ All these WSD systems participated in the SENSEVAL 2 English All Words task. The X-axis is the precision percentage and Y-axis is recall percentage.

In Figure 3.12, supervised systems are presented as gray filled diamonds, unsupervised systems as hollow triangles, gray squares are partially supervised systems,¹⁴ and

¹³The nouns are isolated from the submitted tag set of the different systems and evaluated using `scorer2` in the fine grain mode.

¹⁴The classification into supervised, unsupervised, and partially supervised system is based on the

Condition	P%	R%	COV%	FM
FR-GL	58.1	50.9	87.62	54.26
FR-SYS	58	49	84.43	53.12
SP-GL	57.9	48.6	83.86	52.84
SP-SYS	60	51.5	85.93	55.43
AR-TR	58.3	51.4	88.27	54.63
AR-AM	57.5	49.3	85.74	53.09

Table 3.2: SALAAM performance results on English source SV2AW test data in the default conditions

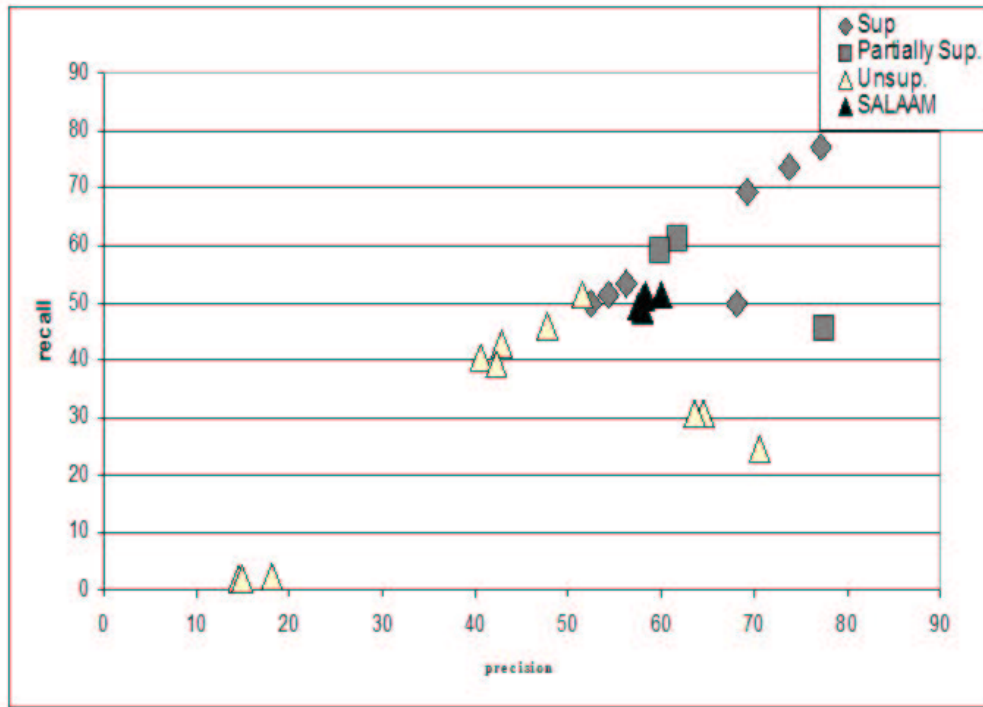


Figure 3.12: SALAAM performance precision & recall results in the default conditions plotted against state-of-the-art WSD systems on the same test set SV2AW

finally black triangles are the results obtained by SALAAM. Unsupervised systems do not rely on annotated data in the process of sense tagging. Supervised systems depend directly on WN17pre sense-tagged training data. Partially supervised systems rely on hand-annotated data from other sources such as the DSO corpus or they back-off to the most frequent sense in WN17pre in case the WSD system could not make a guess.¹⁵

All the results obtained by SALAAM are comparable on both precision and recall performance, i.e., they do not significantly differ from one another according to the Z significance test at $p < 0.05$. SALAAM performance is comparable performance to state-of-the-art WSD systems. None of the unsupervised systems is better than any of the SALAAM conditions on both precision and recall simultaneously. In fact, all SALAAM default conditions are significantly higher than all the unsupervised systems on both precision and recall except for one system which is significantly higher than SALAAM conditions on recall. Three systems are better than SALAAM on P but significantly lower on R. It is also worth noting that the majority of the systems — including all supervised systems — have close to 100% coverage while the highest coverage achieved by SALAAM is 88.27% in the AR-TR condition. This issue is further discussed in Section 3.9 of this chapter.

The FM measure provides a principled way to view and compare the performance of SALAAM against the other WSD systems. Figure 3.13 shows the FM scores obtained by SALAAM and the systems participating in the SENSEVAL 2 All Words task for English. The six SALAAM default conditions are depicted as the black bars in the figure; solid gray bars are supervised systems; the checkered bars are partially supervised systems; and the hollow bars are unsupervised systems. As illustrated by the graph, SALAAM achieves the highest FM compared against other unsupervised systems; moreover, SALAAM rivals both the partially supervised and supervised systems with only three supervised and two partially supervised systems achieving significantly higher FM scores. Therefore, the evidence supports accepting Hypothesis 1.

3.8.2 Testing Hypothesis 2

This hypothesis tests the comparability of SALAAM’s performance when the two MT systems for the target language are interleaved before alignment. Table 3.3 shows the results of condition FR-GLSYS compared against the results obtained by SALAAM in conditions FR-GL and FR-SYS.

As illustrated in Table 3.3, the precision obtained in condition FR-GLSYS is higher than that obtained by either condition single MT system condition. We note an increase

descriptions of the respective systems published in the SENSEVAL workshop proceedings and further confirmed through personal communication with the authors.

¹⁵This is considered partially supervised due to the fact that the frequency information in WN17pre is based on SemCor which comprises 200k words of hand annotated sense running text of the Brown Corpus. Moreover, there is no quantification on the number of cases where the respective system backs-off to the most frequent sense.

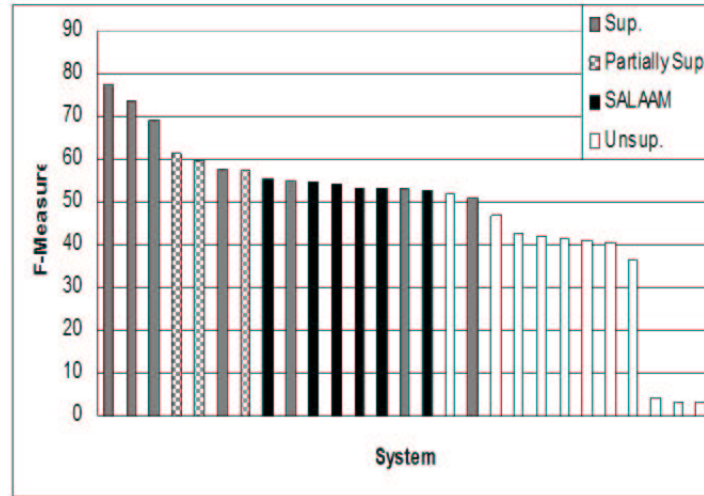


Figure 3.13: SALAAM F-Measure results in the default condition measured against state-of-the-art WST systems on test set SV2AW

Condition	P%	R%	COV%	FM
FR-GL	58.1	50.9	87.62	54.26
FR-SYS	58	49	84.43	53.12
<i>FR-GLSYS</i>	<i>60.6</i>	<i>49.1</i>	<i>81.05</i>	<i>54.25</i>

Table 3.3: SALAAM performance with pre-alignment French target pseudo-translation merge

Condition	P%	R%	COV%	FM
AR-TR	58.3	51.4	88.27	54.63
AR-AM	57.5	49.3	85.74	53.09
AMTR-AR_M	59.1	55.4	93.71	57.19
FR-GL	58.1	50.9	87.62	54.26
FR-SYS	58	49	84.43	53.12
GLSYS-FR_M	59.4	54.5	91.74	56.84
SP-GL	57.9	48.6	83.86	52.84
SP-SYS	60	51.5	85.93	55.43
GLSYS-SP_M	59.8	53.3	89.21	56.36

Table 3.4: SALAAM performance in default condition vs. intralanguage post-alignment merge condition at MAX sense selection criterion

of approximately 2.5%. Recall is at a mid point between the two recall measures for the single MT conditions FR-GL and FR-SYS. We note that the FM score is close to the high end of the range between the FM scores achieved by FR-GL and FR-SYS. The COV scores are less in condition FR-GLSYS than either default condition. The results lend positive support to Hypothesis 2.

3.8.3 Testing Hypothesis 3

We test the hypothesis that merging the source-token alignments before submission to the NG algorithm yields better recall results than recall obtained by SALAAM when the evidence is from a single MT per language. The merge of the two pseudo-translations occurs post GIZA++ alignment. The sense selection criterion is set to MAX. The intralingual post-alignment conditions are GLSYS-FR_M, GLSYS-SP_M, and AMTR-AR_M, for French, Spanish and Arabic, respectively. Table 3.4 illustrates the results obtained.

The highlighted results in Table 3.4 are achieved by SALAAM in the three conditions, GLSYS-FR_M, GLSYS-SP_M and AMTR-AR_M. These results are higher than those obtained by SALAAM in the default conditions. In general, we observe a slight insignificant increase in precision; SALAAM in condition AMTR-AR_M yields better precision results than those obtained by either AR-TR or AR-AM; similarly for GLSYS-FR_M, where we note an increase of 1.3% and 1.4%, over FR-GL and FR-SYS.

As for recall, we observe a statistically significant improvement in the intralanguage conditions over the default conditions across the board. SALAAM achieves an improvement of 4-6% in condition AMTR-AR_M over the default conditions AR-TR and AR-AM. SALAAM in GLSYS-FR_M achieves an improvement of 3.4-4.5% over the individual French default conditions. Notably, we see a significant improvement in

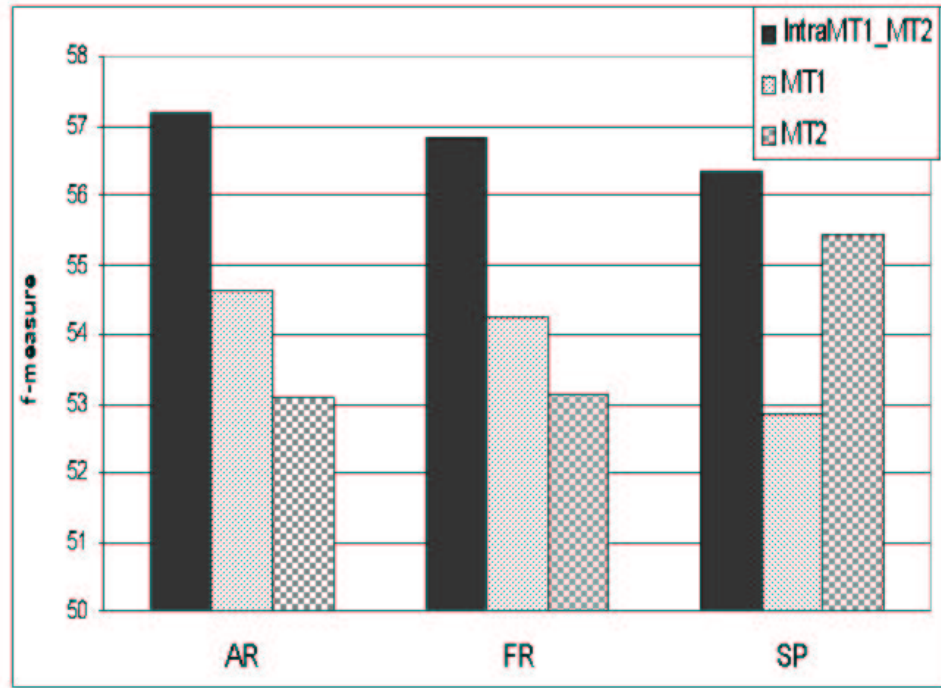


Figure 3.14: SALAAM F-Measure results in the default condition measured against MAX intralanguage condition for the three languages: AR, FR, SP

coverage and FM scores. These improvements are expected as the translation variability increases leading to the creation of more source sets with multiple members. These results support Hypothesis 3.

Figure 3.14 illustrates the significant improvement in FM scores from the single MT default conditions to the intralanguage condition with a sense selection criterion set to MAX. Each cluster of columns pertains to a language. The black column is the intralanguage Condition 3.

We observe that precision score for GLSYS-FR_M is slightly less than precision for FR-GLSYS, however, this is accompanied by a significant improvement in recall from FR-GLSYS to GLSYS-FR_M which leads to a statistically significant improvement in both FM score and COV for the latter condition.

3.8.4 Testing Hypothesis 4

We report results of evaluating SALAAM in Condition 4, an intralanguage condition where the sense selection criterion is set to THRESH, against results obtained from evaluating SALAAM in condition 3 which is also an intralanguage condition but the

Condition	P%	R%	COV%	FM
AMTR-AR_T	64.5	53	82.08	58.19
AMTR-AR_M	59.1	55.4	93.71	57.19
GLSYS-FR_T	65.6	52.1	79.46	58.08
GLSYS-FR_M	59.4	54.5	91.74	56.84
GLSYS-SP_T	65.7	50	76.28	56.78
GLSYS-SP_M	59.8	53.3	89.21	56.36

Table 3.5: SALAAM performance in intralanguage post-alignment merge conditions with sense selection criterion MAX vs. THRESH

sense selection criterion is set to MAX. Table 3.5 illustrates the results obtained.

The precision scores obtained by SALAAM in Condition 4, where the sense selection criterion is set to THRESH, are statistically significantly higher than those obtained by SALAAM with the MAX selection criterion. We observe a significant improvement in precision of 5.4% from AMTR-AR_M to AMTR-AR_T. Similar behavior is observed for the other language conditions. We notice a drop in recall which is expected since the THRESH sense selection criterion removes some valid candidates in the noise removal process. However, the drop is not statistically significant. Furthermore, we observe an expected drop in coverage but an increase in the FM score. The results support Hypothesis 4.

3.8.5 Testing Hypothesis 5

Precision results obtained from applying SALAAM to the test data using evidence obtained from several languages, where the tag sets are intersected, are better than precision results based on evidence pertaining to a single language. We evaluate this hypothesis by comparing results obtained by SALAAM in Condition 5 against results obtained in Condition 4, the intralanguage merge condition. Table 3.6 illustrates the results where the sense selection criterion is set to THRESH.

The first three rows in Table 3.6 illustrate the results obtained by SALAAM in Condition 4. The last four rows indicate the results obtained by SALAAM in Condition 5. As illustrated by the table, the precision increases across the board. For instance, considering condition AR-FR_I_T, we note that its precision, at 66.5%, is higher than that of AMTR-AR_T at 64.5% and GLSYS-FR_T at 65.6%. similarly, for all the conditions including condition AR-FR-SP_I_T, at a precision of 66.6%, it is higher than the precision yielded by AMTR-AR_T, GLSYS-FR_T, and GLSYS-SP_T. It is worth noting that the highest precision obtained is from the combination of evidence from French and Spanish which is explainable by the proximity between the two languages.

Owing to the exclusive nature of the intersection, we note the expected decrease in recall and coverage; moreover, we observe a decrease in FM scores for the multilingual

Condition	P%	R%	COV%	FM
AMTR-AR_T	64.5	53	82.08	58.19
GLSYS-FR_T	65.6	52.1	79.46	58.08
GLSYS-SP_T	65.7	50	76.28	56.78
AR-FR_I_T	66.5	50.5	75.89	57.41
AR-SP_I_T	66.9	48.4	72.42	56.17
FR-SP_I_T	67.6	49.6	73.26	57.22
AR-FR-SP_I_T	66.6	48.2	72.42	55.93

Table 3.6: SALAAM performance for conditions 4 where evidence is obtained from monolingual intralanguage pseudo-translation merge vs. evidence obtained from interlanguage pseudo-translation intersection merge in condition 5

Condition	P%	R%	COV%	FM
AMTR-AR_T	64.5	53	82.08	58.19
GLSYS-FR_T	65.6	52.1	79.46	58.08
GLSYS-SP_T	65.7	50	76.28	56.78
AR-FR_U_T	61.6	56.4	91.46	58.89
AR-SP_U_T	61.8	55.3	89.59	58.37
FR-SP_U_T	62.3	53.2	85.37	57.39
AR-FR-SP_U_T	60.2	55.6	92.31	57.81

Table 3.7: SALAAM performance for conditions 4 where evidence is obtained from monolingual intralanguage pseudo-translation merge vs. evidence obtained from interlanguage pseudo-translation union merge in Condition 6

conditions. The results in this section support Hypothesis 6.

3.8.6 Testing Hypothesis 6

Recall results obtained from applying SALAAM to the test data using evidence obtained from several languages, where the tag sets are union merged, are better than recall results based on evidence pertaining to a single language. We evaluate this hypothesis by comparing results obtained by SALAAM in Condition 6 against results obtained in Condition 4, the intralanguage merge condition. Table 3.7 illustrates the results where the sense selection criterion is set to THRESH.

The first three rows in Table 3.7 illustrate the results obtained by SALAAM in Condition 4. The last four rows indicate the results obtained by SALAAM in Condition 6. As illustrated by the table, we note an increase in recall and coverage from Condition 4 cases to Condition 6 cases. For illustration, condition AR-FR_U_T yields a recall score of 56.4% in comparison to the scores of 53% by AMTR-AR_T and 52.1% for condi-

tion GLSYS-FR_T. The increase is expected since the union is an inclusive merge. It is worth noting that the highest recall and coverage are yielded with Arabic both in the monolingual condition AMTR-AR_T with a score of 53% recall, 82.08% coverage and 58.19 score FM; likewise for the interlanguage merge conditions, the highest scores are yielded by the conditions that involve Arabic, AR-FR_U_T, AR-FR-SP_U_T and AR-SP_U_T.

We also note the drop in precision across the board for all the interlanguage union conditions compared to the monolingual conditions.

3.8.7 Overall results

We summarize the best results obtained from the different conditions in this SALAAM evaluation.

The highest precision obtained is from the interlanguage intersection condition at 67.6 % for FR-SP_I_T. The highest recall obtained is 56.4% in the interlanguage union merge condition AR-FR_U_T. The highest coverage score obtained is 92.31% yielded by SALAAM in condition AR-FR-SP_U_T. The highest FM score result is 58.89% obtained in condition AR-FR_U_T.

The highest performance results from any monolingual condition is yielded by AMTR-AR_T where it achieves an FM of 58.19, with a precision of 64.5% and recall of 53% and coverage of 82.08%.

Similar to Figure 3.12, Figure 3.15 plots the FM performance of SALAAM in the conditions that yield the highest scores against state-of-the-art WSD systems.

The black columns in Figure 3.15 show the best FM scores obtained by SALAAM in conditions AR-FR_U_T, AR-SP_U_T and AMTR-AR_T. As we can see illustrated in the graph, SALAAM outperforms all of the unsupervised methods and is on par with the partially supervised and some of the supervised methods.

3.9 Discussion

3.9.1 Summary of the Results

We have established that SALAAM using translational data as a source of evidence is a very successful approach to WSD. It is worth stressing the novelty of the approach where the source of evidence is orthogonal to the traditional sources of evidence used in the field. SALAAM is radically different from any of the other systems in this evaluation.

Results obtained by the default conditions are highly comparable (in most cases better) than those obtained by state-of-the-art unsupervised methods for WSD. Moreover, merging MT systems for obtaining pseudo-translations for the same language yields even better results than the utilization of a single MT system. Precision results

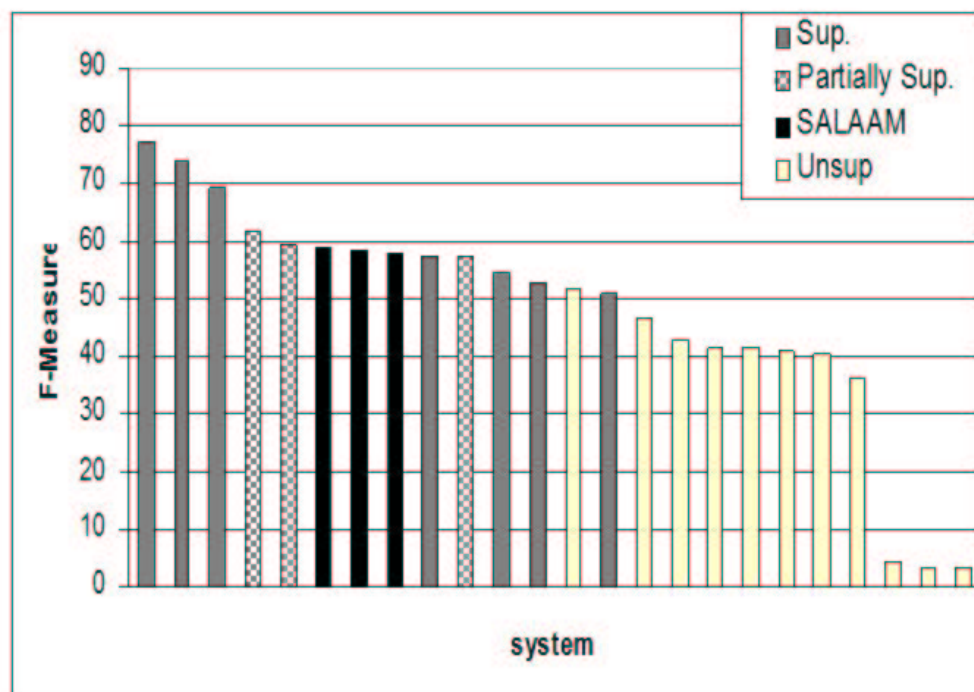


Figure 3.15: SALAAM F-Measure results on test set in SV2AW in the highest yielding conditions depicted against state-of-the-art WSD systems

obtained from all the conditions that are set to the THRESH sense selection criterion are better than those obtained when the sense selection criterion is set to MAX. Intersecting the output of two target languages for SALAAM yields better precision scores than SALAAM applied using a single target language. Union merging the output of two target languages for SALAAM yields better recall values as well as significantly higher coverage of the data. The final paragraph in the results Section 3.8 illustrates that the best results obtained by SALAAM rival a number of supervised methods and are far superior to other unsupervised methods.

The results validate the overall general hypothesis that translations aid in resolving sense ambiguity. These results obtained are very interesting especially given that the target translations are MT based which are orders of magnitude worse than human translations. However, MT has the advantage of rendering relatively good quality alignments owing mainly to the consistency of translation and fidelity to source language word ordering.

3.9.2 Analysis of Results

The results in general are consistent with our hypotheses. There are several issues worth drawing our attention to before qualitatively assessing the merits of the results.

The first issue worth consideration is that of language. As noted earlier, the highest precision results are yielded by intersection merge of French and Spanish, while the highest coverage and recall are yielded by the union merge of either French or Spanish with Arabic. The first result is explainable by the shared ambiguity across French and Spanish indicating that they target the same portion of English data. Therefore, the common noun instances that are tagged by each language independently reinforce each other. On the other hand, Arabic most likely overlaps less with either French or Spanish leading to less reinforcement of evidence, but when we look at recall, we see a boost in that direction exactly for that reason. The unique noun instances in the tag sets pertaining to the merge of Arabic with Spanish or French are more numerous than those found in the merge of the intersection of Spanish and French. This observation may also help explain the lower precision yielded in condition AR-FR-SP_I.T due to the addition of Arabic which is more distant from Spanish and French. Therefore, utilizing both a distant and near language in the sense tagging of data with SALAAM is very beneficial.

The second issue is that of calibration between precision and recall. We experience the very sensitivity of these two measures and how difficult it is to achieve significant improvement on both measures, simultaneously. This observation is supported by the results obtained in conditions 5 and 6 where the union/intersection of tag sets pertaining to two different language is used as a source of evidence. However, SALAAM is able to achieve an improvement on both measures when we compare the default conditions with the intralanguage merge conditions set to THRESH sense selection criterion. Condition AR-AM SALAAM achieves a precision of 57.5%, recall

of 49.3% while AMTR-AR.T yields 64.5% precision and 53% recall yielding a statistically significant gain of 7% in precision and 3.7% in recall. Despite this encouraging result there are endemic problems that are impediments to achieving the best possible precision and recall scores. In the following sections, we analyze some of these issues.

3.9.3 Precision

Inspecting the source type sets, we notice that they comprise many outliers. The outliers exist mainly due to noisy alignment, noisy translations, or both. The problem is aggravated when the outliers are monosemous. A monosemous word will score a confidence level of 1.0 by default according to the NG distance measure, thus biasing the sense tag assignment for the other source set words. If the monosemous word is the wrong word in that set then the bias is detrimental to the sense choice for the other words in the set.

For example, the source word types {ADOLESCENCE, IDOL, TEEN, TEENAGER} form a source type set for the French target word ADOLESCENCE. Obviously IDOL is an outlier: Even though it is related to the other words in the source set,¹⁶ it will have a negative impact on the sense assignment of the other members in the set. Another source of outliers is distant source words that may align with the same target word. For example, AMORCE in French aligns with {INITIATION, BAIT, CAP}, which are all correct translations of the French word but they are distant from one another as AMORCE is a polysemous word in French. Yet, the source type set does not have a homogenous set of words, thus leading to a situation of noise in the tagging process that results in wrong sense assignments.

These problems are mainly a reflection of SALAAM’s implicit simplifying assumption that words in the target language are monosemous by default. Examining the source sets, we observe that this assumption is clearly false. For example, in source sets such as CANON: {CANNON, CANNONBALL, CANON, THEOLOGIAN} and BANDES: {BAND, GANG, MOB, STRIP, STREAK, TAPE} and BAIE: {BAY, BERRY, COVE}, we can clearly find narrower sub sets. Accordingly, the source set corresponding to CANON is split into {CANNON, CANNONBALL} and {CANON, THEOLOGIAN}. Likewise, for BANDES the source set can be split into two subsets: {BAND, GANG, MOB} and {BAND, STRIP, STREAK, TAPE}. These subsets reflect the homonymy of the French words.

The presence of such sub-clusters in the source set, resulting from homonymy, has a definite negative impact on the quality of the sense tagging. One way to resolve this problem is to gather distributional features for the source data and apply automatic clustering techniques in order to distinguish coarse level word distinctions in the source sets [17, 75]. Once sub-clustering is applied, coherent source type sets are discovered

¹⁶Related and similar are different notions: car and tire are related but car and automobile are similar [67]

and, simultaneously, the process discovers — in an automated unsupervised manner — the number of homonymous senses for polysemous words in the target corpus.

In order to verify this hypothesis, we randomly pick target words that have good and coherent source sets and evaluate them using the `scorer2` software.¹⁷ We note a significant improvement in precision scores. Below is an example of target words and their corresponding coherent source type sets.

ABSURDITÉ: {*ABSURDITY, FARCICAL, NONSENSE*}
ACCIDENT: {*ACCIDENT, CRASH, WRECK*}
ACCUSATION: {*ACCUSATION, FRAMING, INDICTMENT*}
ADVERSAIRES: {*ANTAGONISTS, OPPONENTS, CONTESTANTS*}
ACCOMPLISSEMENTS: {*ACCOMPLISHMENT, ACHIEVEMENT, ATTAINMENT, COMPLETION*}

3.9.4 Recall

There are several issues that affect recall negatively. Owing to memory limitations, GIZA++ sets a cap on the maximum possible length allowed for a sentence. Accordingly, 0.5% of the sentences in the test corpus are excluded. This may be fixed in the future by breaking longer sentences into sub-sentences, or simply increasing the memory of the machines in use.

The second factor that affects recall is cross-linguistic lexicalization divergences. The approach as described is limited by the unit cross-linguistic lexicalization alignments. In some cases, a source word is not lexicalized as a unit, therefore creating a one-to-many alignment relation in the target corpus, which will not be handled by SALAAM. For instance, the English word *implementation* is translated into French as **mise en oeuvre**, but since SALAAM does not handle compounds at this stage, a word such as *implementation* is left untagged.

Roughly, 33% of the target nouns are translated into the same source word throughout the corpus. There are several possible reasons for this.

The target language word preserves ambiguity the way the source word does. This could be due to the fact that the target language preserves ambiguity. For instance, the French word **intérêt** which is a translation of *interest* preserves ambiguity, thereby it is ambiguous in the same way the English counterpart is ambiguous.

In other cases, the MT system simply does not have alternatives for the source language word, thereby rendering the same target word for the same source word throughout the corpus. For instance, the source word *priest* is translated as **prêtre** in French, and *vicar* is translated as **curé**. Both are correct translations, however, the end result

¹⁷By visual inspection

is two singleton source type sets, one for **prêtre** and the other for **curé**. The singleton sets can't be tagged by SALAAM. Therefore, translating both *vicars* and *priests* as **curés** would solve this problem. Singleton source sets lead to the exclusion of source words from the tagging process. One solution is to introduce more variability in the corpus genre, leading to more variability in the translation especially if the genre of the test corpus is diversified.

3.9.5 Coverage

The maximum coverage yielded by any of the SALAAM conditions is 92.01%. This coverage figure is expected due to the fact that SALAAM at this stage only processes unit sized entities. In an earlier study by Ide [32], she concludes that only 86.6% of the single lexical units in the novel *Nineteen Eighty Four* correspond to single lexical units in translation when looking for correspondents of words in five different languages pertaining to four different language families. This discovery places an upper bound on the coverage achievable by a system such as SALAAM.

3.9.6 Complementarity with Other WSD Systems

We have repeatedly stressed the radical difference between SALAAM and other state-of-the-art WSD systems. SALAAM relies on an orthogonal source of evidence for its bias toward a specific sense assignment. This leads us to believe that there is qualitative and quantitative evidence for SALAAM's complementarity to other systems. Upon analyzing the data, we find that, indeed, there are some crucial and interesting differences.

For instance, SALAAM when applied in the default conditions GLSYS-FR and/or GLSYS-SP correctly tags over 14% of the cases of polysemous words in the test set that could not be correctly tagged by any of the systems whether supervised or unsupervised. SALAAM is correctly able to sense tag over 49% of the data that none of the unsupervised systems tag correctly. This indicates that SALAAM can definitely complement a traditional monolingual WSD approach to achieve even better results.

3.9.7 Evaluation of Target Language Tagging

As a product of this evaluation, we obtain sense-tagged target data in addition to the sense-tagged source data. At this stage, we do not evaluate the projected tagging quantitatively (see chapter 4, for a thorough evaluation of the target sense tagging). But the following example attempts to give a feel for the sense-annotation quality. We illustrate with a sentence that is randomly chosen from the test corpus SV2AW, with the actual annotations produced by SALAAM. The first line shows the English sentence with the sense-tagged nouns, and the second sentence shows the corresponding French sentence as translated by SYS.

English sentence d01.s57 from SV2AW

*Dr.*_(8044078|8043248) *Vogelstein next turned his attention*₍₄₆₈₆₅₂₎
to colon₍₁₁₇₇₅₈₅₅₎ *cancer*₍₁₁₇₇₅₈₅₅₎, *the second biggest cancer*₍₁₁₇₆₉₀₈₂₎
killer₍₈₂₀₅₄₁₁₎ *in the U.S.*₍₇₂₆₂₀₉₈₎ *after lung*₍₁₁₇₇₉₈₄₈₎ *cancer*₍₁₁₇₇₉₈₄₈₎.

FR-SYS sentence d01.s57 from SV2AW

*Dr.*_(8044078|8043248) *Vogelstein après a tourné son*
attention₍₄₆₈₆₅₂₎ *au cancer*₍₁₁₇₇₅₈₅₅₎ *de deux*_(11775855*)
*points*_(11775855*), *le deuxième plus grand tueur*₍₈₂₀₅₄₁₁₎ *de*
cancer₍₁₁₇₆₉₀₈₂₎ *aux Etats*₍₇₂₆₂₀₉₈₎ *- Unis*₍₇₂₆₂₀₉₈₎
après cancer₍₁₁₇₇₉₈₄₈₎ *de poumon*_(11779848*).

Sense tags marked with an asterisk in the annotated French translation are incorrect. The tagging for **deux** and **points** is incorrect because they are mistranslations. **deux points** is the translation of the *punctuation* sense of *colon* in English. It should have been translated as **colon** in French.

A different problem occurs with **poumon** (lung), which is tagged with the *disease* sense (as a part of *lung cancer*) instead of as an *organ*.

The case of **Etats-Unis** illustrates an instance of mis-tokenization, where a single lexical item is broken to three, yet the tag is correct because each individual token in (**etats**, **_**, **unis**) is aligned with *U.S.*.

3.10 Summary

In this chapter, we present a novel unsupervised method and system, SALAAM, for WST. The method achieves very competitive levels of precision and recall when evaluated against other unsupervised systems on the same test set. SALAAM is novel in its extension of the notion of context to a multilingual dimension. It is complementary to other state-of-the-art systems as it targets contexts that are typically not that amenable to traditional approaches exploiting monolingual contexts. In Chapter 4, we show that this method may be used for bootstrapping sense inventories for a language with scarce resources.

Chapter 4

Extensions to SALAAM

4.1 Introduction

This chapter explores several extensions to the SALAAM system (see Chapter 3). It comprises 3 sections. In the first section, Section 4.2, motivated by the lack of machine translation systems for most languages, we investigate the impact of applying SALAAM to naturally-occurring parallel corpora of genres unrelated to the genre of the test corpus. In Section 4.3, we explore bootstrapping the tagging process for a target language and evaluate the quality of the projected word sense tagging of both Arabic and Spanish. Finally, in Section 4.4, we discuss the feasibility of bootstrapping a WordNet style ontology for Arabic based on SALAAM data.

4.2 Using Human Translations — Naturally-Occurring Parallel Corpora

4.2.1 Introduction

In Chapter 3, we empirically demonstrate that applying SALAAM to parallel corpora is a promising novel approach to WSD; the results obtained are significantly higher than other state-of-the-art unsupervised WSD systems while rivaling some supervised systems when evaluated on the same data set, SV2AW. In this section, we investigate the application of SALAAM to naturally-occurring parallel corpora as opposed to machine translated, pseudo-translated, parallel corpora. The question is how robust is SALAAM as an approach given a naturally-occurring domain specific parallel corpus of a genre that is unrelated to the test corpus genre. Accordingly, we examine the impact of various corpora genre on SALAAM's performance.

4.2.2 Motivation

In order for SALAAM to work, the need arises for variability in translation contexts which produce rich source type sets (see Chapter 3, Section 3.6). The belief is that such a variability in translation contexts should be obtained from naturally-occurring balanced parallel corpora, which unfortunately do not exist. To date, mining such corpora from the web is still a promise that has not materialized, mainly owing to copyright issues. Moreover, there are no naturally-occurring parallel corpora that are tagged. Hence, in Chapter 3, we rely on pseudo-translations as an approximation. But, the fact is, the majority of languages do not have machine translation systems. Nonetheless, a lot of languages do have domain specific texts in translation; for instance, the Bible exists in over 2000 languages. Therefore, in this section, we investigate applying SALAAM to genre specific corpora that are naturally-occurring in lieu of pseudo-translations as in Chapter 3, for the augmentation of the same test corpus SV2AW. In effect, we are essentially measuring the robustness of the SALAAM approach when using corpora that are incongruent with the test corpus and do not possess the expected level of variability in translation contexts.

4.2.3 Hypothesis

We have two general hypotheses:

1. **Augmenting SV2AW with naturally-occurring genre specific parallel corpora while applying SALAAM yields comparable precision results to those obtained by augmenting SV2AW using pseudo-translation corpora .**

Naturally-occurring parallel corpora have more translation variability than pseudo translated corpora as the translation process is subject to the creativity of the human translator. When naturally-occurring parallel corpora are not balanced, we expect a reduction in translation variability which allows for the formation of source type sets comparably variable to those produced via pseudo-translations.

2. **Corpus genre has a significant impact on SALAAM recall results.**

Polysemous words if used in a genre specific corpus will have a bias toward specific senses. The absence of domain specific knowledge of the senses is a problem which is escalated if this genre is not of the same type as that of the test corpus. Furthermore, if the genre is narrow, for example, religious or political to the exclusion of other genres, it tends to be consistent in the translation of its terms, therefore decreasing the variability in translation contexts.¹ This leads to

¹If the corpus genre is narrow yet of the same type as that of the test corpus, and the translator(s) uses variable ways of expressing ideas, then this is a favorable condition for SALAAM. In principle, SALAAM is expected to perform well if the augmenting corpus used is balanced, regardless of the genre of the test corpus — whether balanced or not — as this condition creates variable source type sets; or

the creation of singleton source type sets (see Chapter 3, Section 3.6) that are not amenable to SALAAM to tag, hence, significantly affecting recall.

4.2.4 Evaluation

The evaluation metrics and significance testing used here are the same as those used for SALAAM in Chapter 3, Precision (P), Recall (R), F-Measure (FM) and coverage (COV). The tokenization tools, stochastic alignment software, ontology, gold standard and test set are also the same as those used in Chapter 3. We report the evaluation of applying SALAAM using two sets of corpora for augmenting the same test corpus, SV2AW, as described in Chapter 3, Section 3.7.1: naturally-occurring parallel corpora which are human translation corpora (HT); and both HT corpora and pseudo-translated (MT) corpora which are the corpora used in Chapter 3. In the process, we explore the impact of pruning the alignments in the HT conditions using a bilingual dictionary. All the evaluations are on English-Spanish parallel corpora.

Corpora

In addition to the MT corpora described in Chapter 3, we have two HT corpora. Moreover, we describe the test corpus, SV2AW, here again for convenience. Table 4.1 indicates the relative sizes of the different corpora used. Here follows a description of the three parallel corpora:

- **The Bible (BIB)**

This corpus comprises the Old and New Testaments. The English version is the NIV Bible written in modern English and last updated 1901.² The Spanish Bible is written in modern Spanish.³ BIB has approximately 820K tokens per side. BIB is religious text that is aligned at the verse level [73].

- **Proceedings of the United Nations 1989-1990 (UN)**

The UN text is written in modern day English and Spanish. The portions used in this evaluation specifically date back to the years 1989 and 1990. It is a political and economic genre corpus. The corpus is semi-automatically sentence aligned.⁴ The resulting corpus has approximately 1.7 million words per language side.

alternatively if the augmenting corpus is of the same genre of the test corpus with variable translation contexts.

²<http://www.sni.net/mpj/WEB/index.htm>

³<http://www.mit.edu/afs/athena.mit.edu/activity/c/csa/www/documents/Spanish>

⁴Thanks to Clara Cabezas, a bilingual native Spanish speaker.

- **The SENSEVAL 2 All Words corpus (SV2AW)**

This is the test corpus. SV2AW has three articles from The Wall Street Journal. The articles discuss culture, medicine and education, respectively. Each side has close to 6000 tokens. In this set of experiments, SV2AW is pseudo-translated into Spanish using both GL and SYS machine translation systems in the intralanguage post alignment merge condition. The idea is to maximize the translation variability of the test corpus as this is established a significant improvement relative to a single translation system result (see section 3.7.8, Condition 3), where the English-Spanish parallel tokens are aligned and merged before submitting to the **NG** algorithm for sense assignment.

Corpora	Lines	Tokens
BIB	30427	829031
UN	71672	1734001
SV2AW	242	5815
<i>Total</i>	<i>102341</i>	<i>2568847</i>

Table 4.1: Relative sizes of the English side of corpora used in **HT** Evaluations

Parameters

Similar to the SALAAM evaluation using **MT**, we set the alignment software parameters at 70 tokens per sentence. The sense selection criterion is set to **MAX** (see section 3.7.3) for all the evaluations in this section.

Conditions

We explore the following experimental conditions:

1. **SV2AW alone (SV2AW)**

This condition aims at viewing SALAAM’s raw results on the test set alone with no augmented corpora. We consider this condition the baseline condition; it sets an upper bound on precision⁵ and a lower bound on recall, at **MAX** sense selection criterion.

2. **BIB with SV2AW (BIB+SV2AW)**

This condition examines the results of augmenting the SV2AW test corpus with the Bible corpus, BIB.

⁵Ideally, we can find the true ceiling value for precision if we have human translations and perfect alignment.

3. **UN with SV2AW (UN+SV2AW)**

This condition explores the results of augmenting the SV2AW test corpus with the United Nations corpus, UN.

4. **Fixed UN and SV2AW (Fixed_UN+SV2AW)**

Upon inspecting the token alignment quality of the HT corpora, we realize severe problems due to the inconsistency in sentence length from English to Spanish but also owing to naturally-occurring divergences in syntactic and semantic expression cross-linguistically. Therefore, in this condition, the UN alignments are fixed with some linguistically motivated rules. The rules are heuristics for basic category swapping; it is observed, for instance, that the alignment software consistently swaps adjectives and nouns in Spanish. The correction rules rely on the POS tagging on the English side. Rules are applied in a specific order.

- If an English word is mapped to two Spanish words and the English word following it is mapped to the NULL token then the first Spanish word is assigned to the following English word, rendering the alignment one-to-one in this case.
- If there are three Spanish words aligned with three English nouns in a row, each English noun and its Spanish alignment are checked to see if they share some prefix; if not, then the first and the third Spanish words are switched.
- Spanish translations of English adjectives followed by nouns in English — on the Spanish side indicating that the English words are left untranslated by the MT system — are swapped.

5. **Pruned UN and SV2AW (Pruned_UN+SV2AW)**

An alternative method for fixing the alignments is to use a bilingual dictionary to prune the translations. We use a generic bilingual English-Spanish dictionary which comprises 90K entries. The alignment pairs are filtered so that those that do not occur in the dictionary are removed.

6. **Pruned and Fixed UN and SV2AW (Pruned_Fixed_UN+SV2AW)**

In this case, the UN corpus alignments are fixed according to the correction rules in condition 4 and then pruned according to Condition 5.

7. **UN, MT, and SV2AW (UN+MT+SV2AW)**

For this condition, we merge the UN alignments with those of the pseudo-translated (MT) corpora used in Chapter 3, comprising the Brown Corpus, SENSEVAL1 corpus, SENSEVAL2 Lexical Sample and The Wall Street Journal corpora, in addition to the test corpus SV2AW. Similar to the test corpus for this

evaluation, the pseudo-translated corpora are in the post alignment intralanguage merge condition (see Section 3.7.8, condition 3).

8. **Fixed UN, MT and SV2AW (Fixed_UN+MT+SV2AW)**

Similar to Condition 7, but the UN alignment portion is fixed with the correction rules described in Condition 4.

9. **UN, BIB, MT, and SV2AW (UN+BIB+MT+SV2AW)**

Similar to Condition 7, but the BIB corpus is added to the augmented parallel corpora.

10. **Fixed UN, Fixed Bible, MT and SV2AW (Fixed_UN+Fixed_BIB+MT+SV2AW)**

Similar to Condition 9 but both UN and BIB corpora alignments are fixed based on the rules described in Condition 4. .

Results

Table 4.2 demonstrates the results obtained by SALAAM where the test corpus, SV2AW, is augmented by HT parallel corpora. We include the results for condition GLSYS-SP (see Chapter 3, Section 3.7.8, Condition 3) for comparison of the HT results against an MT result. GLSYS-SP is chosen because it is the closest approximation to a human translation using machine translation systems; moreover, it yields the best results for the Spanish data with the MAX selection criterion. The SV2AW condition illustrates the upper bound on precision and the lower bound on recall.⁶ Both SV2AW and GLSYS-SP are pseudo-translated conditions, MT, hence the bold typeface in the table.

Conditions	P	R	COV	FM
GLSYS-SP	59.8	53.3	89.21	56.36
SV2AW	69.6	24.3	34.9	36.02
UN+SV2AW	57.5	44.5	77.31	50.17
BIB+SV2AW	56.5	36.6	64.82	44.42
Fixed_UN+SV2AW	58.6	44.8	76.45	50.78
Pruned_UN+SV2AW	59.1	32.9	55.72	42.27
Pruned_Fixed_UN+SV2AW	59	33	55.91	42.33

Table 4.2: SALAAM Results on SV2AW for MT & HT parallel corpora independently

⁶Due to the very small size of this test corpus, the SV2AW token alignments are obtained from aligning the entire MT corpus as described in Chapter 3

As expected, the results obtained in the SV2AW condition alone yield the highest precision and the lowest recall across all the different conditions — actually by comparison to all SALAAM conditions, even those of Chapter 3. The overall results indicate that the use of other corpora, whether pseudo-translated or naturally-occurring parallel corpora, plays a significant role in improving the recall values while adding significant noise and thereby reducing precision.

Precision for all HT conditions does not differ significantly from precision of the MT condition GLSYS-SP; according to the `Zscore` statistical significance test the conditions are the same with ($p < 0.0025$) confidence. All HT experimental conditions yield lower FM results when compared with the MT experimental condition GLSYS-SP, yet markedly at statistically significant lower coverage scores. Recall for all HT conditions is significantly lower than recall for GLSYS-SP but at the same time significantly higher than that of condition SV2AW; all HT conditions at least double the coverage achieved by condition SV2AW. Results obtained by conditions that use the UN corpus are better than those obtained using the BIB corpus. Precision and recall obtained by condition Fixed_UN+SV2AW are slightly higher than the UN+SV2AW condition; the improvement is minor, but it shows that fixing alignments is a step in the right direction. We see further improvement in precision for Pruned_UN+SV2AW over Fixed_UN+SV2AW, yet recall is significantly reduced.

Conditions	P	R	COV	FM
GLSYS-SP	59.8	53.3	89.21	56.36
UN+MT+SV2AW	60	54.7	91.18	57.23
Fixed_UN+MT+SV2AW	60.8	55.4	91.18	57.97
UN+BIB+MT+SV2AW	60.1	55	91.46	57.44
Fixed_UN+Fixed_BIB+MT+SV2AW	60.5	55.4	91.46	57.84

Table 4.3: SALAAM results using both **HT** and **MT** for augmenting the test corpus

Table 4.3 illustrates the results of merging the pseudo-translated MT corpora with the HT corpora for augmenting the test corpus SV2AW. We observe a slight improvement in all the results relative to condition GLSYS-SP, yet none of the results yielded by the different experimental conditions is statistically significantly better than condition GLSYS-SP. We note the minor improvement associated with fixing the alignments; we see an increase of 0.8% in precision from UN+MT+SV2AW to Fixed_UN+MT+SV2AW and an increase of 0.7% in recall maintaining the same coverage level; similarly, we note an increase of 0.4% from condition UN+BIB+MT+SV2AW to condition

Fixed_UN+Fixed_BIB+MT+SV2AW. Adding the BIB corpus to the mix seems to slightly improve recall and coverage; For instance, comparing UN+MT+SV2AW and UN+BIB+MT+SV2AW we observe an increase of 0.3% in recall and 0.28% in coverage.

4.2.5 Discussion

As hypothesized, we obtain comparable precision scores when augmenting SALAAM with genre specific naturally-occurring parallel corpora and pseudo-translated corpora. This illustrates the robustness of the SALAAM approach. As expected there is a clear correlation between corpus genre and performance. Even though the difference between precision scores yielded by the conditions UN+SV2AW and BIB+SV2AW is not statistically significant, we observe a drop of 1% from augmenting the test corpus with the UN corpus to augmenting it with BIB. On the other hand, the drop in recall is statistically significant, with a drop of 8% from UN+SV2AW to BIB+SV2AW. The decrease is due to the relative distance of the corpora genre. Qualitatively, the language of the BIB corpus is stylized, which is very different from the language style used in the UN corpus or the test corpus. In fact, just by looking at the dates of the corpora, the UN corpus and the test corpus pertain to the late 20th century, while BIB is early 20th century. This is further supported by the unigram overlap between the BIB corpus and the test corpus of 944 tokens compared to the UN corpus unigram overlap of 1249 tokens, thereby exhibiting an increase of 25% in overlap between the UN corpus and the test corpus.

When the HT corpora are merged with the pseudo-translated corpora, we observe modest improvements in the different measures.

As noted earlier, there is a very sensitive balance between precision and recall, which emerges clearly in all these experimental conditions. It is a challenge to improve on both measures simultaneously. We observe a promising improvement on all metrics in Table 4.3, but less than expected. We believe there are two reasons for this.

The first endemic problem comes from the nature of the HT corpora utilized. There is no genre overlap between the HT corpora and the test corpus. The test corpus has articles about education, medicine and culture; the UN corpus is mostly economic and political in nature; the BIB corpus is religious text. Not surprisingly, these corpora added too much noise to the source type sets. In contrast, the pseudo-translated corpora used in Chapter 3 included text from relevant genres.

Secondly, the automatic token alignments of the HT parallel corpora are much worse by qualitative inspection than those obtained from pseudo-translations. This is an expected drawback. Human translation tends to be more creative: Often translators express sentences in different lengths in different languages; such variations in length cause havoc for the token alignment software. Upon inspecting the HT English Spanish alignments, we find on average 30% of the tokens aligning with the NULL token, compared with 10% of the tokens in the pseudo-translations. An indication of the promising impact an improvement in the alignment would yield is illustrated by the modest improvement in the results from raw alignments to fixed alignments — albeit with ad-hoc rules and heuristics — as presented in Tables 4.2 and 4.3. It is worth noting that fixing automatic alignments is a vast research area which falls outside the scope of this thesis [52, 31]. As expected, pruning has a negative effect on recall; it

eliminates many possible valid members from the source type sets which is probably due to lack of coverage or genre variation between the test corpus and the dictionary utilized. Nonetheless, it has a positive effect on precision.

Looking at the flip side of these results, we believe there are two factors that aid the pseudo-translated version of these experiments. The first factor lies in the genre of the corpora utilized; they cover a myriad of different genres which overlap with the test corpus genre. The second factor is the fact that the pseudo-translations are very consistent translations that render better alignments relative to the HT parallel corpora token alignments.

4.2.6 Summary

In this section, we establish SALAAM’s robustness given naturally-occurring parallel corpora of genre types that are completely unrelated to the test set, SV2AW. The results obtained show no significant difference in performance precision for SALAAM using pseudo-translations of relevant corpora genre versus utilizing unrelated genre corpora. We also note the degraded quality of alignments when using naturally-occurring parallel corpora relative the pseudo-translated parallel corpora.

4.3 Target Language Tagging Evaluation

4.3.1 Introduction

In this section, we discuss the quality of the projected sense tags onto the target language words in SALAAM. We present two quantitative evaluations of the projected tagging on two target languages: Spanish and Arabic. The tagged Spanish target text is automatically evaluated against manually annotated Spanish test data. The tagged Arabic data is manually evaluated. This section is arranged as follows: section 4.3.2 presents the motivation behind evaluating target tagging; in section 4.3.3, we present the underlying hypothesis driving the projected sense tagging evaluation; Section 4.3.4 briefly describes the required resources; section 4.3.5 explores sense tagged target Arabic data; Section 4.3.6 illustrates quantitative evaluation of sense tagged target Spanish data.

4.3.2 Motivation

Given a lexicon and a trained lexicographer, sense tagging texts manually is the guaranteed method of obtaining good quality sense-annotations for words in running text. However, the task is very tedious, expensive, and, by many standards, daunting to the people involved, even when all the required resources are available [25]. The problem becomes ever more challenging when dealing with a language with virtually no

computerized knowledge resources or tools. To date, the only way to obtain sense-annotations in a language with scarce knowledge resources is to do the job manually which constitutes a serious impediment given the sheer number of natural languages in the world.

SALAAM is investigated as a method for resolving this impeding bottleneck. SALAAM provides a bootstrapping method for sense tagging a language with scarce automatic linguistic knowledge resources. As a side effect of applying SALAAM to a parallel corpus and tagging the source side, we obtain a tagged target language corpus automatically, with no extra effort or cost. No target resources are required except for the actual parallel corpus and a simple tokenizer for the target language. The approach serves as an elegant solution to an age old problem and a series of bottlenecks for the acquisition of automated knowledge resources for scarce languages.

4.3.3 General Hypothesis

The application of SALAAM to a parallel corpus should provide a good source for creating seed target language sense-annotations for languages with scarce knowledge resources.

This general hypothesis is based on the premise that people share basic conceptual notions that are a consequence of shared human experience and perception regardless of the languages they speak.

The premise is supported by the fact that we have translations in the first place. People of different linguistic backgrounds are capable of communicating through other modalities. Apart from the empirical value of labelling data with their appropriate senses for computational systems, defining or quantifying senses, first and foremost, aims to make explicit these basic human notions of meaning.

Basing the target sense tagging on a source language involves nothing more than capturing that very idea of shared meaning across languages and exploiting it as a bridge to explicitly define the senses in a target language. Therefore, SALAAM is introducing a bias based on a sound cognitive axiom that languages share basic elements of meaning. When SALAAM is used for tagging a source language, it cashes in on the variation in translation of polysemous words. The flip side of this view is aims at quantifying meaning commonality across two languages.

4.3.4 Required Resources

The current evaluation does not require additional development resources or tools over those used for the evaluation of the SALAAM performance reported in Chapter 3 and Section 4.2 in this chapter. The same corpora utilized for evaluating SALAAM on a source language are used here for the evaluation of the projected sense tagging of the target language. For an evaluation of the target sense tagging, a target test set and target gold standard are identified.

4.3.5 Projected Sense Tagging on Arabic Data

Introduction

In this section, we examine the quality of the projected tags onto Arabic target data. No WordNet ontology exists for Arabic, therefore the evaluation is manual. Arabic is a low density language with scarce automatic linguistic knowledge resources. In terms of data availability, more online corpora including parallel corpora are appearing on the web, yet language specific knowledge resources such as ontologies are virtually non-existent.

Arabic is a Semitic language. It is spoken by at least 200 million people. It is one of the few languages that exhibit diagglossia. Diagglossia is a linguistic phenomenon where a community has two languages operating at the same time. All Arabic speaking countries have at least two main forms of Arabic: Modern Standard Arabic (MSA) and some colloquial form. MSA is predominantly used in written text and speeches, mostly in formal settings. Typically, MSA is understood by the educated class in the different Arabic societies. Furthermore, the language spoken in Malta is a derivative of Arabic yet it is written in Latin script. Arabic script is used by Farsi, Daari and Urdu, as well.

Most words in Arabic have their origins in 3 or 4 letter roots. Most of the roots are verbal roots. A variety of grammatical case and parts of speech are expressed by changing the root into a stem based on one of 13 templates which are variations on the verb **f3l** meaning *to do*.⁷ For example, **ktb**, which means *to write* in the infinitive form. This uninflected form may be changed into the noun **kitab**, meaning *a book* based on the template **f3al**, where the **f**, **3** and **l** correspond to the three consonants, *k*, *t*, *b*. Mainly, the transformation comes with the addition of vowel infixes. There are two types of vowels in Arabic: short vowels and long vowels. Short vowels are often ignored in written text.

Motivation

Motivated by the lack of tools for Arabic and native proficiency in the language,⁸ we examine the projected sense-annotations onto MSA Arabic target tokens.

Evaluation

- Corpora

⁷Throughout this chapter, in describing Arabic data we use **3** to indicate the letter **aiyn**, **2** for glottal stop @, upper case characters for emphatics such as **H** and **D**, corresponding to oand , respectively; and finally **P** is the **sh** sound . The English phoneme *P* does not exist in Arabic.

⁸The author possesses native proficiency in Arabic

The corpus that is evaluated is the SV2AW parallel corpus, pseudo-translated into Arabic using the Al-Misbar (AM) machine translation system.⁹ The corpus comprises 242 lines.

- **Preprocessing**

The Arabic text is transliterated into Latin script. It is tokenized and lightly stemmed;¹⁰ the prefixes and suffixes are separated out from the words; this process results in the reduction of word surface forms to stems.¹¹ Figure 4.1 illustrates the first sentence of the SV2AW English corpus with its translation into Arabic and in turn the Arabic is transliterated in the third sentence in the figure and finally tokenized as presented in the fourth sentence.

The art of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world.
**In fn dqAq tgyyr xAS bAlInjlyz, wmvI Akvr AlxwAS AlInjlyzyp, gyr wADH
 Ila bqyp AlEAlm.**
**In fn dqAq tgyyr xAS bAl Injlyz , wmvI Akvr Al xwAS Al Injlyzyp , gyr wADH
 Ila bqyp Al EAlm .**

Figure 4.1: An example of a transliterated Arabic sentence and its tokenization

- **Ontology**

WN17pre (see Chapter 3 for description) is used for the tokens' sense tagging and projection onto the Arabic text.

- **Application of SALAAM**

SALAAM is applied to the entire corpus as described in Chapter 3. The sense selection criterion is set to MAX. The SV2AW English tokens and their corresponding Arabic token alignments are extracted. The alignments have one-to-many correspondents due to the token alignment software GIZA++, where an English token may correspond to more than one Arabic token.¹² The one-to-many correspondents are compressed and in the process target compounds are

⁹<http://www.almisbar.com>

¹⁰The original tokenization script is developed in collaboration with Kareem Darwish. We do not use a morphological analyzer for Arabic since we assume the availability of minimal resources on the target side of the parallel corpus.

¹¹It is worth noting that this is not equivalent to lemmatization in English. Lemmatization reduces the word to an infinite form devoid of number information; stemming disentangles the words from the associated pronouns; in many cases, stems are left with number and case information.

¹²Note that GIZA++ allows for one-to-many alignments from the source side of the parallel corpus to the target side but not vice versa.

created. The English source nouns are automatically annotated with SALAAM assigned sense tags; the sense tags are then projected onto the corresponding target Arabic tokens.

- **Test Set**

The English noun instances and their corresponding Arabic alignments are extracted from the compressed tagged corpus, SV2AW. The English noun instances are evaluated against the gold standard described in Chapter 3. 581 English noun instances are deemed correct by `scorer2` in the fine grain mode; the correct instances — those that yield a score > 0.000 — are extracted with their corresponding Arabic alignments.¹³ Accordingly, the Arabic tokens are tagged with the projected sense-annotations. The 581 tagged Arabic token instances comprise the test set.

- **Evaluation Results**

The 581 tagged Arabic token instances are manually evaluated. Upon inspection, 526 Arabic word instances are tagged correctly with appropriate senses based on the sense label fit for the Arabic word in question and its surrounding context. Moreover, another 9 Arabic word instances are tagged with approximate senses. A sense is deemed approximate if there are senses used in the tagging that do not fit the Arabic word and its context. For example, the Arabic translation for *evening* is **msAi2yah**; SALAAM tags it with WN17pre sense IDs: 1, 2 and 3; judging by the ontology entries as listed below in Figure 4.2, senses 1 and 3 are appropriate tags for the Arabic word in this context, yet sense 2 is not a good fit, nor is it actually an appropriate sense definition for the Arabic word. Accordingly, only 2 of the 3 possible sense tags are appropriate, therefore resulting in an approximately correct sense projection onto the Arabic word.

1: *evening, eve, eventide: the latter part of the day (the period of decreasing daylight from late afternoon until nightfall); "he enjoyed the evening light across the lake"*

2: *evening: a later concluding time period; "it was the evening of the Roman Empire"*

3: *evening: the early part of night (from dinner until bedtime) spent in a special way; "an evening at the opera"*

Figure 4.2: WN17pre entries for *evening*

¹³We do not evaluate the Arabic correspondents of the incorrectly tagged English instances based on the simplifying assumption that if the English is not correct, then the Arabic probably is not correct.

Of the 581 correct English tagged instances, 38 instances are misalignments with the Arabic. For instance, the English token *cancer* aligns with the target Arabic token **okhra** meaning *other*. In 12 cases, the MT system, Al-Misbar, does not translate the English words into Arabic; the English noun is rendered as is in the translation. 5 instances have the wrong Arabic translation. In one case, the English noun is aligned with an adjective that has the correct meaning, but it is not the correct POS, therefore leading to a misfit between the senses listed in WN17pre and the Arabic word. Table 4.4 summarizes the results.

Evaluation	Number of Instances	Percent Correct
Correct	526	90.5%
Approximate	9	1.5%
Misalignments	38	6.5%
Mistagged	6	1%
Mistranslation	12	2%

Table 4.4: Accuracy results of projected tagging onto Arabic SV2AW data measured against English WN17pre sense definitions

- **Discussion**

These results are promising as a start for the process of bootstrapping sense tagging for Arabic. The Arabic tagged data is a result of applying SALAAM to SV2AW using Al-Misbar translations; this condition is not the highest yielding condition for Arabic; therefore, we may extrapolate that data resulting from applying SALAAM to the merged Arabic MT condition — which yields the highest scores for English tagging using evidence from Arabic in Chapter 3— will accordingly improve on the currently obtained result. Obtaining tagged target data in this manner is very appealing since it virtually comes for free as a side effect of applying SALAAM to a source language.

As shown in Table 4.4, 90.5% of the Arabic projected sense taggings are considered correct at the appropriate granularity level. If we extrapolate — as if we have an Arabic WordNet — from these results to the entire set of English tags, we would expect the overall Arabic performance results to be at approximately 49% precision, given only the correct tags according to our manual evaluation, without taking into consideration the potentially correct tags which could result from misalignments. Such results are extremely encouraging especially if we plot them on the overall performance graph for the English All Words SEN-SEVAL 2 task, we note that the performance on the Arabic data is right in the middle of the graph with many systems’ results for English.

include graph here from presentation

These results are very encouraging as a first pass. We acknowledge here the shortcoming of this evaluation as a post-hoc rather than blind evaluation, therefore, it is subject to inflated agreement rates. No matter how systematic and rigorous the annotator performing the evaluation, s/he tends to agree with the assigned sense more often than if s/he were to pick a sense from a set of senses rather than from a monolingual ontology. One way of solving this problem is by having several annotators perform a manual post-hoc evaluation of the tagging quality. Another method is to translate the corresponding WN17pre entries and glosses into Arabic and then ask an annotator or group of annotators to assign senses to the Arabic words without seeing the English translation, therefore rendering it a monolingual evaluation task.

4.3.6 Projected Sense Tagging on Spanish Data

Given the encouraging results obtained from manually evaluating the projected sense tagging on Arabic data, we perform a blind evaluation on Spanish target data sense-annotations. Spanish is one of the languages used as a target language by SALAAM in Chapter 3. Spanish is chosen as a target language for SALAAM because the utilized MT systems for producing the pseudo-translations claim good quality translations; moreover, computerized linguistic knowledge resources exist for Spanish. Several teams of computational linguists are currently working on building a Spanish WordNet as part of the EuroWordNet initiative [79]. Spanish WordNet is based on the same conceptual structure as English WordNets. The availability of such a resource allows for a blind evaluation of the projected sense tagging of Spanish target tokens.

This section presents an evaluation of the projected sense tagging quality for Spanish tokens, when used as a target language by SALAAM, against a Spanish WordNet gold standard.¹⁴

Gold Standard

This all nouns gold standard (AWGS) is modelled after the gold standard for SV2AW, described in Chapter 3. Our aim is to create a comparable test set/gold standard to the SV2-AW gold standard, which comprises 242 sentences. The idea is to tag all possible noun instances in running text.

A set of 250 sentences is randomly generated from the Spanish SENSEVAL 2 Lexical Sample (SP_SV2LS) corpus provided to participants in the SENSEVAL2 Lexical sample task. The sentences are automatically POS tagged by extracting tags from the output of the Spanish parser Connexor.¹⁵ The resulting sentences are sent to one of

¹⁴In this evaluation, we are tightly bound by the available resources. We have to bridge many resources owing to the fact that we do not have Spanish WordNet.

¹⁵<http://www.connexor.com>

the key sites in Spain, where Spanish WordNet is being developed. All the nouns in the sentences are manually annotated by a human annotator.¹⁶ The sense-annotation is based on the most stable version of the Spanish WordNet, which is partially linked to English WordNet 1.5. The human annotator manually fixes some of the automatically assigned POS tags. All sense tags used in this tag set exist in WordNet 1.5.¹⁷ The human annotator uses "0" to indicate an unassignable tag; mainly, for named entities. Some cases are assigned multiple sense tags, which might include the "0" tag; these cases indicate that the appropriate sense for the noun instance does not exist in the current Spanish WordNet. We will refer to these cases as *approximates*.

For the purposes of this evaluation, noun instances that are assigned a unique "0" tag are excluded. Furthermore, 13 sentences, comprising 183 sense tagged noun instances, are excluded from AWGS as they exceed the 70 token length limit requirement for the GIZA++ stochastic token alignment software. The final AWGS tag set comprises 1279 tagged noun instances corresponding to 233 sentences in SP-SV2LS.¹⁸

Test Set

The test set for this evaluation is the set all noun instances occurring in the 233 randomly generated sentences from SP-SV2LS. Therefore, it is an all words task. We refer to this test corpus as SPSV2AW. SPSV2AW comprises 1279 noun instance test items.

Corpora

Similar to the experimental setup in Chapter 3 and section 4.2 above, the test corpus of 233 sentences is augmented by other corpora in order to apply SALAAM. Similar to the problem faced in Chapter 3, to our knowledge, there are no balanced English-Spanish parallel corpora; therefore, the corpora used for augmentation are the 5 corpora used in the SALAAM evaluation in Chapter 3. The 5 corpora are: Brown Corpus (BC), SENSEVAL 1 Corpus (SV1), SENSEVAL 2 English Lexical Sample Corpus (SV2-LS), The Wall Street Journal Corpus (WSJ), and SENSEVAL 2 All Words Corpus (SV2AW). Throughout the rest of this section we will refer to this corpus as BSSSJ. BSSSJ is pseudo-translated to Spanish using both GL and SYS translation systems, thereby creating two parallel corpora, one corresponding to each MT system. The resulting Spanish pseudo-translated corpus is further augmented with

¹⁶We would like to acknowledge the annotation work by Irina Chugur, who is a native speaker of Spanish and a computational linguist working on the Spanish WordNet project under the supervision of Dr. Julio Gonzalo at UNED, Madrid, Spain.

¹⁷17 tagged noun instances are excluded as their sense IDs do not exist in sense.index file for WordNet 1.5.

¹⁸Three more sentences are excluded as they are repeated sentences resulting from the random generation process.

the Spanish SP_SV2LS corpus. SP_SV2LS comprises training and test data that is provided to participants in the SENSEVAL 2 Spanish language Lexical Sample exercise. SP_SV2LS is a multi-topic collection which comprises the created test set SPSV2AW and the whole corpus contains excerpts from newspapers, fiction, and scientific articles; like BSSSJ, SP_SV2LS does not exist in translation; as manual translation is extremely expensive, we opt for pseudo-translating it into English using both GL and SYS machine translation systems creating source pseudo-translations. Therefore, when augmenting the English side of the utilized parallel corpus with pseudo-translated source English, care is taken that the pseudo-translation on the Spanish side is from the same MT system, i.e. a parallel corpus will have English BSSSJ plus GL translated SP_SV2LS, corresponding to the Spanish side with BSSSJ pseudo-translated using GL plus the original Spanish SP_SV2LS.

BSSSJ is similar in genre to SP_SV2LS; they both cover similar domain topics. Table 4.5 lists the relative sizes of the corpora. The sizes presented are of the corpora in the language in which they originated; the numbers for BSSSJ are those of the English side of the corpus, and those for SP_SV2LS are for the Spanish side of the parallel corpus.

Corpora	Lines	Tokens
BSSSJ	226094	5555039
SP_SV2LS	6815	238339
<i>Total</i>	<i>233129</i>	<i>5793378</i>

Table 4.5: Relative sizes of corpora used in projected Spanish tagging evaluation

Ontology

AWGS is tagged with Spanish WordNet, which has direct links into WordNet 1.5. Therefore, the sense inventory used in this evaluation is WordNet 1.5. WordNet 1.5 is an older version of WN17pre, described in detail in Chapter 3; WordNet 1.5 has the same attributes and structure as WN17pre.

Evaluation Metrics

The evaluation metrics used are the same as those described in Chapter 3. We used precision (P), recall (R), and coverage (COV). The statistical significance test is the *Zscore*, described earlier in Chapter 3, measured at 95% confidence level.

Baseline

Developing an appropriate baseline for this evaluation requires great care. The main issue is the degree of overlap between WordNet 1.5, the inventory used by SALAAM for

sense annotation and projection, and Spanish WordNet, the inventory used for AW-GS. We acknowledge the overlap, yet there are granularity mismatches and cases where the senses in English simply do not have correspondents in Spanish and vice versa. In an ideal world, we would have the human annotator assigning senses from the proper intersection of the two inventories. But since we do not impose that restriction on the human annotator — the human annotator was performing the task monolingually in Spanish — and we do not have access to the actual Spanish WordNet that is used in the task, we create a baseline based on WordNet 1.5 alone. The baseline comprises all the aligned Spanish translation tokens of the English noun instances in the SP_SV2LS corpus. This results in a set of 34878 noun instances.¹⁹ Below, we discuss two possible options for assigning senses to the baseline.²⁰

- **First Listed Sense Baseline (FSBL)**

As the naming indicates, FSBL annotates the Spanish word instance that corresponds to the English noun instance with the first listed sense ID in WordNet 1.5. Similar to other WordNet Ontologies, the first sense listed is the most frequent sense according to the sense frequency in a semantic concordance (SemCor).²¹ FSBL is a questionable baseline for unsupervised methods. We consider FSBL to be a supervised baseline since it is based on sense frequencies in a manually annotated corpus. In this current evaluation, the fact that WordNet 1.5 is a bridge inventory — the actual gold standard comprises sense tags from the Spanish WordNet — may make FSBL more appealing as a baseline. Yet, we argue that first sense frequency effect carries over cross-linguistically owing to the inherent closeness between the Spanish and English languages [74].

- **Random Baseline (RBL)**

For this baseline, a sense is randomly chosen from the set of senses for a given noun instance in WordNet 1.5. The RBL baseline results are based on averaging 10 runs of the random sense generator for each instance in the baseline set. RBL is a more appropriate baseline compared to FSBL for an unsupervised method. In the absence of a Lesk based approach, it is used as the baseline for the current evaluation.

¹⁹We exclude noun instances that align with the NULL token.

²⁰The most appropriate baseline would be a Lesk style annotation of noun instances based on the glosses' word overlap in the Spanish WordNet, but unfortunately we do not have access to the Spanish WordNet.

²¹SemCor is a corpus of roughly 200k manually sense-annotated words in running text extracted from the Brown Corpus.

Experimental Conditions

For all experimental conditions, the SALAAM resulting WordNet 1.5 sense tags of the English corpus are projected onto the Spanish words in SP_SV2LS which includes the test set SPSV2AW. The aim is to measure the quality of sense-annotations of the projected sense tags onto the Spanish tokens in the SPSV2AW test set.

1. **Spanish All-Words with GL (AWGL)**

SP_SV2LS and BSSSJ are pseudo-translated using the GL machine translation system. The pseudo-translated portion of the corpora is GL for both directions, i.e. GL Spanish translations corresponding to source English BSSJ and GL English translations corresponding to the Spanish SP_SV2LS.

2. **Spanish All-Words with SYS (AWSYS)**

Similar to Condition 1 where SP_SV2LS and BSSSJ are pseudo-translated using the SYS machine translation system.

3. **Spanish All-Words with intralanguage pseudo-translation merge post-alignment (AWGLSYS)**

The aligned corpora resulting from SYS and GL are merged before the creation of source sets in the SALAAM tagging cycle. This condition aims at increasing the variability of contexts, thereby allowing for more source type sets.

4. **Spanish All-Words with post-tagging translation intersection merge (AWGLSYS-I)**

The tagged test set resulting from conditions 1 and 2 are intersected where only common sense tags of the shared noun instances are evaluated with the rest of the uniquely tagged words in SPSV2AW. The intersection of the tag sets weeds out some of the possible noisy tags from the tag set. Some noun instances are excluded if they occur in both test tag sets resulting from conditions 1 and 2 and they share no tags in common.

5. **Spanish All-Words with post-tagging translation union merge (AWGLSYS-U)**

The tagged test set resulting from conditions 1 and 2 is merged with a union operation where all sense tags for shared noun instances are evaluated with the rest of the uniquely tagged words in SPSV2AW. This condition allows for more coverage of the data. The union of the tag sets allows for the inclusion of more noisy tags but improves coverage.

Experimental Parameters

The parameters used here are the same as those used for SALAAM in Chapter 3. Moreover, in this evaluation, the sense selection criterion is a parameter, and it is set to MAX.

Hypotheses

1. Hypothesis 1

Results from Condition 1 AWGL and Condition 2 AWSYS will illustrate significant precision improvement over the RBL baseline. SALAAM is more informative in its sense-annotations than random sense choice.

2. Hypothesis 2

Results from Condition 4 AWGLSYS-I will show precision improvement over conditions AWGL and AWSYS. AWGLSYS-I is an exclusive voting scheme that aims at improving the tagging quality in terms of precision. We expect the recall to decrease since several items will be excluded.

3. Hypothesis 3

Results from Condition 5 AWGLSYS-U will show recall improvement over Condition 1 AWGL and Condition 2 AWSYS. AWGLSYS-U is an inclusive voting scheme which aims at maximizing the coverage of the test data. Since `scorer2` rewards partial credit, then we expect precision to decrease owing to the introduction of noisier tags.

4. Hypothesis 4

Results from Condition 4 AWGLSYS-I will illustrate better precision than Condition 5 AWGLSYS-U.

5. Hypothesis 5

Results from Condition 5 AWGLSYS-U will illustrate better recall than Condition 4 AWGLSYS-I.

6. Hypothesis 6

Results from Condition 3 AWGLSYS will show precision and recall improvement over Condition 1 AWGL and Condition 2 AWSYS.

7. Hypothesis 7

Results from Condition 3 AWGLSYS will show comparable precision to Condition 4 AWGLSYS-I and comparable recall to Condition 5 AWGLSYS-U. The key ingredient here is the variability in translation achieved by condition AWGLSYS.

Results

Table 4.6 illustrates the results obtained by applying the different conditions to SPSV2AW. The tagged test set for each condition is evaluated against the gold standard, AWGS, using `scorer2` software set to the fine-grain evaluation mode.

Conditions	P%	R%	COV%
RBL	27.7	21.9	79.01
AWGL	38.6	18.3	47.45
AWSYS	36.9	17.2	46.76
AWGLSYS	39.1	21.5	55.02
AWGLSYS-I	39.8	18.3	45.99
AWGLSYS-U	37.1	21.2	57.18

Table 4.6: Results in % for RBL, and the different evaluation conditions of test set SPSV2AW

Results for conditions AWGL and AWSYS achieve statistically significantly better precision scores than RBL using the `Zscore` significance test. Hypothesis 1 is accepted. We note the significant drop in coverage from 79.01% for RBL to 47.45% and 46.76% for AWGL and AWSYS, respectively. It is worth noting the relatively low coverage of the baseline in general. This RBL coverage score demonstrates the fact that some sentences are not aligned as they exceeded the cap of 70-token per sentence length set by the automatic alignment software. More importantly, the coverage level reflects the fact that not all tagged Spanish nouns in AWGS corresponded to nouns on the English side of the corpus.

AWGLSYS-I condition shows better precision results than AWGL and AWSYS conditions, which allows us to accept hypothesis 2. However, AWGLSYS-I maintains the same level of recall with a slight loss in coverage relative to the coverage achieved by conditions AWGL and AWSYS.

In accordance with hypothesis 3, AWGLSYS-U condition exhibits statistically significantly better recall results than conditions AWGL and AWSYS, respectively. Moreover, we observe a significant improvement in coverage, from a maximum of 47.45% for the AWGL condition to 57.18% for the AWGLSYS-U condition. We also note that precision for condition AWGLSYS-U is at an expected midpoint between precision for AWGL and AWSYS.

Results obtained in condition AWGLSYS-I achieve better precision than AWGLSYS-U supporting hypothesis 4.

A significant increase in recall and coverage is obtained in condition AWGLSYS-U over AWGLSYS-I, which supports hypothesis 5.

The intralanguage post-alignment merge condition, AWGLSYS, produces the best results on all three measures relative to the individual MT system conditions AWGL

and AWSYS. We see an improvement in precision and a significant improvement in both recall and coverage supporting hypothesis 6.

AWGLSYS scores a precision of 39.1% as opposed to a precision of 39.8% for condition AWGLSYS-I. There is no significant difference between the two conditions on precision. Similarly for recall, conditions AWGLSYS and AWGLSYS-U achieve comparable results with no significant difference. These results support Hypothesis 7. In terms of coverage, AWGLSYS condition yields comparable results to AWGLSYS-I.

In addition to the results reported in Table 4.6, FSBL yields a precision of 43.2% and a recall value of 34.1% at a coverage of 79.01%. FSBL is not included in the results table as it is not an appropriate baseline for this task, as discussed earlier in the baseline section. We note, however, that FSBL achieves significantly better results than any of the SALAAM conditions on all measures.

Discussion

This evaluation lays the basis for a robust system of bootstrapping sense tagging for a new language with scarce automatic knowledge resources. We practice caution because this evaluation has many approximations due to the limitation on resources available, nonetheless, to our knowledge, this is the first attempt at bootstrapping sense tagging automatically for a language with limited resources. Despite the modest results when compared to SALAAM source tagging as discussed in Chapter 3, precision results are encouraging as they indicate a significant departure from the RBL precision results, which are at a noticeably lower coverage level than those seen in Chapter 3. The recall and coverage are very low compared to the results achieved for these measures on the source language.

But before discussing the details of these results, the results from FSBL beg the question of why use SALAAM at all if FSBL achieves higher scores on all measures. The response lies in language distance and homonymy. In cases where source and target languages are close and one of the languages has an ontology that is arranged with the most frequent sense listed first,²² using FSBL as a bootstrapping method is worthwhile. Spanish and English are relatively close languages as they have a shared ancestor among other things. The common origin results in preserving ambiguity, leading to many cases of semantic overlap. In these cases polysemous nouns tend to be used in the same way with regular polysemy and metonymy, even with homonymy. For example, the polysemous word *interest* in English has the same meaning as the word **interés** in Spanish. The problem arises when using FSBL where languages are distant. The languages grow apart and pragmatic differences start playing a significant role in the correct ordering of senses. It is especially worse where an English word is homonymous. For instance, the first sense of *bank* in WordNet 1.5 is *side of the river* sense. Yet, when *bank* is aligned with **bnk** in Arabic, the appropriate sense for the

²²Or default sense or most typical sense

Arabic word is the *financial institution* sense of *bank*.

The following is a discussion of the different factors that affect precision, recall and coverage.

- **Precision**

The first factor that affects precision is translation quality. Both MT systems produce close to gisting quality translations on the English side of the parallel corpus where many of the ambiguous words are not even translated. Qualitatively inspecting the English translation output suggests that the quality of translation into English is worse than the translation from English for both MT systems. For example, the following Spanish sentence:

Las artes caminan hasta que se produce una quiebra; entonces su presencia rompe con lo que fueron modelos arregostados en cómodas repeticiones.

is translated into English as follows:

The arts walk until a crash takes place; then your presence breaks up with what you/they were model arregostados in comfortable repetitions.

The presence of the Spanish words in the source type sets is noisy resulting in deflated precision. Moreover, some ambiguous words that are homonymic in nature are translated into the wrong word in English simply because the MT system defaults to the most common sense for an ambiguous word.

pseudo-translating the source side of the parallel corpus has a cascading negative effect on the automatic POS tagging quality since many of the words are not translated. Many tokens are mistagged as nouns which are eventually included in source type sets – tokens being identified as nouns which are not nouns.

Another issue that affects precision is the presence of *faux amis*. This is a phenomenon that is present in languages that are close to one another. *Faux amis*, occur when a word in Spanish is left untranslated, and it exists in English in the same orthographic form. For example, **sensible** in Spanish corresponds to *sensitive* in English, not *reasonable*, which is the meaning of the English word *sensible*.

Misalignments constitute a huge bottleneck that seriously affect precision. Admittedly, as mentioned in Section 4.2, MT alignments are more consistent than HT alignment, yet MT is still a source of considerable noise in the source sets. For example, in the source set (*ABANDON ABANDONMENT DERELICTION DROPOUT FEELING NEGLECT*), *FEELING* is an obvious outlier even though

it is a related word; it results from a misalignment. As discussed in Chapter 3, such misalignments, especially if they are monosemous, could yield bias in the wrong direction for the **NG** sense selection algorithm.

Sense granularity mapping between the Spanish WordNet and the English WordNet 1.5. is an issue in this evaluation. As mentioned in Section 4.3.6, there are senses in AWGS that do not exist in WordNet 1.5, and vice versa, which is reflective of the different granularity size of the concepts in these two languages. For example, for AWGS, there exist 50 approximate cases as described in the section describing the gold standard. Twenty five of these cases are tagged by some SALAAM condition with only one of them tagged correctly.

- **Recall**

Several factors affect recall. Due to the quality of the pseudo-translations, many of the words are left untranslated, which leaves them un-amenable to forming source type sets; consequently, they are left untagged. Approximately 20% of the potential noun instances form singleton source type sets, which means they are not passed onto the **NG** sense selection algorithm. For example, out of the total 57791 noun instances in experimental condition AWGLSYS, more than 9753 noun instances form singleton source type sets, therefore, they are excluded from the tagging process.

Three of the sentences, comprising 24 sense tagged noun instances, are excluded from SPSV2AW since they exceed the length limit set by the token alignment software.

Divergences in the POS tags between Spanish and English lead to low recall. These divergences result from both poor quality of the automatic POS tagging of the pseudo-translated English, and genuine divergences where some nouns in English are translated into other POS tags in Spanish and vice versa.

Furthermore, the human annotator manually altered some of the POS tags in the corpus. She also changed the tokenization of several instances, thereby creating compound nouns in Spanish. This resulted in 21 cases of compound nouns in AWGS. Only one of these compound nouns was found and correctly tagged in experimental condition AWGLSYS.²³

- **Coverage**

The same factors that affect recall affect coverage. The coverage scores obtained from FSBL and RBL clearly indicate an upper bound on coverage achieved in this evaluation. The scores indicate that more than 20% of the Spanish noun instances tagged in AWGS do not exist for SALAAM. This is mainly due to the

²³Compounds are automatically created on the target side of the parallel corpus when there is a one-to-many correspondence between the source and target alignments produced by GIZA++.

POS divergences, which is discussed above as one of the factors affecting recall. The even lower scores yielded by the SALAAM conditions are a reflection of the nouns that are excluded due to sentence length problems or singleton source type set issues.

4.3.7 General Discussion

We note the difference in performance for the Arabic and Spanish projected tagging. Arabic yields better results in terms of overall precision. Yet, it is hard to compare across both evaluations.

As an experimental setup, the Spanish evaluation is blind where the annotator relies on a monolingual resource for tagging the Spanish text without having access to the English translations at all. Yet, this evaluation suffered the effect of relying on a pseudo-translated source corpus. A more realistic approach would be to have the SP_SV2LS manually translated to English and then perform the same evaluation with good quality English source data. Nonetheless, this section provides a rigorous framework for performing the task of evaluating projected sense tags on the target language side of a parallel corpus.

4.3.8 Summary

In summary, SALAAM is devised as a new technique for word sense tagging a target language with source language resources. The quality of tagging of the target language using SALAAM is evaluated for two languages: Arabic and Spanish. The results obtained from Arabic demonstrate that of 90.5% of the correct tags for English noun instances are correct tags for Arabic zoning in on the commonality of sense usage cross-linguistically, in effect, quantifying meaning characterizations for a language with poor resources via its shared sense usages with rich resources. On the other hand, we perform a fully automated blind evaluation of the quality of projected tagging for Spanish data. The results obtained are modest even though they significantly improve on a random baseline. The main reason for the modest performance is attributed to the use of source pseudo-translations accompanied with inconsistencies in alignments, therefore detrimentally affecting the quality of the tagging. But nonetheless, the technique presented is a new technique that is fully automated and, except for the parallel corpus and gold standard set, requires minimal resources.

4.4 Feasibility of bootstrapping a WordNet style ontology for Arabic

4.4.1 Introduction

Efforts in the domain of ontology creation have mostly been manual. EuroWordNet [79] exists for several languages: Dutch, Spanish, French, Czech, Italian and Estonian; EuroWordNet interfaces these different Ontologies with the Internal Language Index (ILI). The bootstrapping method starts with monolingual dictionaries for the new language, and an ontology is created in the WordNet format. Apart from the immense time investment in the bootstrapping phase, the researchers are faced with the challenge of linking the created WordNet with existing WordNets and dealing with sense granularity issues which is one of the biggest challenges facing such an endeavor.

Having a method that leverages existing resources is a big plus as the manual task of creating an ontology such as WordNet is extremely expensive and genuinely daunting. The problem becomes even more challenging when the language in question is a language with scarce automatic knowledge resources such as Arabic. The method we are proposing here, in fact, a side effect of applying SALAAM to a parallel corpus, automatically bootstraps a WordNet for a new language by obtaining the mappings cross-linguistically, thereby bootstrapping the conceptual mapping. Given a large and diverse enough parallel corpus with good quality token alignments, this method can help bootstrap a large ontology for a new language from scratch.

In this section, we investigate the feasibility of bootstrapping a WordNet ontology for Arabic. The appeal of building a WordNet for Arabic is not only based on empirical grounds for computational linguistic applications, but also it allows for an exploration of interesting lexical semantic cross-linguistic variations — albeit at this stage exclusively paradigmatic. Like other languages, Arabic lexemes exhibit the full range of ambiguity attributes from regular polysemy to metonymy and homonymy. Lexical ambiguity in Arabic is further compounded by the writing system; as mentioned earlier, written texts in Arabic typically omit the short vowels leading to more ambiguity, creating false homonyms. For instance, the word **klya** in the written form could refer to *kidney*, *faculty* — college sense — or *completion*. In fact, **klya** is pronounced differently depending on the intended meaning; therefore, when it is referring to *kidney*, it is pronounced **kilya**, and when it is referring to *faculty* it is pronounced *koleya*. Yet, the writing system does not capture this difference. Context is constantly used by speakers and readers of Arabic text to resolve this ambiguity online. In this particular example, *faculty* and *completion* is a case of genuine homonymy, though the Arabic *completion* sense is more of an adjective than a noun. Methods relying on context and/or vowel restoration are very useful in this level of lexical ambiguity resolution.

4.4.2 Evaluation

With that intent in mind, we evaluate the 526 word instances of Arabic that are deemed correctly tagged using the English WN17pre (see Section 4.3.5).

- **Same level sense granularity: Arabic and English words are equivalent**

We observe that a majority of the ambiguous words in Arabic are also ambiguous in English; they preserve ambiguity in the same manner; in Arabic, 368 noun tokens corresponding to 162 noun types,²⁴ are at the closest granularity level with their English correspondent;²⁵ For instance, all the senses of *care* apply to its Arabic translation **E3nAyA**; this is illustrated in Figure 4.3.

1: *care, attention, aid, tending: the work of caring for or attending to someone or something; "no medical care was required"; "the old car needed constant attention"*
2: *caution, precaution, care, forethought: judiciousness in avoiding harm or danger; "he exercised caution in opening the door"; "he handled the vase with care"*
3: *concern, care, fear: an anxious feeling; "care had aged him"; "they hushed it up out of fear of public reaction"*
4: *care: a cause for feeling concern; "his major care was the illness of his wife"*
5: *care, charge, tutelage, guardianship: attention and management implying responsibility for safety; "he is under the care of a physician"*
6: *care, maintenance, upkeep: activity involved in maintaining something in good working order; "he wrote the manual on car care"*

Figure 4.3: English WN17pre entries for *care*

It is worth noting that the cases where ambiguity is preserved in English and Arabic are all cases where the polysemous word exhibits regular polysemy and/or metonymy. The instances where homonymy is preserved are borrowings from English. Metonymy is more pragmatic than regular polysemy [14]; for example, *tea* in English has the following sense:

This sense of *tea* in Figure 4.4 does not have a correspondent in the Arabic **shay**. Yet, a word like *lamb* in English has the metonymic sense of MEAT and this is preserved in Arabic. Researchers building EuroWordNet have been able to devise a number of consistent metonymic relations that hold cross linguistically such as *fabric/material*, *animal/food*, *building/organization* [78, 82]. In Arabic

²⁴Arabic words are not lemmatized; therefore, some cases are included as both plural and singular forms.

²⁵This means that all the English senses listed for WN17pre are also senses for the Arabic word.

3: *a reception or party at which tea is served; "we met at the Dean's tea for newcomers"*

Figure 4.4: Metonymic sense of *tea* in WN17pre

these defined classes seem to hold, yet this specific case of *tea* and *party* does not hold. In Arabic, the specific sense is expressed as a *tea party* or **Haflet shay**.

- **Arabic word equivalent to English word subsense**

In this evaluation set, there are 122 instances where the Arabic word is equivalent to a subsense only of the English word. The 122 instances correspond to 78 word types. An example is illustrated in Figure 4.5; the correct sense tag assigned by SALAAM to *ceiling* in English is sense 1, which is correct for the Arabic word **sqf**. Yet, the other 3 senses are not correct translations for **sqf**; for instance, sense 2 would be translated as **Irtifa3** and sense 4 as **3low**.

1: *ceiling: the overhead upper surface of a room; "he hated painting the ceiling"*
2: *ceiling: (meteorology) altitude of the lowest layer of clouds*
3: *ceiling, cap: an upper limit on what is allowed: "they established a cap for prices"*
4: *ceiling: maximum altitude at which a plane can fly (under specified conditions)*

Figure 4.5: English WN17pre senses for *ceiling*

This case is particularly dominant where the English word is homonymic. By definition, homonymy is when two independent concepts share the same orthographic form, in most cases, by historical accident. Homonymy is typically preserved between languages that share common origins or in cases of cross-linguistic borrowings. Owing to the family distance, preserving homonymic ambiguity holds the least between English and Arabic. For example, *tower* in English has the following sense illustrated in Figure 4.6, which does not exist at all for the Arabic word **brj**.

3: *a powerful small boat designed to pull or push larger ships*

Figure 4.6: Homonymic sense for *tower* in WN17pre

Therefore, for most homonymic polysemous words in English, the Arabic translation corresponds to one of the homonymic senses only.

- **English word equivalent to Arabic subsense**

35 instances, corresponding to 18 type words in Arabic, are manually classified as more generic concepts than their English counterparts. For these cases, the Arabic word is more polysemous than the English word. As an example, Figure 4.7 shows the word *experience* listed with 3 senses in WN17pre. All 3 senses are appropriate meanings of the Arabic word **tjrba** but they do not include the SCIENTIFIC EXPERIMENT sense covered by the Arabic word.

1: experience: the accumulation of knowledge or skill that results from direct participation in events or activities; "a man of experience"; "experience is the best teacher"
2: experience: the content of direct observation or participation in an event; "he had a religious experience"; "he recalled the experience vividly"
3: experience: an event as apprehended; "a surprising experience"; "that painful experience certainly got our attention"

Figure 4.7: WN17pre senses for *experience*

From the above points, we find that 62% of the ambiguous Arabic words evaluated are conceptually equivalent to ambiguous English words. This finding is consistent with the observation of the builders of EuroWordNet. Vossen, Peters, and Gonzalo (1999) find that approximately 44-55% of ambiguous words in Spanish, Dutch and Italian have relatively high overlaps in concept and the sense packaging of polysemous words [78]. 31% of the ambiguous Arabic words correspond to specific subsenses of the English word and 7% of the Arabic words are more generic than the English words.

The encouraging results obtained from the manual analysis of a sizeable sample of the Arabic tagged data suggests that bootstrapping an Arabic WordNet style ontology is a feasible task.

4.4.3 Levels of representation

As mentioned earlier, Arabic has a templatic syntax; roots are transformed into stems based on a templatic fit. For example, the root **ktb** becomes **kitab** based on the template **fi3al**. Stems are usually embedded with prefixes and suffixes creating surface forms that are the words as they appear in text. Reducing a surface form to a stem is relatively easy given a light stemmer [16]. In traditional Arabic monolingual dictionaries, the entries are in root form. Yet, the writing system hardly ever has the roots in raw form. In the following discussion, we examine issues regarding the appropriate representation level for an Arabic WordNet.

- **Roots**

As mentioned earlier, words in Arabic, as a Semitic language, have roots. Roots are the underlying forms from which stems and surface forms generate. The dynamic role attributed to roots might be a result of pedagogical factors: language is taught in schools with an emphasis on roots; dictionary entries are indexed by their roots.

Most words in Arabic can be reduced to 3 or 4 letter roots. Roots are typically consonant based. Arabic has generative templates that lead to the creation of stems. Roots are highly generative and typically very ambiguous. For instance, a word like **sh3r** means *hair*, *poetry* or *to feel*. This could be treated as a case of homonymy that is resolved by applying the appropriate template; therefore, the stem for *hair* is **shA3r**, for *poetry* **shi3r** and for *to feel* it is **shaA3rA**. Likewise, the root **Hrm** generates **Haram** as in *shrine*, *sanctuary*, *wife* or *forbidden*; it is also the root for the *clothes worn by pilgrims* as in **iHram**, as well as the root for thief as in **Haramy**.

Due to the pervasive ambiguity in the root representation, one would expect a huge overlap between the different POS databases in an Arabic WordNet.

We find the option of creating an ontology based on roots theoretically elegant, especially if the templates are not ambiguous. A root based ontology will have to be generative and underspecified. The main bottleneck is extracting the root from a surface level representation since words do not occur in their root form in written nor spoken Arabic. Several off-the-shelf morphological analyzers may be utilized to reduce surface forms to their corresponding roots, yet coverage remains a severe bottleneck [16, 11].

- **Stems**

A stem based ontology is a more direct approach to building an enumerative WordNet style sense inventory. Empirically, Arabic stems are more accessible by computational systems. Texts are written in surface form but easily transformed to stems (see Section 4.3.5 above). Stems are more distinguishable as different POS tags based on the templates they correspond to in Arabic. The main problem with stems is normalization; the same words meaning the same thing may be written in various ways. For example, the word for *schools* in Arabic maybe **madares** or **madrasat**. The second form is mostly predictable but the former form is not. Issues also arise with infixing, depending on the case of the word in question; words may have different endings. For example, *authors* in Arabic is either **mo2leffeen** or **mo2leffwn**. These minor hurdles are surmountable with the availability of good tokenizers and morphological analyzers.

Choosing the appropriate level of representation is an issue worth in-depth investigation. Our preliminary qualitative assessment calls for using the most direct approach at the beginning and then refining the ontology with some form of hybrid representation of both roots and stems in a multidimensional WordNet representation.

4.4.4 Summary

SALAAM is explored as a method for seeding a WordNet style ontology for Arabic. By quantitative inspection, the approach seems promising. We discuss different issues of representation for Arabic specifically. We conclude that stems as a first step are the appropriate level of representation for the entries in such an ontology.

Chapter 5

Exploration into Bootstrapping Supervised WSD

5.1 Introduction

It has been established that supervised WSD systems yield better results than unsupervised systems [40]. Yet, tagged data is not always available for training. Indeed, lack of training data is a very severe bottleneck for supervised systems. One of the goals of the SENSEVAL exercises is to create large amounts of sense-annotated data for supervised systems [40]. The problem is ever more challenging when dealing with a language with scarce knowledge resources. Typically, when confronted with a new low density language, researchers are preoccupied with building tools and knowledge resources before seriously observing WSD issues. Despite its central role to most NLP applications for any language, WSD is deemed too complicated.

One of the goals of SALAAM is to provide large amounts of sense-annotated data in several languages simultaneously to bootstrap supervised WSD systems, thereby, loosening the bottleneck on data acquisition for training supervised systems. Sense annotations yielded by SALAAM are noisy compared to manually tagged data but in the absence of an alternative they serve as a good initial launching board. Explicitly, this chapter explores the nature of the trade-off between small amounts of cleanly tagged data versus large amounts of noisy data for training in a supervised setting.

Most supervised WSD systems follow the canonical training-testing paradigm. The key idea in most supervised WSD systems is that senses are viewed as classes, rendering the problem an explicit classification problem. The goal of the WSD system is to assign test data items to the correct classes based on learning properties of the classes in a training period. Most supervised systems utilize machine learning algorithms. The machine learning paradigm may be briefly described as follows:

In the **Training Phase**, given sufficient training examples per class, the system extracts relevant features from the context of the word in question creating a feature vector;¹ valid features could be part of speech tags [10], syntactic features [25], con-

¹The features are at the crux of any classification system; different types of classifiers and ensemble classifiers have different merits, however, the importance of the features can not be over stressed.

text n-grams [61, 63], or a combination of the different contextual features [12]; the machine learning system — the *learner* — learns estimated parameters based on associating a class (sense) or set of classes with features extracted from the training examples. In summary, the *learner* learns parameters from explicit associations between the class and the features, or combination of features, that characterize it.

In the **Testing Phase**, given a new test item, the supervised WSD system extracts features based on the same conditions that are used in the training phase. According to the learned estimated parameters acquired in the training phase, a prediction process takes place where the *learner* predicts the best class for the new test item. Consequently, the test item is annotated.

Needless to say, such systems are very sensitive to the training data. Training data should provide ample coverage of the potential classes. Majority of approaches for WSD within a supervised framework attempt to ensure that the training and test data are from the same genre, domain, and that the training data provides sufficient coverage of the possible senses.

In this chapter, we examine various issues in connection with bootstrapping a typical supervised method for WSD using SALAAM annotated data for training. This method aims at alleviating the training data annotation bottleneck for most supervised systems. We investigate the first phase in an iterative approach to bootstrapping a supervised system using unsupervised sense-annotations. In a bootstrapping approach, the need arises for examining the different factors affecting the supervised system’s performance. Accordingly, we discuss different parameters as components in a fitness function that can potentially be automatically applied to the unsupervised training data in order to ensure/predict good classification performance for the supervised WSD system. It is worth emphasizing that training on data that results from an unsupervised approach renders the whole approach here unsupervised for this task even though it utilizes the canonical learning paradigm.

The layout for this chapter is as follows. After stating the problem being addressed and the motivation behind this work in Section 5.2, we discuss related work in the area of bootstrapping WSD systems in Section 5.3; Section 5.4 reviews the particular supervised model and application that is used as a test bed for the bootstrapping technique presented here; Section 5.6 describes an empirical investigation into the feasibility of such an approach; this is followed by a general discussion of the results with a close look at the different parameters affecting the performance of the bootstrapping approach in Section 5.8.

5.2 Motivation

The availability of sense-annotated training data is a serious bottleneck for supervised WSD systems. Previously, researchers have investigated using dictionaries and supervised methods with clean data for iteratively bootstrapping the tagging effort [58, 85].

These approaches do a good job when the resources are available for a language. The problem is ever more challenging when we migrate to a new language with scarce knowledge resources.

Unlike previous approaches, we propose to start the supervised tagging system with an unsupervised seed set. Unsupervised systems do not have as much knowledge/tool requirements as supervised systems; moreover, they are less language and corpus dependent. Therefore, in this chapter, SALAAM is presented as a means of providing large amounts of sense-annotated data with the aim of relieving supervised systems from the training data acquisition bottleneck.

SALAAM is an appealing approach as it not only provides sense tagged data in one language, rather in two simultaneously. In Chapter 4, we obtain encouraging results for the projected sense tags on a second language; furthermore, the results are significantly better than a random baseline and therefore should provide the appropriate signal for supervised learners amidst the noise. Accordingly, SALAAM has the advantage of providing a multilingual framework for solving this problem.

5.3 Related Work

This chapter relates to work in monolingual bootstrapping by Gale, Church and Yarowsky [27], Yarowsky [85], and Mihalcea [58].

The first study by Gale et al. (1992) is the earliest study to our knowledge which directly discusses the feasibility of bootstrapping a WSD system using noisy data. They give an empirical evaluation of the level of degradation in the WSD system's performance as they introduce different levels of error. (For a review of that paper see Section 2.4.2, Chapter 2) They conclude that their system is tolerant to noise in the training data. It is worth noting that they look at 6 data items each with 2 senses only and with a bounded number of examples per item. In this chapter, we ask the same question: can we bootstrap supervised WSD systems using noisy examples for training.

Research by Yarowsky and later Mihalcea is different from the research presented by the previous study. The question is asked differently. The focus is more on the bootstrapping technique rather than on the quality of the data. They address the issue of data quantity while maintaining the good quality level of the training examples. Both investigations present algorithms for bootstrapping supervised WSD systems using clean data based on a dictionary or ontology resource. The general idea is to start with a clean initial seed and iteratively increase the seed size to cover more data.

In Yarowsky's work [85], he starts with a few tagged instances to train a decision list approach for tagging unlabeled data. The initial seed is manually tagged with the correct senses based on entries in Roget's Thesaurus. The approach is unsupervised. He reports very successful results — 95% — on a handful of data items.

A directly comparable study to our exploration in this chapter, however, is work by

Mihalcea [57, 58]. She bases her bootstrapping approach on a generation algorithm, GenCor. GenCor creates seeds from monosemous words in WordNet, Semcor data, Sense tagged examples from the glosses of polysemous words in WordNet, and other hand tagged data if available. This initial seed set is used for querying the Web for more examples and the retrieved contexts are added to the seed corpus. The words in the contexts of the seed words retrieved are then disambiguated. The disambiguated contexts are then used for more querying of the Web for more examples, and so on. It is an iterative algorithm that incrementally generates large amounts of sense tagged data. The words that are found are restricted to either part of noun compounds or internal arguments of verbs.

Mihalcea reports results of 69.3% precision on the English SENSEVAL 2 allwords task using the bootstrapped corpus for training an instance-based-learning supervised WSD system. When applying GenCor results as the training examples for her supervised system in the SENSEVAL 2 Lexical Sample English exercise, she shows that the approach yields results comparable to those obtained when training with hand tagged data. Mihalcea reports the results for 6 items of the 29 items in the Lexical Sample exercise. Table 5.1 compares her results obtained by training the learning supervised system on hand annotated examples against those obtained by training on the automatic generated corpus using GenCor. We show only the precision percentages using `scorer2` in the fine grain mode; we have added a column here in the table where we calculate the Performance Ratio (PR) (see equation (5.2) below) of scores obtained using GenCor examples to those using hand tagged examples.

	Hand Tagged Data		GenCor Data		
Nouns	Training Size	Prec.	Training Size	Prec.	PR
art	123	65.4%	265	73.1%	1.12
chair	121	82.5%	179	87.3%	1.05
channel	78	34.1%	1472	40.9%	1.19
church	81	63.9%	189	58.3%	0.91
detention	46	87.5%	163	83.3%	0.95
nation	60	73.1%	225	69.5%	0.95

Table 5.1: Comparative results obtained by Mihalcea’s bootstrapping system when training an instance-based learning supervised WSD system using both human tagged data and GenCor tagged data as training examples

5.4 Empirical Layout

Similar to the experimental design presented by Mihalcea [58] and described in Section 5.3 above, we compare results obtained by a supervised WSD system for English using

human tagged examples against SALAAM tagged examples for training. We use the same test set used by Mihalcea, the data from the SENSEVAL 2 English Lexical Sample task. The supervised system we use is the system developed and tested by University of Maryland for the SENSEVAL2 English Lexical Sample exercise, UMSST [12]. This supervised system utilizes a Support Vector Machine (SVM) learning paradigm for the classification and tagging of the items in question.

5.5 University of Maryland Supervised Sense Tagging system (UMSST)

The UMSST system is created in the classic supervised learning framework. Each word — test item — that will be tagged is considered an independent classification problem. In the learning phase, the training examples pertaining to an ambiguous word are analyzed, creating feature-value pairs labelled with the correct class (sense). These data are used to estimate the parameters for the learner in the supervised system producing a trained classifier. The classifier is then applied to the unseen test data instances; for each instance, the classifier predicts the appropriate sense tag.

The features used for UMSST are contextual features with weight values associated with each feature. The features are extracted from the immediate context of the labelled word. The text is tokenized with an English specific tokenizer. Then three types of features are extracted from the tokenized text: wide context features; narrow context features; and grammatical features.

The wide context features use all the tokens in the paragraph where the labelled instance occurs. The narrow context feature is a collocational feature; it only takes the tokens within a fixed window size n surrounding the labelled word. In this implementation of UMSST, n is set to 3 tokens on each side of the labelled instance. As for grammatical features: the context of the instance is parsed using a dependency parser [48, 51]; syntactic tuples such as verb-obj, subj-verb, etc. are extracted.

For example, if the word *feature* in the second sentence in the preceding paragraph is the word of interest, its narrow collocational features will be $L = \{context, narrow, the, is, a, collocational\}$ for the words to the left and right of the word *feature* when n is specified as 3 tokens. The grammatical features are *subj-of*(*be, feature*) and *mod-n*(*feature, context*). As for the wide context features, those are all the words in the paragraph as follows, $\{the, wide, context, feature, use, all, token, in, paragraph, where, the, label, instance, occur, ., narrow, is, a, collocational, ;, it, only, take, ..., etc.\}$.

Each feature extracted is associated with a weight value. The weight calculation is a variant on the Inverse Document Frequency (IDF) measure in Information retrieval. The weighting in this case is an Inverse Category Frequency (ICF) measure where each token is weighted by the inverse of its frequency of occurrence in the specified context of a specific labelled instance. For example, if the feature is the token *this* and it co-occurs with all instances of the senses of a given word, the ICF value for *this* is

very low, indicating that *this* does not contribute significant information as a discerning feature for any of the senses of the ambiguous word. On the other hand, if the ICF of a feature is high, it indicates that feature has discriminatory power among the senses of the polysemous word.

The learning approach used by UMSST is a Support Vector Machine (SVM) algorithm.² Similar to other learning paradigms, the system takes in the training instances for the word in question and yields a classifier; it takes a feature vector as input and produces a confidence function over all possible categories observed in the training data.

SVM is chosen as an appropriate learning framework because it can achieve high performance with very large numbers of features. Moreover, SVMs are known for their interpretability and robust theoretical basis. The version that is used in the UMSST experiments is an off-the-shelf implementation, *SVM – light*TM by Joachims [36].³

For each word in the Lexical Sample task, a family of classifiers is constructed, one for each of a given word’s senses. All the positive examples for a sense (S_i) are considered the negative examples of (S_j) where $i \neq j$. In the testing phase, similar feature vectors are created for the test data; the feature vectors are run with the SVM classifiers based on the parameters estimated in the training (learning) phase. The sense that yields the strongest positive response is selected. UMSST trains on the hand-tagged provided English Lexical Sample training data and is tested on the test data for the same exercise. The contexts for the English Lexical Sample data are defined by the organizers of the SENSEVAL2 task; A typical context spans 2 to 4 sentences in length on average.

The approach as described so far yields the results for the UMSST trained using human annotated examples. For purposes of this current evaluation, we are only interested in nouns.

5.6 Bootstrapping Evaluation

In this evaluation, we use the large amounts of English data sense tagged by SALAAM as described in Chapter 3. We create a system that generates the contexts for the SALAAM tagged examples automatically in the format of the SENSEVAL2 Lexical Sample training and testing data.

²see <http://www.computer.org/intelligent/ex1998/pdf/x4018.pdf> for interesting discussions about the advantages of Support Vector Machines.

³*SVMlight* is available at <http://www.ai.cs.uni.dortmund.de/svmlight>.

5.6.1 Test data

The SENSEVAL2 Lexical Sample test data (SV2LS-test) comprises 29 nouns with varying numbers of example contexts per noun. The nouns range in polysemy from 2 senses to 19 senses per noun. Table 5.2 lists some of the test items and their characteristic features.

Nouns	# Senses	Perp	# Cont	Nouns	# Senses	Perp	# Cont
art	17	4.92	104	fatigue	6	1.88	40
authority	9	3.73	99	feeling	5	2.83	52
bar	19	6.5	136	grip	6	3.25	50
bum	4	1.78	40	hearth	3	2.14	33
chair	7	1.66	63	holiday	6	1.73	31
channel	7	4.92	71	lady	8	2.83	52
child	7	2.14	63	material	16	4.92	74
church	6	2.46	65	mouth	10	4	71
circuit	13	9.19	85	nation	4	1.64	36
day	16	3.73	162	nature	7	4.59	52
detention	4	2.64	32	post	12	5.66	72
dyke	2	1.4	28	restraint	8	4.92	47
facility	5	2.83	58	sense	8	4.92	54
spade	6	2.3	32	stress	6	3.25	43
yew	3	1.85	28				

Table 5.2: Test items for SV2LS-test

The perplexity is calculated as $(2^{Entropy})$. Entropy is measured as follows [40]:

$$H(X) = - \sum_{x \in X} p(x) \log_2(p(x)) \quad (5.1)$$

where x is a sense for a polysemous noun and X is the set of its senses.

Like entropy, perplexity is a measure of the bias in the distribution of the contexts — the training examples — for the different senses. The less perplexity, the less entropy, the more bias there is, i.e. the higher the perplexity the more the distribution of contexts is closer to a uniform distribution. Perplexity is thought of as the weighted average sum of choices a random variable has to make [37]. therefore if the senses are distributed uniformly, then every sense is equally likely to be chosen. For example if there are 3 senses to choose from and the training contexts are the same for all 3, then the perplexity is going to be 3 which means that the learner is choosing from all 3 senses with equal probability.

The average perplexity in the test set is 3.47. The average number of senses is 7.93. The total number of contexts for all senses of all words in the test set is 1773.

5.6.2 Hand-Tagged Training Data

The hand-tagged training data is obtained from the SENSEVAL2 Lexical Sample data provided by the organizers of the SENSEVAL exercise.⁴ This training data corpus comprises 44856 lines and 917740 tokens. Table 5.3 illustrates the characteristics of the hand-tagged training data.

Nouns	# Senses	Perp	# Cont	Nouns	# Senses	Perp	# Cont
art	17	6.5	219	grip	6	3.48	100
authority	9	4.92	196	hearth	3	2.46	66
bar	19	7.46	271	holiday	6	1.93	65
bum	4	1.99	81	lady	8	2.64	99
chair	7	2	138	material	16	6.06	149
channel	7	5.28	138	mouth	10	3.48	130
child	7	2.83	130	nation	4	1.8	78
church	6	2.46	136	nature	7	4.59	100
circuit	13	8.57	174	post	12	5.66	145
day	16	4.59	321	restraint	8	4.92	94
detention	4	2.3	63	sense	8	4.92	111
dyke	2	1.61	59	spade	6	2.83	65
facility	5	3.25	120	stress	6	3.48	89
fatigue	6	2.14	77	yew	3	1.65	57
feeling	5	2.83	116				

Table 5.3: Characteristics of hand-tagged training data for the SENSEVAL2 English Lexical Sample task

As depicted in Table 5.3, similar to the test set, there are 29 nouns ranging from 2 senses to 19 senses per noun; they average 7.93 senses per noun. Perplexity is measured by equation (5.1), the average perplexity across the different nouns in the hand-tagged training data is 3.75. The total number of contexts is 3587 contexts for the whole set.

As expected, there is a close affinity between the test set and the hand-tagged data used for training; The perplexity values are close with 3.47 in the test set and 3.75 in the hand-tagged training data. The number of contexts in the training data nearly doubles that of the test set with 3587 contexts in the hand-tagged data and 1773 contexts for the test set. Figure 5.1 plots the Pearson R correlation coefficients' of the sense distributions for the test set and the hand-tagged training data across the 29 nouns. As depicted by the graph, the correlations are all positive ranging from 0.90 to 1.0.

Figure 5.2 below further emphasizes the relationship between the hand tagged training data and the test data. In the graph, we plot the trend lines with the test

⁴<http://www.senseval.org/>

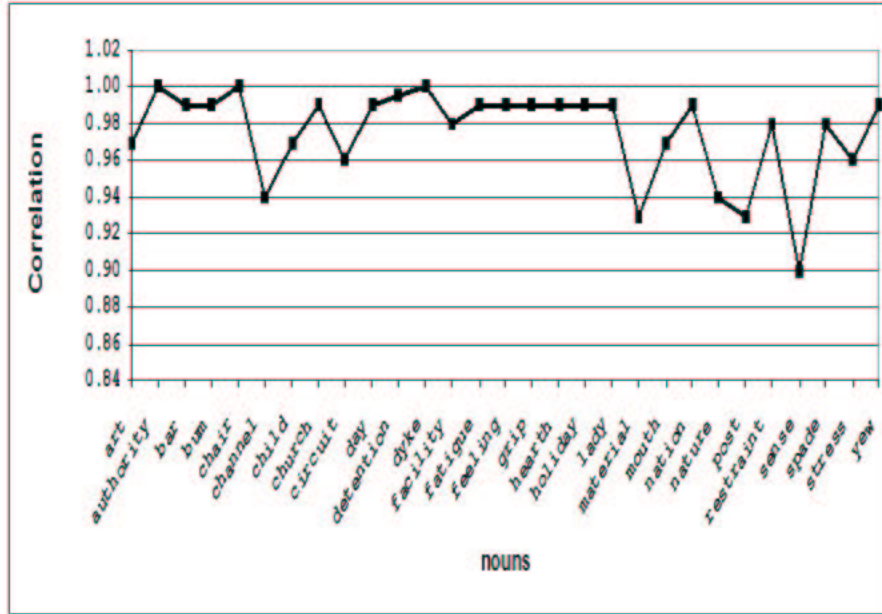


Figure 5.1: Sense distribution correlations across different nouns in the test data and hand-tagged training examples

data perplexity values depicted with a solid line and the hand-tagged training data perplexity values represented as the hashed line. The X-axis represents the nouns in alphabetical order while the Y-axis the perplexity values. As shown in the graph, the two lines almost overlap.

The close correspondence between the training and the test data in terms of the sense distribution and perplexity shows that the former is a very good representation of the latter.

5.6.3 Gold Standard

In this evaluation, we are comparing the performance of the same supervised system UMSST trained on SALAAM tagged data (UMSST-SALAAM) against the performance of the system when trained on hand-tagged data (UMSST-human). Therefore, UMSST-human’s performance is the gold standard. It is worth noting that in terms of the SENSEVAL 2 results, UMSST-human is a very representative supervised system as its scores are *middle-of-the-pack* scores. Table 5.4 shows the scores obtained by UMSST-human on the test data described in Section 5.6.1 when trained on the data described in Section 5.6.2. The scores are calculated using `scorer2` in the fine grain mode. The average precision score over all the items is 65.3%. Throughout the rest of

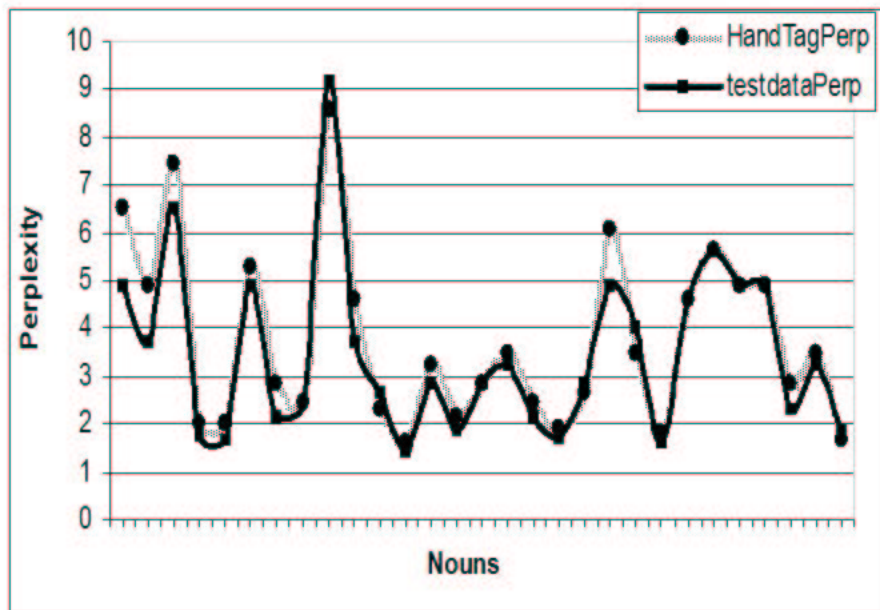


Figure 5.2: Trend lines of the perplexity measure for test data and hand-tagged training data

this chapter, we only show precision scores; coverage is at a 100%, recall and precision are the same.

5.6.4 SALAAM Training Data Corpora

We use the corpora sense tagged in Chapter 3 and the sense tagged HT corpora from Chapter 4 as training data. For the purposes of this evaluation, we are using the English corpora only. The training data in this case is not hand annotated. We exclude the SENSEVAL2 Lexical Sample test data (SV2LS-test)(see Section 5.6.1) from the training data in this evaluation. The corpora are tagged with senses from WN17pre — the ontology used for SV2LS-test data. We create three sets of SALAAM tagged examples based on the different corpora utilized:

- **SV2LS-TR**: English SENSEVAL2 Lexical Sample trial and training corpora
- **MT**: The English Brown Corpus, SENSEVAL1 trial, training and test corpora, Wall Street Journal corpus, and SENSEVAL 2 All Words corpus
- **HT**: UN English corpus

Nouns	UMSST-human%	Nouns	UMSST-human%
art	47.9	grip	58.8
authority	62	hearth	75
bar	60.9	holiday	86.7
bum	85	lady	72.5
chair	83.3	material	55.9
channel	62	mouth	55.9
child	58.7	nation	78.4
church	73.4	nature	45.7
circuit	62.7	post	57.6
day	62.5	restraint	60
detention	65.6	sense	39.6
dyke	89.3	spade	75
facility	54.4	stress	50
fatigue	80.5	yew	78.6
feeling	56.9		

Table 5.4: Current evaluation gold standard precision results obtained by UMSST-human

Table 5.5 shows the relative sizes of the different corpora SALAAM-tagged and used as training data.

Corpora	Lines	Tokens
SV2LS-TR	61879	1084064
MT	151762	37945517
HT	71672	1734001
<i>Total</i>	<i>285313</i>	<i>40763582</i>

Table 5.5: SALAAM-tagged training corpora sizes

5.6.5 SALAAM-tagged Training Data Creation

The goal is to render the SALAAM-tagged corpora in a similar format to that of the human-tagged training corpus provided by the organizers of SENSEVAL2 and also utilized by UMSST. These training corpora are sense tagged using SALAAM. The sense tagging for MT and SV2LS-TR is based on using SALAAM with machine translated parallel corpora. The HT corpora are tagged based on using SALAAM with the English Spanish parallel corpus where the alignments are fixed (see Chapter 4, experimental condition 4).

The SALAAM-tagged corpora are divided into instance contexts for the different nouns in the test set. A context ranges from 2-4 lines long while respecting document boundaries. For example, in the Wall Street Journal the context for a word in the last line of a document within the corpus includes the two lines before the line where the word of interest occurs; the context does not span the beginning of the following document. Instance contexts may overlap.

Once the corpora are in the appropriate format, features are extracted as described in Section 5.5 for training the SVM learning algorithm.

5.6.6 Experimental Conditions

We have several factors that are varied to create the different experimental conditions for this evaluation. The factors are based on the SALAAM tagging parameters.

- **Corpus**

We vary the corpus used for training. We have 4 different combinations for the training corpus: MT and SV2LS-TR (MT+SV2LS-TR); MT and SV2LS-TR and HT (MT+HT+SV2LS-TR); HT and SV2LS-TR (HT+SV2LS-TR); or SV2LS-TR alone (SV2LS-TR).

- **Language**

This is the context language of the parallel corpus used by SALAAM to obtain the sense tags. We have 3 options: French (FR), where the sense tags are obtained using SALAAM in condition GLSYS-FR_M or GLSYS-FR_T (see condition 3, Chapter 3); Spanish (SP), where the sense tags are obtained using SALAAM in condition GLSYS-SP_M or GLSYS-SP_T (see condition 3, Chapter 3); or, Merged languages (ML), where the results are obtained by merging the output of FR and SP.

- **Threshold**

This is the threshold to which SALAAM's sense selection criterion is set. We have two options: MAX (M); and THRESH (T) (see Section 3.7.7 in Chapter 3 for detailed description).

- **Merge Type**

For the ML cases, we merge the results of two languages in 4 different ways:

- *Intersection (I)*

The tagged noun instances that are common to the results of the sense tagging of the two merged languages are intersected. Unique noun instances to either language are kept in the tagging set.

- *Union (U)*

The tagged noun instances that are common to the results of the sense tagging of the two merged languages are union merged. Unique noun instances in both languages are kept in the tagging set.

- *Strict Intersection (SI)*

This is similar to the *I* case above in that it is the intersection of the results of sense tagging the two languages, but it is more restrictive in that only noun instances that are common to both languages are included in this set. i.e. Unique instances to either language are excluded from the tag set.

- *Strict Union (SU)*

This is similar to the *U* case above, the SALAAM tagging results of the two languages are union merged, but only noun instances that are common to both languages are included in this set.

These factors result in 48 conditions.⁵ We exclude 7 of these conditions due to the extreme sparsity of the contexts, i.e. nearly none of the nouns in the test set have contexts for training. We are left with 41 conditions.

5.6.7 Evaluation Metric

In this evaluation, we use the Performance Ratio (PR) between two precision scores on the same test data. Precision is measured as described in Chapter 3.

$$PerformanceRatio(PR) = \frac{(P)AutomaticTagging}{(P)HandTagging} \quad (5.2)$$

5.6.8 General Experimental Hypothesis

Training a supervised WSD system using large amounts of SALAAM-tagged data for training results in *good* PR when measured against the same supervised system using hand-tagged data.

For the purposes of this evaluation, we consider a PR of ≥ 0.65 an acceptable, *good*, performance ratio, as this is the average score of UMSST's performance in SEN-SEVAL2 exercise.

⁵We are not listing the conditions here for space considerations. An excel workbook instanceStats.xls that details the different conditions for the different nouns in the test set may be viewed at <http://www.umiacs.umd.edu/mdiab>.

5.7 Results

In this evaluation, we view the different conditions as independent automatic taggers, that may be activated simultaneously.⁶ Therefore, we present the results in two different ways. In Table 5.6, we illustrate the maximum results obtained by any of the conditions (taggers). In Table 5.7, we show the five best individual conditions.

Table 5.6 shows the max results achieved by any of the taggers. The first column in table shows the precision (P) obtained on the test data SV2LS-test when training UMSST using SALAAM-tagged data, UMSST-SALAAM. These are the best results obtained per noun.⁷ These results are compared against the second column in the table, which demonstrates results of UMSST using hand-tagged training data, UMSST-human. The third column is PR measured according to equation (5.2); PR is measured between UMSST-SALAAM and UMSST-human precision results. The rows are sorted in descending order based on PR, for convenience. The last row in the table illustrates the overall performance average. UMSST-SALAAM achieves 45.1% average precision on the test set SV2LS-test; UMSST-human yields 65.3% average precision across the different nouns on the same test set. This results in an overall PR of 0.69 for UMSST-SALAAM against UMSST-human.

As shown in Table 5.6, 9 nouns yield the highest ratio of 1.00. In fact, in the first two nouns *detention* and *chair*, UMSST-SALAAM scores better results than UMSST-human leading to ratios of 1.05 and 1.02, respectively. 3 nouns, *art*, *child* and *material*, yield ratios above 0.9, followed by 4 nouns yielding ratios above 0.65. If we are to include only nouns that achieve ratios of ≥ 0.65 — the first 16 nouns in Table 5.6 — the overall precision of UMSST-SALAAM is significantly increased to 63.8% and the overall precision of UMSST-human is increased to 68.4% leading to a PR of 0.93.

The following Table 5.7 shows the results as obtained by the best 5 individual conditions. The top 16 test nouns in Table 5.6 achieve the highest PR across these 5 conditions. The maximum number of high PRs — ≥ 0.65 — yielded by any individual condition is 12.

In Table 5.7, the ratios that exceed 0.65 are typed in italics. The last row in the table shows the average PR across the different conditions. The five conditions yield similar average PRs. We note the presence of extremes in these results, for instance, *mouth* has a PR of 0.73 for condition 1 but approximately zero PR across the board, which is a reflection on the lack of training data for that item in the respective condition. It is worth noting that they are not as high as the average ratio taken over the top 16 test nouns in Table 5.6 since some of the test nouns in these conditions do not yield the maximum possible precision which is reported in Table 5.6; for instance, none of these 5 conditions achieve the max precision for the noun *child* of 57.1% at a ratio of

⁶Even though there are conditions that seem to be very highly related, these conditions are independent.

⁷The results in these columns are from applying different conditions

Nouns	UMSST-SALAAM P.	UMSST-human P.	PR
detention	68.8	65.6	1.05
chair	84.8	83.3	1.02
bum	85	85	1.00
dyke	89.3	89.3	1.00
fatigue	80.5	80.5	1.00
hearth	75	75	1.00
spade	75	75	1.00
stress	50	50	1.00
yew	78.6	78.6	1.00
art	46.9	47.9	0.98
child	57.1	58.7	0.97
material	51.5	55.9	0.92
church	56.2	73.4	0.77
mouth	40.7	55.9	0.73
authority	43.5	62	0.70
post	37.9	57.6	0.66
nation	45.9	78.4	0.59
feeling	33.3	56.9	0.59
restraint	33.3	60	0.56
channel	32.4	62	0.52
facility	27.6	54.4	0.51
circuit	27.7	62.7	0.44
nature	19.6	45.7	0.43
bar	18	60.9	0.30
grip	15.7	58.8	0.27
sense	9.4	39.6	0.24
lady	11.8	72.5	0.16
day	4.9	62.5	0.08
holiday	6.7	86.7	0.08
Average	<i>45.1</i>	<i>65.3</i>	<i>0.69</i>

Table 5.6: Precision scores and PR of UMSST-SALAAM and UMSST-human on SV2LS-test

Nouns	1	2	3	4	5
detention	<i>1.05</i>	<i>1.05</i>	<i>1.05</i>	<i>1.05</i>	<i>1.05</i>
chair	<i>1.02</i>	<i>1.02</i>	<i>1.02</i>	<i>1.02</i>	<i>1.02</i>
bum	0.09	0.15	<i>0.94</i>	0.03	0.03
dyke	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>
fatigue	0.03	<i>1.00</i>	<i>1.00</i>	0.03	0.03
hearth	<i>1.00</i>	<i>1.00</i>	0.17	<i>1.00</i>	<i>1.00</i>
spade	<i>0.92</i>	<i>1.00</i>	<i>1.00</i>	<i>0.96</i>	<i>1.00</i>
stress	0.11	0.05	0.05	<i>1.00</i>	<i>1.00</i>
yew	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>
art	<i>0.98</i>	<i>0.98</i>	<i>0.98</i>	<i>0.98</i>	<i>0.98</i>
child	<i>0.89</i>	<i>0.87</i>	<i>0.89</i>	<i>0.87</i>	<i>0.89</i>
material	<i>0.87</i>	<i>0.82</i>	<i>0.58</i>	<i>0.89</i>	<i>0.84</i>
church	<i>0.75</i>	<i>0.75</i>	<i>0.75</i>	<i>0.72</i>	<i>0.75</i>
mouth	<i>0.73</i>	0.00	0.03	0.03	0.03
authority	0.60	0.60	0.58	<i>0.67</i>	<i>0.70</i>
post	0.63	<i>0.66</i>	<i>0.66</i>	0.47	0.58
Ave. PR	0.73	0.75	0.73	0.73	0.74

Table 5.7: PRs of the best individual conditions using UMSST-SALAAM training data on the top 16 test nouns in Table 5.6: 1 is condition MT+HT+SV2LS-TR_THRESH_SP; 2 is condition HT+SV2LS-TR_THRESH_ML_I; 3 is condition HT+SV2LS-TR_THRESH_ML_U; 4 is condition MT+HT+SV2LS-TR_MAX_ML_I; and 5 is condition HT+SV2LS-TR_MAX_ML_I

0.97, in all these conditions the PRs for this noun range from 0.87 to 0.89.

We note that these five conditions use the HT corpus and four of the five conditions are the result of merged languages in the tagging using SALAAM.

Comparing our approach to that of Mihalcea, she reports success for her bootstrapping approach — achieving ≥ 0.90 ratio — on 6 nouns out of the 29 nouns in the SV2LS-test set using clean data as presented in Table 5.1. In our study with noisy SALAAM training data, we achieve similar success rates with 12 nouns out of the 29 nouns in the test set, including 4 of the nouns used in her study. Moreover, 4 additional nouns yield ratios ≥ 0.65 .

Figure 5.3 illustrates the results, specifically PR values, obtained using SALAAM training data versus PR values obtained by Mihalcea’s system. The hashed bars in the graph are the PR values obtained by Mihalcea while the solid bars are those of SALAAM.

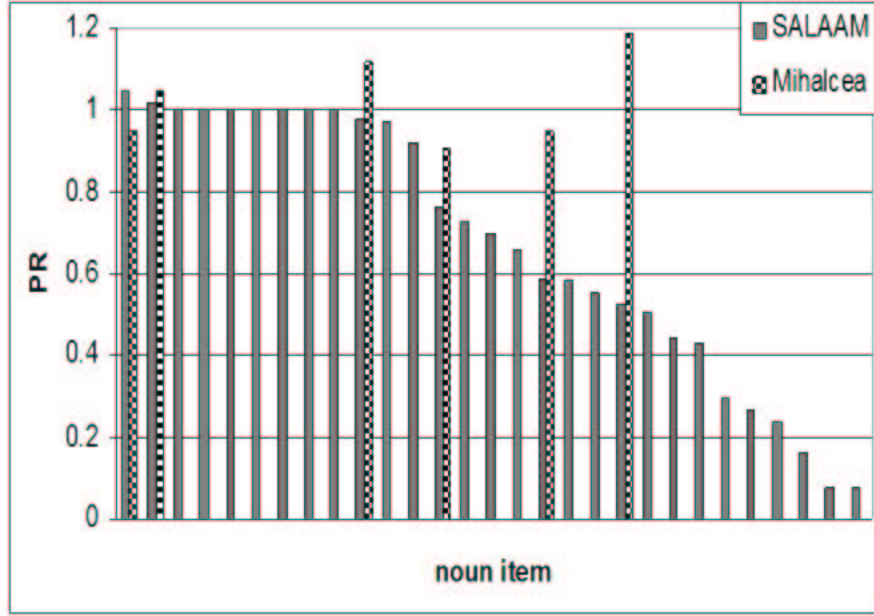


Figure 5.3: Comparison between Mihalcea's results and SALAAM results on the same test set

5.8 Discussion

The results obtained are very encouraging, especially when compared against clean data approaches such as the study by Mihalcea (see Section 5.3 above).

An interesting observation worthy of noting is the precision obtained on the SV2LS-TR corpus for the 29 nouns using SALAAM. SV2LS-TR is common to all the training conditions in this evaluation. We obtain an evaluation of the quality of the tagging in the training data by evaluating the SV2LS-TR tagging since it constitutes a significant sample of the UMSST-SALAAM training data set. The idea in carrying out this comparison is to get a feel for how noisy is the tagging for the different items under investigation.

Since we have the hand tagged annotations for the different contexts of the nouns for the SV2LS-TR corpus provided by the organizers of the SENSEVAL2 Lexical Sample task, we calculate the precision of the SALAAM tagged words using `scorer2` in the fine grain mode. Figure 5.4 plots the trend lines for the results presented in Table 5.8.

Figure 5.4 shows the trend lines of the performances of both the supervised UMSST-SALAAM system and the quality of the SALAAM unsupervised (SALAAM-SV2LS-TR) annotations on a portion of the training data used for the 29 nouns in this eval-

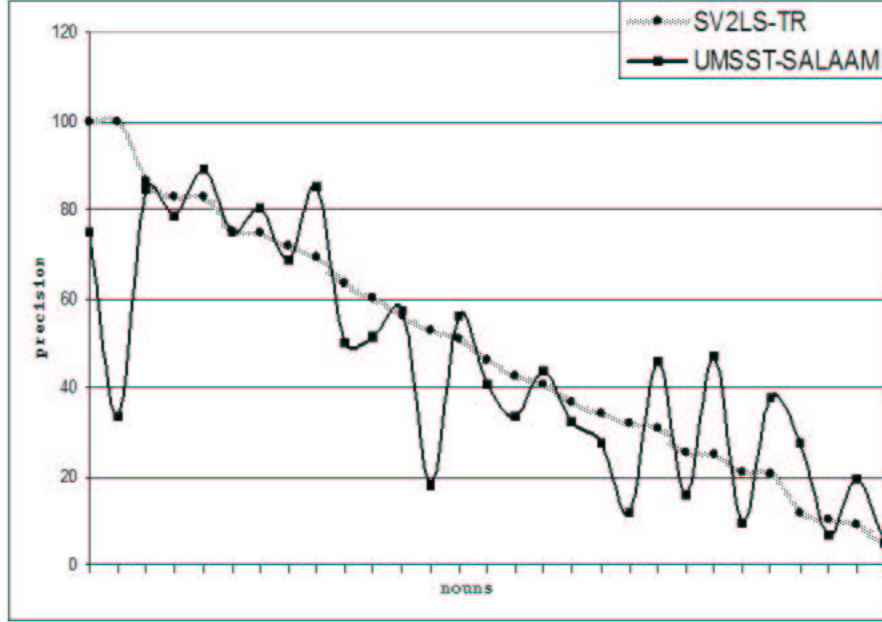


Figure 5.4: Trend lines for the precision obtained by SALAAM-SV2LS-TR and UMSST-SALAAM

uation. The X-axis is the nouns and the Y-axis is the precision score. The solid line is the trend line for the precision scores obtained by UMSST-SALAAM-tagged in the conditions that yield the max results illustrated in Table 5.6; the hashed line represents the precision of SALAAM-SV2LS-TR. As expected, the trend lines nearly overlap for several nouns such as *chair*, *channel*, *day*, *detention*. Yet surprisingly, we observe nouns where UMSST-SALAAM outperforms SALAAM-SV2LS-TR, i.e. the precision obtained by UMSST-SALAAM is higher than the precision of the sample it trains on. This is further illustrated in Table 5.8 below.

The rows in Table 5.8 are sorted in a descending order by the SALAAM-SV2LS-TR column. 11 out of 29 nouns have lower precision in SALAAM-SV2LS-TR training data sample than that obtained by the supervised system UMSST-SALAAM. The rest of the nouns illustrate higher precision in the sample training data.

We expect noisy training data to produce good quality results such as the case with *detention*, *spade*, *stress* and *yew* where the training data is noisy but still the precision in SALAAM-SV2LS-TR is higher than that of UMSST-SALAAM. Nouns such as *dyke* and *fatigue* exhibit intriguing behavior as they both yield PRs of 1.0 according to Table 5.6, but SALAAM-SV2LS-TR's precision is less than that of UMSST-SALAAM; for *dyke*, SALAAM-SV2LS-TR has a precision of 82.9% while UMSST-SALAAM achieves a precision of 89.3%. All the more surprising are the scores for

Noun	SALAAM-SV2LS-TR P	UMSST-SALAAM P
hearth	100	75
restraint	100	33.3
chair	86.7	84.8
yew	83	78.6
dyke	82.9	89.3
spade	75.4	75
fatigue	75	80.5
detention	71.9	68.8
bum	69.5	85
stress	63.6	50
material	60.3	51.5
child	56.2	57.1
bar	52.8	18
church	50.9	56.2
mouth	46.2	40.7
feeling	42.7	33.3
authority	40.7	43.5
channel	36.7	32.4
circuit	34.1	27.7
lady	31.9	11.8
nation	30.9	45.9
grip	25.3	15.7
art	25	46.9
sense	20.9	9.4
post	20.6	37.9
facility	11.8	27.6
holiday	10.3	6.7
nature	9.1	19.6
day	4.9	4.9

Table 5.8: Precision % scores obtained for SALAAM-SV2LS-TR and UMSST-SALAAM

the word *art*; for UMSST-SALAAM, *art* achieves a PR value of 0.98 against UMSST-human, yet the SALAAM-SV2LS-TR precision is very low at 25%. We count, all in all, 11 nouns where this phenomenon occurs.

Such seemingly contradictory results allow us to entertain two hypotheses about the robust performance of the supervised WSD system. The first hypothesis lays the burden of the difference in performance on the rest of the tagged corpora used for training. The idea is that the augmented corpora, MT and HT, have better quality tagging for those noun items than SV2LS-TR, therefore leading to an increase in precision for the UMSST-SALAAM on the test set for the 11 nouns where the performance in UMSST-SALAAM exceeds that of SALAAM-SV2LS-TR; similarly, augmenting SV2LS-TR with HT and MT introduces noise to the tagging quality for nouns such as *restraint* — which has a SALAAM-SV2LS-TR precision of 100%, yet UMSST-SALAAM only achieves 33.3% — contributing to the reduced performance in UMSST-SALAAM.

The second hypothesis is in tune with an interesting observation noted by Yarowsky [85] about noise in a machine learning environment, which states that noise is usually tolerated. In such an environment, correct parameters pertaining to a certain class — sense — are obtained from all of its occurrences, while incorrect parameters are distributed among all the different classes, therefore they do not produce statistically significant patterns. The results obtained in this evaluation lend support to Yarowsky’s observation as they show the robustness of the learning algorithm and the discriminatory power of the utilized features.

One can easily visualize using other types of noisy data — for example a Lesk based approach [43] — as training data and obtaining similar performance.⁸ Nonetheless, using SALAAM annotated data is interesting for two main reasons:

- SALAAM is able to annotate large amounts of source and target language data, therefore, we can visualize a system of bootstrapping a supervised WSD system for Arabic using the tags projected from the English WordNet ontology (see Chapter 4).
- Owing to the exploitation of multilingual evidence, SALAAM is able to sense tag words that are not typically accessible to methods that use monolingual contexts alone. This is illustrated by the complementary results presented and discussed in Chapter 3, Section 3.9.

5.8.1 Analysis of factors affecting PR

The results obtained are very encouraging, indeed, as an initial investigation into bootstrapping a supervised learning WSD system using noisy data. But we would like to

⁸It is worth stressing the inability to use first sense approaches to yield good results in classical machine learning approaches to WSD. Machine learning approaches need the variation in environments — negative and positive — of training examples in order to make predictions.

take this approach one step further. Our goal is to automatically predict, given noisy training data, which data items are *taggable* by this approach and which are not. We quantify *taggable* as possessing the potential to yield acceptable PR values, where acceptable is set at ≥ 0.65 . The approach aims to alleviate the tagging acquisition bottleneck. Therefore, if we have a system that is capable of predicting ahead of time which candidate items need to be hand annotated and which could be annotated automatically in the manner described in this chapter. The following analysis closely examines the different factors that affect the bootstrapping effort. This exploration aims at garnering a better understanding of the factors in order to be able to utilize them in a future automated system.

We classify the factors into two types: First, characteristics of the training data alone, such as number of training examples, number of senses per noun, perplexity of the senses, and semantic translation entropy; second, factors that are attributes of the relation between the test data, SV2LS-test, and SALAAM training data, such as the correlation between their respective perplexity measures, and correlation between their sense distributions across the different nouns. Finally, we address an attribute of the senses of the nouns in question and attempt to devise a measure of their context confusability based on sense similarity within a word. We discuss these factors by focusing on the best results obtained from all the automatic taggers (experimental conditions) as illustrated in Table 5.6. In the next portion of this section, we explore the effect of these different factors with respect to PR as reported in Table 5.6 against the best PR values obtained by UMSST-SALAAM.

1. Number of senses & Number of training examples in SALAAM training data

It is well known within supervised learning paradigms that there is a close relationship between the number of examples given and the performance of the learning algorithm. Table 5.9 shows the number of training contexts and the PR yielded. Also listed in the table is the number of senses for each word.

In Table 5.9, the average number of senses is 7.9 and the median is 7 senses. The cases where there are many senses such *art* with 17 senses, *material* with 16 senses, *mouth* with 10 senses and *post* with 12 senses, exhibit good performance at PRs of 0.98, 0.92, 0.73 and 0.66, respectively. The linear Pearson R correlation coefficient between number of senses and PR of UMSST-SALAAM is -0.31 which is not significant ($F(1, 27) = 2.9, p > 0.1$); it is a weak negative correlation indicating that when the number of senses increases PR — weakly — tends to decrease.

The correlation coefficient between the number of contexts and PR is even weaker at -0.15 which is not significant ($F(1, 27) = 0.637, p > 0.4$) indicating the lack of any significant linear correlation between the number of training contexts and performance. More interestingly, we observe cases where there are only 5 training examples yet PR is 1.0 as the noun *hearth* shown in Table 5.9.

Noun	# Senses	# Contexts	PR
art	17	111	0.98
authority	9	656	0.70
bar	19	175	0.30
bum	4	122	1.00
chair	7	432	1.02
channel	7	213	0.52
child	7	694	0.97
church	6	342	0.77
circuit	13	213	0.44
day	16	1247	0.08
detention	4	148	1.05
dyke	2	37	1.00
facility	5	358	0.51
fatigue	6	95	1.00
feeling	5	217	0.59
grip	6	195	0.27
hearth	3	5	1.00
holiday	6	177	0.08
lady	8	404	0.16
material	16	681	0.92
mouth	10	55	0.73
nation	4	4350	0.59
nature	7	457	0.43
post	12	365	0.66
restraint	8	11	0.56
sense	8	300	0.24
spade	6	72	1.00
stress	6	304	1.00
yew	3	64	1.00

Table 5.9: List of test nouns with their corresponding number of senses and sense contexts in the SALAAM-tagged training data

The lack of any significant correlation between the number of training instances and the PR contradicts the hard held belief in machine learning circles, that there is a direct correlation between supervised systems performance and the number of examples seen by the learner.

2. Perplexity of senses in SALAAM training data

Perplexity is measured according to equation (5.1). As explained earlier, there is a direct relation between perplexity and entropy. Entropy is a measure of confusability in the senses' contexts distributions, if it is uniform, entropy is high. A skew in the sense contexts distributions indicates low entropy, therefore low perplexity. The lowest possible perplexity value is 1 indicating an entropy of 0. Perplexity, accordingly, is the number of senses that are confusable due to the level of uncertainty in the sense contexts distributions. This characteristic is directly measurable on the SALAAM-tagged training data. For example, *bar* has the highest perplexity value of 9.85 for its 19 senses; and *day* with 16 senses has a relatively much lower perplexity of 1.3. Figure 5.5 illustrates the trend lines for the different sense contexts distributions of *bar* and *day*. The solid line depicts the distribution for *bar* which is almost a straight line indicating the close to uniform distribution, thereby reflecting an expected high perplexity. The hashed line depicts the sense contexts distribution for *day*; as we can see the spike in the graph reflecting the skew in *day*'s senses contexts distribution, hence the low perplexity.

Table 5.10 illustrates the perplexity per noun in the SALAAM-tagged training data and PRs obtained by UMSST-SALAAM. The rows are sorted in an ascending order by the Perplexity values. On the one hand, we observe nouns with high perplexity such as *bum*, with a perplexity value of 3.03, yet, achieving a high PR value of 1.0. On the other hand, nouns with relatively low perplexity values such as *grip* yield a very low PR of 0.26. Moreover, nouns with the same perplexity and similar number of senses yield very different PR scores. For example, *bum* and *feeling*, both have a perplexity value of 3.031, *bum* has 4 senses and *feeling* has 5, but the former yields a PR of 1.0 while the latter achieves a PR of 0.59 only. Furthermore, *nature* and *art* have the same perplexity of 2.297; *art* has 17 senses while *nature* has 7 senses only, however, *art* yields a PR of 0.98 while *nature* yields a PR of 0.44 only. Similarly, examining *holiday* and *child*, both nouns have the same perplexity of 2.144 and the number of senses is close with 6 senses for *holiday* and 7 senses for *child*, yet the performance is very different; the performance ratio for *holiday* is 0.08, while that of *child* is 0.97.

Consequently, the data is inconclusive. It does not support the existence of a correlation between the perplexity measure and the performance ratio, PR. These observations are further solidified by the low negative linear Pearson correlation coefficient of -0.12 , which is not significant ($F(1, 27) = 0.45, p > 0.5$)

Noun	# Senses	Perplexity	PR
dyke	2	1.000	1.00
facility	5	1.050	0.51
grip	6	1.058	0.27
church	6	1.149	0.77
channel	7	1.189	0.52
nation	4	1.301	0.59
day	16	1.347	0.08
fatigue	6	1.357	1.00
chair	7	1.385	1.02
yew	3	1.580	1.00
holiday	6	1.682	0.08
child	7	1.693	0.97
sense	8	1.853	0.24
detention	4	1.932	1.05
circuit	13	1.959	0.44
spade	6	2.144	1.00
stress	6	2.144	1.00
lady	8	2.144	0.16
nature	7	2.297	0.43
art	17	2.297	0.98
authority	9	2.462	0.70
hearth	3	2.639	1.00
post	12	2.639	0.66
mouth	10	2.828	0.73
bum	4	3.031	1.00
feeling	5	3.031	0.59
restraint	8	3.732	0.56
material	16	3.732	0.92
bar	19	9.849	0.30

Table 5.10: Test nouns, the corresponding number of senses, perplexity values and PR

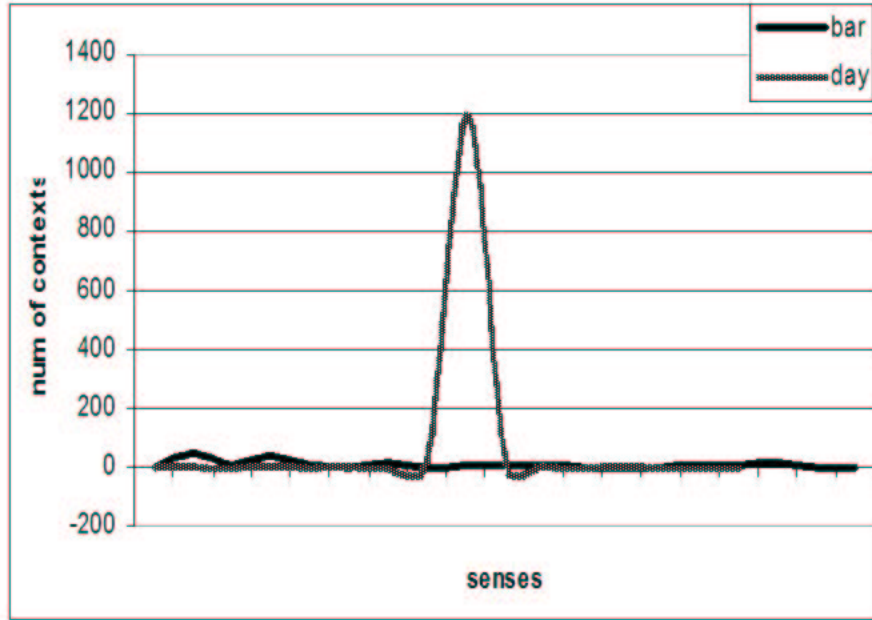


Figure 5.5: A plot of the distribution of the senses' contexts of *bar* and *day*

between perplexity and PR. At first blush, one is inclined to hypothesize that, maybe, the combination of low perplexity associated with a large number of senses — since it is an indication of high skew in the distribution — is a good indicator of high PR, but reviewing the data, this hypothesis is dispelled by *day* which has 16 senses with a perplexity of 1.3 yet still yields a very low PR of 0.08.

In summary, the data does not suggest that there exists any correlation between PR and perplexity. As it stands, perplexity alone is not a good indicator of performance.

3. Semantic Translation Entropy in SALAAM training data

Semantic translation entropy is a special characteristic of the SALAAM-tagged training data. Since the source of evidence for SALAAM tagging is multilingual translations, it is natural to evaluate the impact of translation on the tagging. Semantic translation entropy measures the amount of translational variation for the source word in the target language. Semantic Translation Entropy is introduced by Melamed [54]. This measure is a variant on the entropy measure described in Section 5.6.1. The equation utilized is expressed as follows:

$$H(T|s) = - \sum_{t \in T} p(t|s) \cdot \log_2(p(t|s)) \quad (5.3)$$

where t is a translation in the set of possible translations T ; and s is a source word.

The probability of a translation t is calculated directly from the alignments of the source nouns with their target language translations. It is the probability calculated via the maximum likelihood estimate of the translation for the word in question.

As mentioned in Chapter 3, variation in translation is a desirable feature for SALAAM tagging. Therefore, we would expect there to be a positive correlation between the quality exemplified by precision of tagging SV2LS-TR and semantic entropy, since the variation in translation indicates that the source word has several possible translations in the target language. Indeed, we do obtain a positive correlation of 0.33 between the precision of the unsupervised tagging of SALAAM-SV2LS-TR and semantic entropy. The correlation coefficient value is expected to be higher if we have good quality translations and good quality alignments.

Table 5.11 shows the obtained semantic entropy values per noun.

The row entries in Table 5.11 are sorted in a descending order based in the Semantic Translation Entropy column. Based on the values presented in the table, there exists no clear correlation between Semantic Translation Entropy and Performance Ratio, PR. The linear correlation coefficient is 0.22, which is not significant ($F(1, 27) = 1.31, p > 0.26$).

Several nouns in the table that have a high semantic entropy value exhibit a high PR. This is the case for *bum*, *detention*, *dyke*, *stress*, and *yew*. There are data points that exhibit very low Semantic Translation Entropy and PR such as *child* and *holiday*. Examining the latter two nouns individually, we observe that *child* has a semantic translational entropy of 0.08 and it yields a very high performance ratio of 0.97. The low semantic entropy indicates lack of translational variation for this word even though it has 7 senses. Based on condition MT_THRESH_ML_U, which rendered this result for *child*, we see that *child* is translated to $\{\textit{enfant}, \textit{enfantile}, \textit{niño}, \textit{niño-pequeño}\}$ which preserve the ambiguity in both French and Spanish. Moreover, from Table 5.8, SALAAM-SV2LS-TR precision for *child* is only 56.1%, but the perplexity is low at 1.69, probably contributing to the good performance ratio. Examining *holiday*, on the other hand, we notice that it has a relatively high Semantic Translation Entropy value of 0.66, yet it yields the lowest PR of 0.08. Furthermore, upon inspecting the precision results obtained by SALAAM-SV2LS-TR from Table 5.8, *holiday* has

Noun	Semantic Translation Entropy	PR
bum	0.799	1.00
detention	0.758	1.05
holiday	0.664	0.08
restraint	0.623	0.56
post	0.615	0.66
stress	0.542	1.00
spade	0.492	1.00
yew	0.491	1.00
dyke	0.474	1.00
hearth	0.462	1.00
circuit	0.42	0.44
nature	0.418	0.43
fatigue	0.385	1.00
authority	0.381	0.70
grip	0.364	0.27
channel	0.361	0.52
sense	0.332	0.24
chair	0.327	1.02
feeling	0.317	0.59
mouth	0.299	0.73
material	0.297	0.92
bar	0.251	0.30
lady	0.249	0.16
church	0.246	0.77
art	0.236	0.98
day	0.236	0.08
facility	0.202	0.51
nation	0.187	0.59
child	0.0856	0.97

Table 5.11: Test nouns with their corresponding Semantic Translation Entropy values and performance ratios, PR

a precision of 10.3% which is partially explainable by the quality of the alignments which are very noisy. A sample of the alignments is listed as follows: *{fiesta, vacances, fête, día-fiesta, preserve, nouveau-engagé, holiday, congè, los, las, les, assistance}*.

Therefore, it seems that Semantic Translation Entropy alone is not a good indicator of PR.

4. Sense Distributional Correlation between test data and SALAAM-tagged training data

Sense Distributional Correlation (SDC) is an attribute that results from comparing the test data sense context distributions with SALAAM-tagged data sense context distributions. We noted earlier, in Section 5.6.3, that the correlation between the human tagged sense distributions and those of the test data is very strong, with correlations ranging from 0.9 to 1 correlation coefficients across the different nouns in this evaluation.

In this subsection, we compare the correlation coefficients of the SALAAM-tagged training data and the test data, SDC, against the performance ratio, PR.

Table 5.12 presents those results. Row entries in the table are sorted descending by the SDC column. Observing the data in the table, we notice a strong correlation between SDC and the performance ratio, PR. This is further confirmed by Pearson's correlation coefficient of 0.87 which is significant ($F(1, 27) = 80, p < 0.0001$).

Figure 5.6 further illustrates the strong correlation. The hashed line in the figure is a plot of the SDC values against the solid line depicting the performance ratios.

Upon close inspection of nouns in the table, we observe that the nouns that have a high performance ratio have high SDC values. Yet, it is not always the case that high SDC values predict high performance ratios. For example, *circuit* and *post* have relatively very high SDC values, 0.794 and 0.859, respectively; but they score lower performance ratios than *detention* which has a comparatively lower SDC value of 0.776. Examining these 3 data items in previous tables, we notice that both *circuit* and *post* have many senses, 13 and 12, respectively, while *detention* has 4 senses only. *detention* has a higher semantic translation entropy as illustrated in Table 5.11; moreover, it has a lower perplexity as shown in Table 5.10.

Therefore, we conclude that SDC is a very good indicator of PR but it still lacks vital information if used alone in order to make the correct prediction consistently.

5. Absolute Difference between Perplexity of senses of test data and SALAAM tagged training data: Perp Diff

Noun	SDC	PR
dyke	1.000	1.00
bum	0.999	1.00
chair	0.998	1.02
fatigue	0.995	1.00
hearth	0.990	1.00
yew	0.989	1.00
spade	0.975	1.00
mouth	0.964	0.73
nation	0.962	0.59
material	0.887	0.92
authority	0.860	0.70
post	0.859	0.66
art	0.830	0.98
child	0.800	0.97
church	0.795	0.77
circuit	0.794	0.44
detention	0.776	1.05
stress	0.770	1.00
channel	0.647	0.52
restraint	0.532	0.56
feeling	0.329	0.59
lady	0.215	0.16
bar	0.200	0.30
sense	0.169	0.24
facility	0.156	0.51
nature	0.013	0.43
grip	-0.020	0.27
day	-0.038	0.08
holiday	-0.087	0.08

Table 5.12: Test nouns with their corresponding SDC values and PRs

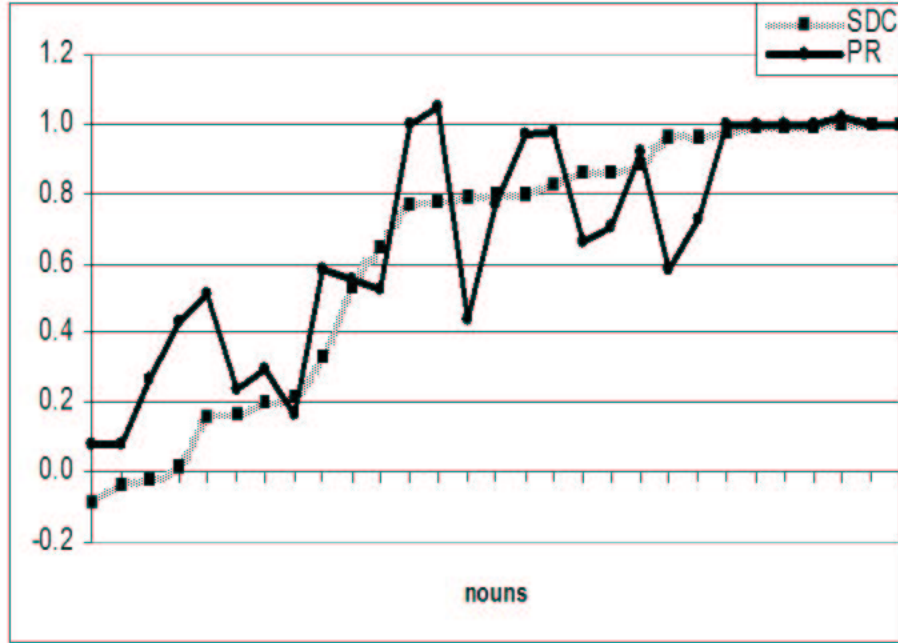


Figure 5.6: A plot of the SDC and performance ratio on the 29 nouns

There exists a very high Pearson's correlation coefficient of 0.96 between the perplexity measure of the test data and the human-tagged training data. In contrast to a relatively low correlation coefficient of 0.43 between the SALAAM-tagged training data and that of the test data. The two perplexity measures are illustrated in the following graph.

In Figure 5.7, the hashed line is the SALAAM-tagged perplexity and the solid line is the test data perplexity.

Table 5.13 illustrates the absolute difference between the SALAAM-tagged perplexity and the test data perplexity, PerpDiff, and the performance ratios, PR, per noun.

In Table 5.13, row entries are sorted by PerpDiff values descending. Examining the data, the correlation between PerpDiff and PR is at -0.4 . PerpDiff alone is not a good predictor of PR. We observe cases such as *holiday* with a very low difference indicating that the two perplexity measures are close for that data item, yet the PR is also very low, 0.08. While *circuit* has a perp diff of 7.23 but it achieves a relatively higher PR of 0.44. On the other hand, items such as *art* and *bum* have a relatively high PerpDiff but achieve very high PR scores.

6. Sense Context Confusability

Noun	PerpDiff	PR
holiday	0.05	0.08
spade	0.16	1
feeling	0.20	0.59
yew	0.27	1
chair	0.28	1.02
nation	0.34	0.59
dyke	0.40	1
child	0.45	0.97
hearth	0.50	1
fatigue	0.52	1
lady	0.69	0.16
detention	0.71	1.05
stress	1.11	1
mouth	1.17	0.73
material	1.19	0.92
restraint	1.19	0.56
bum	1.25	1
authority	1.27	0.7
church	1.31	0.77
facility	1.78	0.51
grip	2.19	0.27
nature	2.29	0.43
day	2.38	0.08
art	2.62	0.98
post	3.02	0.66
sense	3.07	0.24
bar	3.35	0.3
channel	3.73	0.52
circuit	7.23	0.44

Table 5.13: Test noun items with the absolute difference between SALAAM-tagged perplexity and test data perplexity, PerpDiff, against the performance ratios, PR

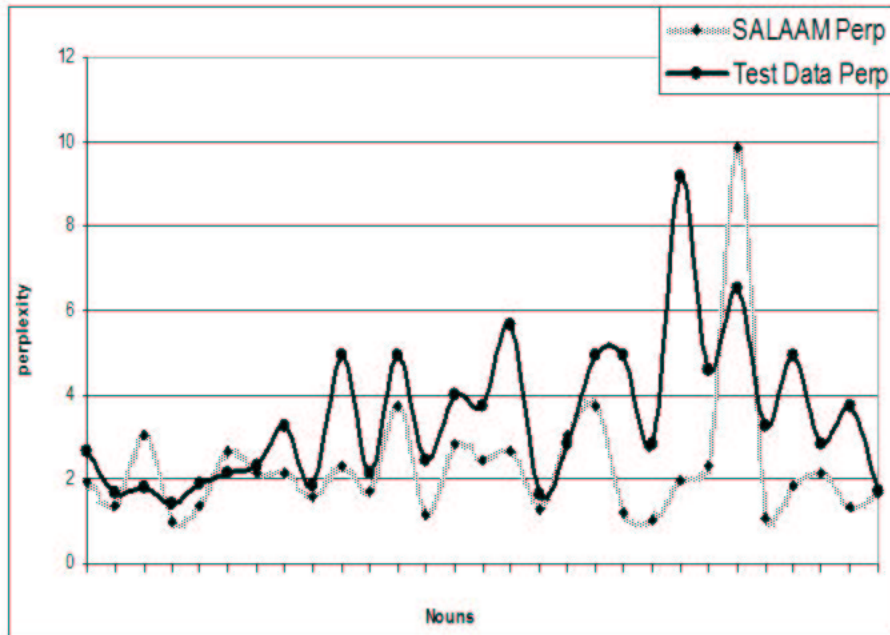


Figure 5.7: A comparative view of the different perplexity measures in SALAAM-tagged training data and the test data for the 29 nouns

This is a characteristic of the words in the training and test sets. Many of the senses of words in WordNet are similar as illustrated by the *sims* relationship in WN17pre [24, 13]. Similar senses typically lead to similar usages, therefore similar contexts. Such a similarity in contexts causes problems for the learning algorithm since the features extracted from the corresponding contexts for the polysemous word instances will tend to be very similar thereby, detracting from the learning algorithm’s discriminatory power. Such cases are referred to here as confusable sense contexts. A situation of sense context confusability arises when two senses are confusable and they are highly uniformly represented in the training corpus.

Upon examining the 29 polysemous nouns in the training and test sets, we realize that a significant number of the words have similar senses according to a manual grouping provided by Palmer, in 2002, as part of the SENSEVAL2 data distribution.⁹ For example, senses 2 and 3 of *nature*, meaning *trait* and *quality*, respectively, are considered similar by the manual grouping. Not all the senses pertaining to words in this test set are considered by the manual grouping,

⁹<http://www.senseval.org/sense-groups>. This manual sense grouping comprises 400 polysemous nouns including the 29 nouns in this evaluation.

since the test set has subsenses included in the tag set. For instance, the manual grouping only considers senses 1, 2 and 3 from WN17pre for *spade* — which are indeed homonymic —, while, in the current test set, *spade* has 6 senses. We find senses 2, 4 and 5 in this test set to be similar according to the following definitions as extracted from WordNet:

- Sense 2: hand shovel
- Sense 4: garden spade
- Sense 5: ditch spade

Table 5.14 illustrates the nouns sorted by PR values descending. The second column in the table presents the SDC values (see Sense Distribution Correlation above for description). The third column shows the manual grouping of similar senses; some cells in the third column are left blank indicating that the senses for the particular corresponding noun do not comprise groups i.e., the senses are not similar. The last column shows the performance ratio, PR.

Inspecting the nouns in Table 5.14, we see the majority of the nouns have multiple groupings. As established earlier, SDC has a significant impact on performance ratio. In this section, we argue that SDC is not a good predictor of performance ratio without taking Sense Context Confusability into consideration.

First, we consider the performance ratio of words that do not have sense groupings, the blank cells in Table 5.14, *detention*, *dyke*, *spade*, and *church*. They all achieve high performance ratios of 1.0 except for *church* which has a performance ratio of 0.77. *detention*, *dyke*, and *spade* all have SDCs above 0.97. Moreover, the four words have a close perplexity to the perplexity observed in the test data as illustrated in Table 5.13 with PerpDiff of 0.71, 0.4, 0.16 and 1.31, respectively.

However, upon inspecting *spade*, we note the existence of similar senses, as mentioned above, senses that are not considered in the manual grouping by Palmer since they are subsenses in WN17pre. All the automatic taggers (experimental conditions) in this investigation achieve an SDC above 0.95 for *spade*, and none of them have multiple instances in the similar senses, i.e. all the taggers have either 0 or 1 contexts for sense 4 and 5, with over 40 contexts for sense 2, indicating that there is always a skew in the distribution of the contexts for the potential similar sense contexts. Furthermore, the PR scores for all the taggers is ≥ 0.9 .

Accordingly, this observation leads us to believe that if the senses are not confusable and there is a sufficient number of contexts available, with a relatively close perplexity to that of the test data, then regardless of the SDC value, a tagger should be able to achieve high performance ratio; this is indeed the case with

Noun	SDC	Grouped Senses	PR
detention	0.776		1.05
chair	0.998	(1,4)(2,3)	1.02
dyke	1.000		1.00
bum	0.999	(1,2,3)	1.00
fatigue	0.995	(1,3)	1.00
hearth	0.990	(1,3)	1.00
yew	0.989	(1,2)	1.00
spade	0.975		1.00
stress	0.770	(1,4)(2,5)	1.00
art	0.830	(1,4)(2,3)	0.98
child	0.800	(1,3)(2,4)	0.97
material	0.887	(1,4)(2,3)	0.92
church	0.795		0.77
mouth	0.964	(1,2)(3,4,8)	0.73
authority	0.860	(1,6)(2,5)(3,7)	0.70
post	0.859	(1,2)(4,6)(5,7,8)	0.66
nation	0.962	(1,3)	0.59
feeling	0.329	(2,6)(4,5)	0.59
restraint	0.532	(1,3,4)	0.56
channel	0.647	(1,7)(2,4,6)	0.52
facility	0.156	(1,4,5)(2,3)	0.51
circuit	0.794	(2,3)(5,6)	0.44
nature	0.013	(1,4)(2,3)	0.43
bar	0.200	(1,2,13)(3,5,10)	0.30
grip	-0.020	(1,6)(2,3)	0.27
sense	0.169	(1,3)	0.24
lady	0.215	(1,2,3)	0.16
day	-0.038	(1,8,9)(3,5)(6,7)	0.08
holiday	-0.087	(1,2)	0.08

Table 5.14: Test nouns with SDC, manually grouped similar senses, and performance ratios, PR

detention, where some of the taggers have a SDC of 0.26, yet they still yield a performance ratio of 1.0.

As for the case of *church*, the perplexity of SALAAM-tagged training data is much lower than that of the test data as illustrated in Table 5.13, SALAAM-tagged perplexity is 1.15 while that of the test data is 2.83. This observation may lead us to conclude that perplexity is the determiner of the PR in this case. However, comparing *stress* to *church*, we note that both exhibit similar behavior with regards to the different factors examined here as illustrated in Table 5.15, yet *stress* achieves a PR of 1.0.

Factor	stress	church
SDC	0.770	0.795
Semantic Translation Entropy	0.54	0.246
PerpDiff	1.11	1.31
Number of Senses	6	6
Number of Contexts	302	342

Table 5.15: Characteristics of the nouns *stress* and *church*

As we can see in the Table 5.15, the characteristics for *stress* and *church* are very similar with the major difference being the Semantic Translation Entropy and PerpDiff. We inspect the contexts for *stress* for confusability and we discover no confusability in the contexts exists. Therefore, we may conclude that indeed the predicting factors are a combination of PerpDiff and Semantic Translation Entropy.

The nouns that are intriguing in Table 5.14 are the ones that have relatively high SDC values yet their performance ratios are low such as *post*, *nation*, *channel* and *circuit*. For instance, *nation* has a very high SDC of 0.962, a low perplexity of 1.3, relatively close to the 1.6 perplexity of the test data, a sufficient number of contexts (4350), yet its performance ratio is at a relatively low 0.59. According to the manual sense grouping listed in the table, senses 1 and 3 are similar, and indeed when we inspect the context distributions, we find the bulk of the senses' instances from senses 1 and 3 which create confusable contexts for the learning algorithm.

We have established the importance of taking Sense Context Confusability into consideration when attempting to predict the performance of a tagger. But we are faced with the problem of quantifying this factor. Fortunately, we are able to use Resnik's information theoretic similarity measure to quantifiably measure the similarity between the senses of polysemous words (see Chapter 3 for details) [68].

We conducted a preliminary experiment to investigate to what extent does the automatic similarity measure concur with the manual similarity grouping of senses. We evaluated 7 words, ranging from 2-7 senses per word: *yew*, *dyke*, *spade*, *church*, *holiday*, *nature*, and *nation*.¹⁰ The automatic measure assigned high similarity scores to 7 out of the 8 groups deemed similar by manual groupings including our additional manual grouping of senses 2,4, and 5 for *spade*, as well as senses 4 and 5 for *church*. In fact, the one case that was not considered similar by the automatic measure was a case of metonymy for *yew* where sense 1 means *tree* and sense 2 *wood*, which is a debatable case. But crucially the automatic measure did not group any senses that were not manually deemed similar. Therefore, one can easily use an information theoretic based similarity metric to quantify sense similarity. Using such a measure in conjunction with inspections of the uniformity of the distributions among the similar senses, allows for the quantification of the Sense Context Confusability.

5.9 Combining factors

We have analyzed the different potential factors that could have an impact on the PR score. Our ultimate goal is to be able to automatically predict which words are good candidates for bootstrapping. Implementing a learning model to automatically predict good performance ratios is a matter of future work but we will consider several relevant features based on our discussion above.

A first step would be to assign weights to the different factors affecting the performance ratio, PR, by training a learning model on the relevant ones. Such a learning framework will have to be implemented for each word independently. The learning model deduces the significant weight for each predictor.

The overarching question is which predictors are relevant. From our detailed exploration, the data suggests that SDC, Sense Context Confusability have a direct impact on PR, yet, Semantic Translation Entropy and PerpDiff play an indirect role in the prediction. The impact of the number of examples seems more relevant in a clean training/testing environment; we observe cases that only have 5 training examples, yet achieved a PR of 1.0; moreover, there is the notion of sufficient number of examples, we could not deduce from the data what that magic number is except to note that it has to be more than 5 given an SVM learning paradigm for this specific application type. Therefore, we would want to combine the number of contexts as a factor. The number of senses does not seem directly relevant based on our analysis.

Fortunately, there are several learning candidates robust enough for such an investigation that range from a simple regression model to algorithms such as Decision

¹⁰We use Resnik’s measure as a first step, but for this task the Lin [51] information theoretic measure is also appropriate. The difference between the two measures lies in normalization, the Lin similarity measure produces a similarity score between 0 and 1; while that of Resnik is not normalized.

Trees, Instance Based Learning, and Naïve Bayes frameworks.

The predictors in such a framework will be a combination of nominal and numeric values. The nominal predictors for this evaluation are: language, for example, we have French, Spanish and Merged languages; corpus type, for instance, HT+SV2LS-TR, MT+SV2LS-TR, SV2LS-TR, or SV2LS-TR; and sense selection criterion, MAX or THRESH. The numeric predictors are SDC, PerpDiff, Number of Contexts, Semantic Translation Entropy, and Sense Context Confusability. The learning framework will be given the predictors based on all the taggers and the value the learner is trying to predict is an acceptable performance ratio. This may be cast in a binary framework by setting a threshold on the acceptability value.

It is worth emphasizing that two of the identified factors are dependent on the test data, SDC and perp diff. Given the fact that the test data size is small relative to the hand tagged training data size required by a classical supervised system for WSD, SALAAM-tagged training is still a viable solution to the annotation acquisition bottleneck.

5.10 Summary

In this chapter, we have introduced a new approach that combines an unsupervised and supervised learning method for WSD that makes significant strides toward easing the annotation bottleneck. This is accomplished by means of a trade-off between quality and quantity. SALAAM produces large amounts of noisy data for training. We demonstrate the value of this approach using a precision ratio metric, PR, comparing a supervised WSD system trained on hand tagged data against that same system trained on SALAAM tagged data. The bootstrapping approach evaluated yields superior results to those obtained by the only comparable approach which is tested on the same data set but bootstrapped using clean data. Essentially, the method introduced here is entirely unsupervised yet it is able to rival results obtained by a supervised method for 12 out of 29 noun items.

Moreover, we explore, in depth, the question of when it is safe to use SALAAM tagged data as training data, since the approach works less optimally for several data items. We render a detailed analysis of the different factors affecting the performance ratio, PR. Finally, we make suggestions on how to combine the relevant factors toward the goal of automatically predicting good performance ratio.

Chapter 6

Facets of Similarity

6.1 Introduction

The notion of similarity is endemic to most scientific endeavors. The research agenda boils down to seeking generalizations about the world; in most cases generalizations are made based on explicit or implicit groupings of phenomena or ideas. Word Sense Disambiguation (WSD) is not different from any other scientific pursuit in that respect. Similarity plays a vital role in this field; majority of WSD systems have within them formulated some variant on a similarity measure, a way for mapping observables to a set of predetermined or undetermined classes. All WSD systems, whether supervised or unsupervised, have an embedded similarity (generalization) function that maps a set of features from some defined *Context* onto a set of classes.

In this chapter, we examine ways of automatically modeling semantic similarity; we are interested, in particular, in how they compare to human similarity judgments. There is a close relationship between understanding how linguistic representations are used and acquired and the manner in which semantic similarity is modeled. Models differ in their assumptions about features. Some models of similarity, such as Tversky's (1977), assume an explicit set of features over which a similarity measure can be calculated; some distributional methods for measuring word similarity may be viewed as an empirical implementation of such a model [9, 75]; in these methods, distributional features of words are acquired from the analysis of large corpora. Other semantic models define the features implicitly focusing more on relations among lexical items in a semantic network type framework, à la Quillian (1968); methods that utilize such models compute similarities among words represented in a taxonomy exploiting its hierarchical structure. In several of these methods, the measure takes into account some corpus-based features such as frequency information (e.g., Rada, Mili, Bicknell, & Blettner, 1989; Lin, 1999; Resnik, 1999).

There are a myriad of semantic similarity measures, which tend to look at different sources of evidence. Resnik [73] proposes using human similarity judgments as a reference point for comparing the different measures. Indeed, humans think in multi-dimensional space, there is plenty of evidence supporting the hypothesis that people

tap into different knowledge resources to make judgments about the world [42]. In Resnik’s 1999 study, he focuses on nouns. The key question asked is how do different automatic similarity measures compare with one another against human judgments. In this chapter, we present a similar type of investigation but the focus is on verbs. We design an experiment for human similarity using verbs and we compare the results obtained from different automated similarity measures against the obtained human judgments.

This chapter is laid out as follows: in Section 6.2, we draw attention to the different intrinsic facets of verbs and motivate the experiment; Section 6.3 examines the different automated similarity models that are evaluated in this investigation; Section 6.4 describes the experiment, discusses the results obtained from both the human ratings independently and then in relation with the automatic measures; Section 6.5 describes a framework for incorporating the different automated measures as an approximation to human judgments in SALAAM; finally, the chapter concludes with a summary of the findings in Section 6.6.

6.2 Motivation

Upon inspecting the performance of the different state-of-the-art WSD systems on the set of English verbs in the SENSEVAL 2 All Words task, we note a severe drop in the results obtained when compared to the results yielded for the nouns by those same systems. The average precision scored for the verbs is 30.8 % with the highest precision score at 55.4% and the lowest precision score at 6.5%, with a standard deviation of 14.2. Likewise, recall scores are significantly lower with a range of 0.2% to 49.4%. These results, if nothing else, are indicative of the complexity of the task. It is hardly surprising that results obtained from both supervised and unsupervised systems alike are low. This is attributed mainly to the sheer number of senses for verbs in WordNet; but also more importantly such results are a reflection of the fact that verb senses differ along more dimensions than simply paradigmatic ones. If a WSD system is not at least implicitly sensitive to these variations, it is subject to mis-tagging verb instances.

Verbs are different from nouns in many respects.¹ They vary paradigmatically, like nouns, however, in addition, verbs possess syntagmatic attributes. Typically, syntagmatic properties of verbs are characterized in terms of different dimensions. Being more relational in nature, verbs lay restrictions on the type and properties of the words that are associated with them. Some of these syntagmatic attributes are defined in terms of syntactic subcategorization properties and thematic restrictions, aspectual class attributes, and selectional preferences. These characteristics are not independent from each other in most cases.

¹We acknowledge that the distinction between verbs and nouns is not that discrete; we understand that a range does exist; but for purposes of this chapter we are not discussing the full spectrum.

Given the complex nature of verbs, we design an experiment to obtain human judgments on verb semantic similarity. The design and choice of the experiment items pay close attention to the different dimensions of meaning associated with verbs. Such an experiment is intended to serve as a point of reference of automatic semantic similarity measures; moreover, guided by insights from such a study, we lay a more cognitively salient framework for utilizing the different automated similarity measures within the area of WSD. In the process, we explore different semantic similarity measures and compare them with human ratings on verb similarity.

6.3 Models of Verb Similarity

Automated semantic similarity measures based on different similarity models typically take advantage of and are sensitive to one of or more of the many paradigmatic and syntagmatic attributes of verbs. Such methods may do so either implicitly or explicitly. For example, methods that depend on word adjacency collocations in text, implicitly take advantage of the word order of the verb and its arguments, yet depending on the prespecified window of text of interest, they may not be able to capture long distance syntactic relations. Methods that depend on WordNet and WordNet style Ontologies tend to be sensitive to the IS-A relationship within the taxonomy. The IS-A relation captures the *manner* dimension of meaning for the verbs' semantics. Such methods are known to yield good performance with nouns. Unlike the noun taxonomy however, the verb hierarchy is very shallow and broad. Similarity methods that measure similarity via syntactic relations may be sensitive to adjunction relations as well as locative and temporal modifiers when measuring verb similarity.

Accordingly, in this chapter we consider three classes of similarity measure, corresponding to three types of lexical representation ranging in level of syntactic depth from paradigmatic (syntactic-light) to being heavily dependent on syntax. In the first class, verbs are associated with nodes in a hierarchical ontology. At one end of the spectrum, the first class is shallowest in explicit syntactic representation. In the second, distributional syntactic cooccurrence features obtained by parsing a large corpus represent verbs. Finally, in the third class, verb entries are represented according to a theory of lexical conceptual structure. This class represents the other end of the syntactic spectrum with explicit coding of syntactic facets of meaning.

6.3.1 Class 1: Taxonomic Models

WordNet represents taxonomic models in this study. As mentioned earlier, this type of model focuses on the paradigmatic aspects of meaning for verbs. It is in clear contrast to efforts that classify verbs based on their syntagmatic behavior. For the purposes of this investigation, we use WordNet 1.5 (see Chapter 3 for a full description of a WordNet style Ontologies). We present three measures of verb similarity as follows:

- **Edge Count Similarity**

Given the hierarchical nature of the WordNet verb taxonomy, the simplest method of calculating similarity between two verbs is to count the number of intervening edges. The total number of edges is subtracted from the maximum possible number of edges in the taxonomy. Accordingly, *edge_sim* is calculated for two verbs v_1 and v_2 as follows:

$$wsim_edge(v_1, v_2) = (2 \times max) - [_{c_1, c_2} len(c_1, c_2)] \quad (6.1)$$

where c_1 ranges over all the senses (synsets) of verb v_1 , and c_2 ranges over all the synsets of verb v_2 ; max is the maximum depth in the taxonomy, and $len(c_1, c_2)$ is the length of the shortest path from c_1 to c_2 . Intuitively, if the synsets for the two verbs are not from the same portion in the taxonomy, the value of *edge_sim* is 0.

Edge counting is well known for its problems in over estimating and under estimating similarity between concepts in WordNet. This is mainly owing to the fact that subtrees vary in their bushiness, i.e. some trees are shallower than others.

- **Information Theoretic Similarity**

Information theoretic based similarity measures address the problem associated with the edge counting measure. Information based measures typically assign weights to nodes in the taxonomy. The weights are quantified in terms of information content. We describe two variants on information based similarity: *res_info*, devised by Resnik [73] and *lin_info* devised by Lin [51]. The inherent structure of the taxonomy arranges the nodes such that the more abstract concepts in the tree are higher than the more specific concepts. For instance, the concept of FRUIT is higher than the concept APPLE. Both measures exploit this feature of hierarchical taxonomies by assigning lower information content to the broader concepts than the amount of information content assigned specific concepts. Intuitively, this is a reasonable assumption since the amount of informativeness — information contribution — is lower in broader concepts than the amount of information in a more specific concept. Both measures calculate information content based on the unigram frequencies of the concepts in a corpus. The amount of information in a node is calculated as follows:

$$InformationContent = -\log_2(p(c)) \quad (6.2)$$

where c is a concept in the tree.

The more frequent a concept in a taxonomy the lower its information content. However, not all concepts in the tree occur in the corpus. Accordingly, the frequencies of more specific nodes are propagated upwards in the tree such that a parent node's frequency is the aggregate of the frequencies of all its children, in addition to its own frequency in the corpus if it happens to occur. Consequently, the broader concepts in the tree are automatically assigned lower information content.

Based on this characterization of information content, the similarity between two concepts in a taxonomy is measured according to *res_info* as

$$sim_{res_info}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log_2(p(c))] \quad (6.3)$$

where $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2 , and $-\log_2(p(c))$ is the information content of node c obtained in the manner described above.

In *res_info* measure, the most informative subsumer of the two concepts is the subsumer with the highest information content among all possible subsumers.

The *lin_info* measure between two concepts closely resembles the *res_info* measure but it normalizes the shared information content by using the sum of the unshared information content of the concepts compared. Therefore, *lin_info* similarity between two concepts c_1 and c_2 is calculated as follows:

$$sim_{lin_info}(c_1, c_2) = \frac{2 \times \log_2 p(\bigcap_i C_i)}{\log_2 p(c_1) + \log_2 p(c_2)} \quad (6.4)$$

where C_i are the superclasses that are maximally specific for concepts c_1 and c_2 . The range of this similarity is 0 to 1.

The most important feature of both measures is the definition of similarity as a function of the shared information content. Similarity measured in this manner is not sensitive to the number of edges intervening between two concepts in a tree, therefore, it is not prone to the problems associated with *edge_sim*.

The equations as mentioned above, characterize the measure of similarity for two concepts (synsets/senses) in the taxonomy not verbs. In order to obtain the similarity of two verbs v_1 and v_2 , the following calculation is utilized by both *res_info* and *lin_info*:

$$wsim_{info}(v_1, v_2) = \max_{c_1, c_2} [sim_{info}(c_1, c_2)] \quad (6.5)$$

where $sim_{info}(c_1, c_2)$ is either *res_info* or *lin_info*.

According to equation (6.5), the most informative subsumer is the subsumer with the highest information content across all sense pairings for the two verbs v_1 and v_2 .

6.3.2 Class 2: Distributional Co-occurrence Model

Lin [50] demonstrates the generality of information theoretic based similarity by showing how such measures can be used to measure not only taxonomic distance but also string similarity and the distance between feature sets. This approach is illustrated by representing words as collections of syntactic cooccurrence features obtained by parsing a large corpus. For example, both the noun *duty* and the noun *sanction* would have feature sets containing the feature *subj_of(include)*, but only *sanction* would have the feature *adj_mod(economic)*, since *economic sanctions* appears in the corpus but *economic duties* does not. Because these features include both labeled syntactic relationships and the lexical items filling argument roles, the underlying representational model can be thought of as capturing both syntactic and semantic components of verb meaning. Lin computes the quantity of shared information as the information in the intersection of the distributional feature sets for the two items being compared. This yields the following measure *lin_dist*:

$$lin_dist(w_1, w_2) = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))} \quad (6.6)$$

where $F(w_i)$ is the feature set associated with word w_i , and where $I(S)$, the quantity of information in a feature set S , is computed as $I(S) = - \sum_{f \in S} \log_2 p(f)$.

Lin assumes that the features are independent, therefore allowing for the summation of the log probabilities in equation (6.6). In the experiments described here, we use similarity values obtained for verb pairs using Lin's implementation of his model, with his feature sets and probabilities obtained by analyzing a 22 million word corpus of newswire text, The Sun Jose Mercury.

6.3.3 Class 3: Semantic Structure Model

The third automated method for assessing the semantic similarity of verbs relies on detailed representations of verb semantics according to the theory of lexical conceptual structure, or LCS [19, 35]. In the theory of LCS, a verb representation is defined in terms of its semantic structure and semantic content. The semantic structure represents the different dimensions of syntactic (syntagmatic) features for a verb, while semantic content is defined as the idiosyncratic (paradigmatic) information associated with it. In this model, a clear distinction is made between semantic content and semantic structure. This same distinction plays a central role in current studies of lexical representations [28, 64, 66].

We take advantage of the explicit distinction between semantic structure and semantic content to derive a measure that focuses exclusively on similarity of semantic structure independent from semantic content.

To illustrate LCS representations, we present the definition of the verbs *run* and *jog* which may have the same semantic structure, indicating a change of location depending on the senses of the verbs, e.g.,

$$\begin{aligned} & (go_{loc} \times \\ & \quad (to_{loc} \times (at_{loc} \times y)) \\ & \quad \quad (from_{loc} \times (at_{loc} \times z)) \\ & \quad \quad \quad (\text{manner } (M))), \end{aligned}$$

The verbs differ only in the value (M) — an element of semantic content within the semantic structure — indicating the manner of motion (either (*jogging*) or (*running*)). Such regularities in semantic structure are argued to provide an explanation for systematic relationships between meaning and syntactic realization [47, 44, 45, 46].

Given such LCS structures, we observe the patterning in semantic structure between verbs that are highly similar. We devise an algorithm inspired by Lin’s adaptation of his information theoretic measure to calculate similarity on parsed tuples.² In his approach he decomposes parse trees into (pseudo) independent features and uses his information theoretic based similarity measure, *lin_info* to calculate similarity. The algorithm we devise performs the same task: it converts each LCS entry into a series of tuples, conflating hierarchical information in the LCS structure. We recursively create an independent feature from each primitive component of the LCS representation and the **head** of its subordinates. So, for example, the feature set representation of *run* contains six features:

$$\begin{aligned} & (go_{loc} \ to_{loc} \ from_{loc} \ \text{manner}) \\ & \quad (to_{loc} \times at_{loc}) \\ & \quad \quad (at_{loc} \times y) \\ & \quad \quad (from_{loc} \times at_{loc}) \\ & \quad \quad \quad (at_{loc} \times z) \\ & \quad \quad \quad (\text{manner } \textit{RUNNING}). \end{aligned}$$

The created features of *jog* are identical except for the manner feature expressed as the last feature above, which instead would be (manner *JOGGING*). Therefore, we observe a complete overlap between the feature sets for the two verbs, which captures the fact that the semantic distinction between this particular pair of verbs rests entirely on semantic content, and not semantic structure.

²We are grateful to Dekang Lin for the idea (personal communication).

We have available to us a large lexicon of LCS representations for verbs in English [20], containing thousands of lexical entries. The probability of each feature is estimated by counting feature occurrences within the lexicon. We acknowledge that the probability estimate calculated in this manner counts features within a set of entries in a large lexicon (types) rather than verb instances in a large corpus (tokens), but inspection of the estimated probabilities suggests that frequent features are relatively discounted, having low information content, and rare features have high information content. Accordingly, *lcs_sim* is calculated for two entries verb v_1 and verb v_2 using the shared information content of their feature sets:

$$lcs_sim(v_1, v_2) = I(F(v_1) \cap F(v_2)) \quad (6.7)$$

where $I(F(v_1) \cap F(v_2))$ is measured as in equation (6.6).

Then the similarity *lcs_sim* between two verbs is calculated over all their entries corresponding to the different senses as the maximum yielded value *lcs_sim* taken over the cross product of all the verbs' lexical entries.

Therefore, this similarity measure considers only semantic structure, not semantic content; accordingly, only syntagmatically relevant features take part in the computation. When comparing, *run* and *jog*, in the specific example mentioned above, they only differ in their semantic content, their paradigmatic characteristic idiosyncrasies which are not captured by this model.

6.4 Human Judgment Experiment

We design a human experiment to collect ratings on semantic similarity between verbs. The intent of the experiment is to establish a reference point for comparing different automated similarity measures. The design follows that of Miller and Charles (1991), which is a design for noun similarity. However, when comparing verbs, considering their multidimensional nature and the complex intertwining of their attributes, choice of experimental material has to be controlled. Therefore, we pay close attention to syntactic subcategorization, thematic grids, and aspectual class information, as described below, in order to limit the possible dimensions across which the two verbs in a pair could differ and to focus on semantic similarity.

Moreover, we create two of versions of the experiment: one with the verbs presented to the participants with *No-Context*, and the second with the verbs presented in context. The idea is to examine to what extent *Context* has an effect on verb similarity ratings.

6.4.1 Participants

We have a cohort of 10 subject volunteers, 5 women and 5 men ranging in age from 24 to 53. They are all native speakers of English who participated by email. None of the participants has significant background in psychology or linguistics.

6.4.2 Materials

As mentioned earlier, in order to capture semantic similarity between verbs, we need to control for the different dimensional variations associated with verbs. Fortunately, we have available to us a large lexicon of English LCS structures comprising 4900 entries [20]. We control for three syntagmatic dimensions in the choice of the verb pairs:

- **Aspectual Class**

Each verb entry in the LCS lexicon contains information about its aspectual features: where the verb entry is dynamic, durative or telic [21]. Verbs are classified into four main aspectual classes: Activities, States, Achievements and Accomplishments. The combination of aspectual features predicts the aspectual class of the verb. Table 6.1 is taken from (Dorr& Olsen, 1997) [22], where a 1 indicated presence of a feature and 0 indicates its absence.

Aspectual Class	Telicity	Dynamicity	Durativity	Example Verbs
State	0	0	1	<i>know, have, be</i>
Activity	0	1	1	<i>march, paint, dance, chase</i>
Accomplishment	1	1	1	<i>destroy, eat, build</i>
Achievement	1	1	0	<i>notice, win, break</i>

Table 6.1: Aspectual features determining aspectual class for verbs

- **Thematic Grid**

The thematic grid information identifies whether or not a verb takes an **Agent**, **Theme**, **Goal**, etc.

- **Subcategorization Frames**

This identifies whether a verb takes an object or two objects, for example. For information on subcategorization, we used the subcategorization frame for the first listed verb sense provided in the Collins Cobuild Dictionary [76].

Accordingly, a verb such as *broil* requires both an **Agent** and a **Theme**, and is marked as both **Durative** and **Telic** but not **Dynamic** and has the subcategorization frame (v+o).

In order for us to build verb pairs, we first remove all verbs whose thematic grids do not require a theme, so as to limit the range of variation in thematic grids. All verbs require an agent, so the remaining variation lies in the presence or absence of oblique roles such as goal. The next phase is to group the full set of verbs into lists in correspondence with the eight possible combinations for the three aspectual features, then the lists are reduced to the four most numerous ones which are {**Durative**}, {**Durative, Dynamic**}, {**Dynamic, Telic**}, and {**Durative, Dynamic, Telic**}. Verbs may and do appear on multiple lists. Within each of those four lists, all possible pairings of verbs that matched in terms of subcategorization frames are created. 12 pairs are selected that range from low to high similarity pairings.

In summary, a set of 48 verb pairs is constructed such that:

1. both verbs in every pair require a theme,
2. both verbs have the same subcategorization frame, and
3. both verbs come from the same aspectual class.

Verbs on the list are all presented to the participants in the past tense. In order to avoid ordering effects, the order of the verb pairs is randomized; half the subjects in either condition of the two conditions (see Section 6.4.3) is shown items in a specific order, and the other half is shown the same items in a reverse order.

6.4.3 Conditions

We create two experimental conditions: *Context* and *No-Context*. The materials as just described are duplicated in order to create two distinct sets of conditions. The conditions are exactly the same with the exception that in the *Context* condition, each verb in the verb pairings presented to a participant is within an example sentence which demonstrates the verb's intended sense. The contextual sentences are taken from the corresponding verb entry in the Collins Cobuild Dictionary. For example, the sentence for *enrich* is *They enriched the library with new books*.

6.4.4 Procedure

Human Ratings Experiment

The 10 participants are divided evenly into two groups corresponding to the experimental conditions *Context* and *No-Context* groups. Subjects in the *No-Context* group are given the set of 48 verb pairs, without example sentences. They are asked to compare their meanings on a scale of 0-5, where 0 indicates that the verbs are not similar at all and 5 indicates maximum similarity.

Participants are explicitly asked to ignore similarities in the sound of the verb and similarities in the number and type of letters that make up the verb.

Participants are also asked distinctly to rate similarity rather than relatedness, with the instructions giving an example of the distinction. For instance, *spend* and *eat* are related since they are associated with shopping malls, but they are not semantically similar. As some verbs in the set are of low frequency, a *don't know* option is included for subjects to mark if they are unsure of the meaning of either verb. We impose no time limits on the subjects to carry out the task, however, the experiment tends to take approximately 20 minutes.

From the full set of 48 verb pairs, 10 are excluded because some participant did not know the definition/intent of one or the other verb in a verb pair item. Furthermore, upon inspecting the materials closely after the experiment, we discover that 11 items did not match the strict rules that we controlled the variability in the set with. The 21 pairs of verbs excluded are distributed evenly across the four aspectual classes we consider for this experiment, therefore we do not believe that the exclusion has a significant impact on the general observations. Accordingly, we are reporting only on 27 verb pairs. Table 6.2 shows the final 27 verb pairs.

bathe	kneel	loosen	open
chill	toughen	neutralize	energize
compose	manufacture	obsess	disillusion
compress	unionize	open	inflate
crinkle	boggle	percolate	unionize
displease	disillusion	plunge	bathe
dissolve	dissipate	prick	compose
embellish	decorate	swagger	waddle
festoon	decorate	unfold	divorce
fill	inject	wash	sap
hack	unfold	weave	enrich
initiate	enter	whisk	deflate
lean	kneel	wiggle	rotate
loosen	inflate		

Table 6.2: The final verb pairs used in the human judgments experiment

Participants in the *Context* group are given exactly the same task, but using the *Context* materials as described before.

Automatic Experiment

The different computational similarity measures are calculated based on the descriptions rendered in Section 6.3. The set of 48 verb pairs are submitted to the respective similarity measure with no contextual information. In summary, each verb pair gets a set of 5 similarity scores computed based on the different automated measures.

6.4.5 Results

We calculate the correlation coefficients of each automated similarity measure with both human conditions, in order to assess the extent to which sets of similarity ratings can predict one another. The correlation metric utilized is the Pearson's r . Table 6.3 show the resulting correlations.

<i>Sim Measure</i>	<i>Context</i>	<i>No-Context</i>
edge_sim	0.720	0.675
res_info	0.779	0.658
lin_info	0.768	0.668
lin_dist	0.453	0.433
lcs_sim	0.313	0.385
Combined	0.872	0.785
InterRater	0.793	0.764

Table 6.3: Comparing the different automated similarity measures to the two human conditions

The **Combined** row of Table 6.3 shows the value of a linear Multiple Regression R when the five computational measures are compared with human ratings (see below); and the InterRater row of the table shows human average InterRater agreement, measured by r , using *leaveoneout* resampling according to Weiss & Kulikowski, (1991).

Inspecting each of the similarity measures individually, we observe that for both conditions *Context* and *No-Context* the taxonomic measures *res_info*, *lin_info* and *edge_sim* outperform the distributional measure *lin_dist* and the lcs *lcs_sim* measures in their correlations with the human ratings.

6.4.6 Discussion

We are not surprised by the low correlation attained by the LCS measure, *lcs_sim*, since we control in our experimental design for the most salient features that make LCSs interesting; we control the variation along the semantic structure dimension.

The results obtained by *lin_dist* are also quite low relative to both human conditions. This measure depends on the syntactic analyses of a large corpus. The low correlation is a reflection of the fact that the measure is dependent on the corpus being utilized and the frequency of the selected verbs for this experiment in that corpus. For example, some low frequency verb pairs such as {*decorate*, *embellish*} and {*dissolve*, *dissipate*} show wide differences with the human ratings.

Examining the correlations obtained via the taxonomic measures, we note the superiority of the information measures to the edge counting measure in the *Context* condition, which is in congruence with the results obtained by Resnik(1999) on nouns;

however, in the *No-Context* condition *edge_sim* is no different from both of the information theoretic measures. At this point we are not certain why this is the case, but it will need to be investigated with a larger set of verb pairs before the results can be conclusive.³ There is no significant difference between the two different information theoretic measures, *lin_info* and *res_info*, they have a Pearson's *r* of 0.96.

Quantitative analysis

Several interesting observations come to the fore when comparing human judgments. First, a comparison of the *Context* and *No-Context* mean ratings by human participants yields $r = 0.89$, which provides some reassurance that participants in the *No-Context* condition are generally interpreting the verbs in the same sense as participants in the *Context* condition — where, as previously stated, the *Context* sentence encouraged interpretation according to the first listed verb sense in Collins Cobuild Dictionary. These results also indicate that the first sense listed in the dictionary is indeed the default sense for participants.

Secondly, average interrater agreement in the two conditions (0.79 and 0.76) is much lower than that obtained in a noun ratings experiment using the same method, where leaveoneout resampling yields an estimate of $r = 0.90$ [73]. This supports our hypothesis that judging verb similarity is a harder task than judging noun similarity; owing to the multidimensional facets of meaning for verbs quantifying their similarity is a more involved task.

Thirdly, we find that participants in the *No-Context* condition have a very strong tendency to assign higher similarity ratings to the same pair when compared to participants in the *Context* condition, as determined using a *Paired TTest* ($N = 27$; $t(26) = 4.49$; $p < 0.002$).

This last observation keeps in line with the notion that participants in the *No-Context* condition are accommodating verb comparisons — providing room for more flexible interpretations of verb meaning — in a manner that is not conveniently accessible to participants in the *Context* condition since their interpretations are constrained by the context sentence.

Finally, we combine the five measures in a Multiple Regression model, *R*, where the measures predict the human ratings. The basic idea is to validate the notion that combining different sources of information (the different models) for verb similarity will yield higher correlations with human ratings. Therefore, we use the similarity scores yielded by each measure as independent variables, and the human ratings in each condition independently, as the dependent variable. The score obtained in the

³For the time being we entertain the hypothesis that the edge counting is more correlated with the *No-Context* condition since people are more liberal in assigning similarity scores in the *No-Context* condition more than in the *Context* condition, which is the case with the edge counting measure which is not sensitive to the weighting of the edges, i.e. small distances such as the distance between *pen* and *ballpen* has the same weight as the edge between *toy* and *artifact*.

Combined row is the multiple regression value resulting from using all $2^5 - 1 = 31$. Looking across the different correlation scores obtained, those yielded by the combination of all five measures is the best predictor of the human ratings. Although the *lcs_sim* and *lin_dist* do not yield high correlations with the human ratings, they do seem to contribute to the predictive power of the regression model since they rely on different sources of information.

In summary, supported by the performance of the models, as well as the improved predictive power of the multiple regression, we interpret the outcomes as evidence that human ratings of similarity are sensitive to both paradigmatic and syntagmatic facets of verb representation, and we posit that the computational models are securing important aspects of verb representation in order to make predictions about similarity judgments.

Qualitative Analysis

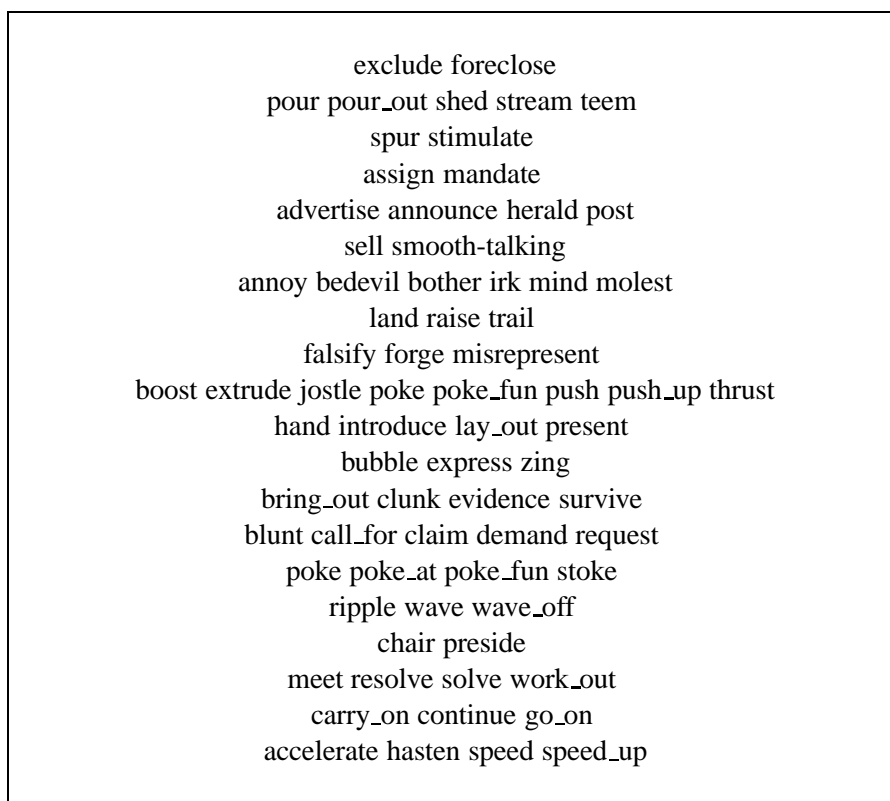
We qualitatively examine the cases where none of the similarity measures assign similarity scores and compare those with the ratings assigned by participants in the experiments; we speculate that the subjects are tapping into dimensions of meaning that are not captured by any of the similarity models and possibly not yet characterized or fully formalized. For example, verb pairs such as *unfold/divorce*, *chill/toughen*, *initiate/enter*, all five measures assign them zeroes, yet the human mean ratings are (average 1.6, 1.4, and 3.2, respectively, in the *No-Context* condition). The ratings are low but they are still higher than some other pairs that did get more than a zero from one of the automated similarity measures such as the verb pair *open/inflate* which gets a human mean rating of (0.6). In many of the cases, the apparent sense extensions seem to verge on the metaphorical: one can describe *divorce* as the *unfolding* of a marriage, a person may *chill* and *toughen* when being insulted, *enter* a group by being *initiated* into it. Attempting to integrate these dimensions of similarity in our models requires a better understanding of how word meanings are portrayed and intertwined, which is a very intriguing line of inquiry for future research but falls outside the scope of our current research.

6.5 Application to SALAAM

Given the very interesting insights obtained from the relation between the different automated similarity measures and the human ratings, we devise a framework for combining different automated measures of verb similarity within SALAAM. As mentioned above, the results obtained on the verb portion, in general by all participating systems, are significantly lower than those yielded for the nouns.

For SALAAM, the results yielded follow the same significant drop when comparing verbs to nouns. We examine the results obtained by condition 4, in Chapter 3,

GLSYS-SP_T, which is the intralanguage merge of the Spanish with the sense selection criterion set to THRESH. We use a variant on the NG algorithm — described in Chapter 3 — adapted for the verb WN1.7pre sense inventory called Verb Grouping (VG), also implemented by Resnik. VG is based on the same exact information theoretic similarity measure used for nouns *res_info* described in Section 6.3.1. It specifically targets the IS-A taxonomy in WN1.7pre. SALAAM yields a precision of 32.4% and recall of 8.9% for verbs, in striking contrast to the 65.7% precision score and 50% recall score for nouns in the same experimental condition. We examine the source type sets, they are very clean, in fact, qualitatively better than noun source type sets within the same condition. We observe that they do not have the pervasive noise attested for in the noun source type sets. Figure 6.1 illustrates a random sample from the verb source sets yielded by SALAAM.⁴



exclude foreclose
 pour pour_out shed stream teem
 spur stimulate
 assign mandate
 advertise announce herald post
 sell smooth-talking
 annoy bedevil bother irk mind molest
 land raise trail
 falsify forge misrepresent
 boost extrude jostle poke poke_fun push push_up thrust
 hand introduce lay_out present
 bubble express zing
 bring_out clunk evidence survive
 blunt call_for claim demand request
 poke poke_at poke_fun stoke
 ripple wave wave_off
 chair preside
 meet resolve solve work_out
 carry_on continue go_on
 accelerate hasten speed speed_up

Figure 6.1: Random sample of verb source type sets yielded by SALAAM

⁴An observation worthy of noting, many of the source sets have the same members since we do not lemmatize the words on the target side of the parallel corpus.

6.5.1 Integrating Human Ratings in SALAAM: A Cognitive Based Feasibility Study

The obtained results are a reflection of the utilization of a similarity model that relies on a single dimension of meaning for verbs, the paradigmatic dimension. As explained earlier, verbs have a very rich multidimensional set of characteristics that are of significant relevance when comparing verbs. Such a reduction of the multidimensionality of the verbs to a single facet of meaning does not capture the possible fine variations between the different senses of the verb. The different senses of a verb are represented in the taxonomy, however, with no explicit distinction for the different syntagmatic characterizations of the verb senses, at least within an IS-A style taxonomy such as WordNet, which is the target of the similarity measure.

Framework

Guided by the observation based on the human experiment that combining different automated similarity measures is a good approximation of human judgments, we set out to describe a framework that is based on the human experiment above for improving verb similarity, primarily for the benefit of SALAAM.

Crucially, the combination of the different measures, in order to be effective, has to rely on various sources of information, various models of the different dimensions of verb meaning. In the experiment described in Section 6.4, we compare human ratings against verb similarity measures that consider paradigmatic and syntagmatic information; the measures differ in their source of data, from WordNet sense inventory to a parsed corpus to LCS representations.

We obtain the correlations between the different automated similarity measures based on a linear regression model. Linear regression models are predictor models. Put simply, the linear regression model yields the best fit of the data to a straight line.⁵ It estimates the weights contributed by a set of random variables X in order to predict a variable y . Regression models are compelling because they are easy to understand, compute and quite often they outperform more complicated prediction models.

The regression equation is as follows:

$$\hat{y} = a_o + \sum_{j=1}^p a_j x_j \quad (6.8)$$

where \hat{y} is the predicted values of the response variable y , a_o is the intercept on the regression plane, a_j are the *regression coefficients* (weights) and x_j are the variables

⁵A valid argument could be that the relation between the different sources of information is not necessarily linear; however it is proven that when we expand a continuous function in a Taylor series, often the lowest terms, which are the linear terms, are the most important, resulting in the best simple approximation yielded by a linear model. [29]

trying to predict y . [29]

Accordingly, the framework we are proposing is applying this model to the SALAAM similarity calculation phase, to the VG algorithm. The idea is using different similarity measures on the verb source type groups and combining the resulting similarity values, crucially, weighted by the coefficients obtained based on a human judgment experiment. As we note earlier in this chapter, the various similarity measures contribute differently to the overall correlation with the human ratings. Therefore, weighting the different similarity measures based on their expected predictive value for the human similarity judgment, renders a cognitively based framework for automatically measuring semantic similarity.

We set out to explore the requirements for testing the proposed framework in order to investigate its impact on SALAAM's performance.

Hypothesis

Combining different automated similarity measures modeling different aspects of verb semantics according to linear coefficients obtained based on a human judgment experiment should yield better SALAAM performance over its performance if the different measures are used separately or simply added together with equal weight.

Resources

Since the end goal is to produce sense tagged data, the need arises for obtaining similarity values between the different senses of the verbs in the verb source type groups. This does not constitute a problem for the *res_info*, *lin_info*, or *edge_sim* since they are applied to the verb taxonomy in WN1.7pre. The problem is encountered by *lin_dist* since is based on parsing a large corpus, it calculates the similarity between verbs at a coarser level than the sense granularity level.

In order to apply *lin_dist* directly, we require the availability of a large enough corpus tagged with verb senses, which does not exist, therefore, *lin_dist* is excluded from this experiment.

As for LCS similarity, we would need a large lexicon of LCS entries, crucially marked with WN1.7pre verb sense ids.

Feasibility Experiment

We conduct a feasibility experiment to test our hypothesis. We use the same metrics defined for SALAAM in condition 4 in Chapter 3, GLSYS-SP.T. The test set is the set of 543 verb instances in the SV2AW English corpus (for details, see Section 3.7.1, in Chapter 3). The used ontology is WN17pre since it is the sense inventory used for the gold standard.

- **Experimentation**

We adapt the VG algorithm to calculate several different similarity values and combine them according to the regression equation, equation (6.8).

We use the similarity measures *res_info*, *lin_info* and *edge_sim* as implemented in the publicly available package `WordNet-Similarity0.03` by Patwardham & Pederson (2003).⁶

We add another similarity measure not mentioned above, an adapted Lesk measure, *lesk* [3]. The adapted Lesk measure uses the basic Lesk algorithm of measuring the amount of overlap in the definitions of two words in a dictionary as a measure of their similarity. In *lesk*, the algorithm is applied to the glosses of the synsets in the WordNet ontology. The rationale behind using such a measure in this *Context* is as an approximation to the *lin_dist* since *lesk* considers the words surrounding the verb in the calculation, therefore implicitly coding syntactic features.⁷

Fortunately, for the LCS measure, *lcs_sim*, many of the approximately 10K lexicon entries are marked with WordNet 1.6 sense ids. We map the sense ids to WN1.7pre using the publicly available mapper.⁸ The entries are expanded to roughly 27K entries indexed by the WN1.7pre sense ids. The entries are converted to the tuple format described in Section 6.3. The measure is applied to the pairs of senses of the verbs according to equation (6.7).

- **Conditions**

1. **Default**

We examine the performance of SALAAM with the individual similarity measures: SALAAM-*edge_sim*, SALAAM-*lin_info*, SALAAM-*res_info*, SALAAM-*lesk*, and SALAAM-*lcs*.

2. **Combined-Equal**

In this condition, the different measures are combined with the same weight, therefore, the coefficients equation (6.8) are $a_j = 1$ where $j = 1 - p$ and the intercept $a_o = 0$.

3. **Combined-Weighted**

⁶<http://www.d.umn.edu/cs/tpederse/research/>

⁷We may obtain data that can be used for the *lin_dist* measure through the sense tagged training data from the SENSEVAL exercises in addition to the verbs' glosses in WordNet and SemCor, yet the size of such a corpus will still be relatively.

⁸<http://www.lsi.upc.es/nlp/tools/mappings.html>

In this condition, the coefficients obtained from the regression model applied to the different similarity measures predicting the human ratings are used.⁹

• Results

Table 6.4 shows the regression coefficients obtained for the different similarity measures applied to the 27 verb pairs used in the experiment described in Section 6.4. In this preliminary study, we only consider the *Context* condition from the human judgments experiment for comparison with the different similarity measures. *lesk*'s correlation with the human *Context* condition is 0.44, which is close to the correlation of *edge_sim* with the same human condition. It is worth noting that the combined correlation coefficient excluding *lin_dist* and including *lesk* is decreased to 0.85. The value for the intercept a_o is -0.482 .¹⁰

Measure	Coefficient
edge_sim	0.247
lin_info	-0.353
res_info	0.325
lcs_sim	-0.014
lesk	-0.001

Table 6.4: Regression Coefficients for the automatic similarity measures

We note that the three taxonomic similarities *edge_sim*, *res_info* and *lin_info* have the highest coefficients corresponding to their correlations with the human judgments as illustrated in Table 6.3. *lcs_sim* follows with a very tiny coefficient, then *lesk* is almost negligible.

The SALAAM performance results for the different conditions described in Section 6.5.1 are shown in Table 6.5.

The results depicted in Table 6.5 are inconclusive due to the lack of statistical significance. But we note some qualitative phenomena, SALAAM-*lcs_sim* and SALAAM-*edge_sim* yield identical performance though they rely on different sources of information.

As expected, the two information theoretic similarity measures SALAAM-*res_info* and SALAAM-*lin_info* yield extremely similar results.

SALAAM-*lesk* produces the best recall results.

⁹We calculate the coefficients for *lesk* using the 27 pairs of verbs used in applying the other automatic similarity measures.

¹⁰We experimented with removing either *res_info* or *lin_info* to avoid co-linearity effects since there was a high correlation between both measures, the results yielded are the same with either metric.

Condition	Precision	Recall
SALAAM-edge_sim	43.2%	2.9%
SALAAM-lcs_sim	43.2%	2.9%
SALAAM-res_info	32.1%	7.9%
SALAAM-lin_info	31.1%	7.9%
SALAAM-lesk	28.4%	8.8%
Combined-Equal	43.2%	2.9%
Combined-Weight	45.8%	2.0%

Table 6.5: SALAAM performance results with different similarity measure conditions

We observe the precision of the Combined-Weight condition exceeds all the other measures in precision.

Despite the modest results, we note the increase in precision from Combined-Equal to Combined-Weight condition.

Discussion

In fact, the precision results obtained by Combined-Weight and Combined-Equal are close to the high end of the scores obtained by state-of-the-art WSD as exemplified by the SENSEVAL 2 exercise. The VG assigns high confidence to several senses leading to partial credit by `scorer2`, which affects the precision negatively. Nonetheless, the precision score obtained by Combined-Weight qualitatively exceeds any of the individual similarity score conditions as well as the Combined-Equal.

The drop in recall for both combined conditions is potentially explainable by the fact that all the measures return similarity values for the same verb senses, but given that the coefficient weights are not absolute, they neutralize each other leading to a loss in confidence scores. This hypothesis is supported by the Combined-Equal results since all of the measures are given equal weight.

The recall in general is extremely low which is mainly attributed to the shallowness and bushiness of the taxonomy. This is reflected in all the recall values obtained. This observation is confirmed by the results obtained by SALAAM-edge_sim. The edge counting is the most affected by the shallow depth of the ontology. If all the senses are on similar levels, they will be equidistant, then there is no bias to choose one sense over the other, which is reflected in the uniform confidence scores obtained by VG.

As for SALAAM-lcs_sim, most likely it is a problem of coverage of the WN1.7pre. However, majority of the senses did not exist in the LCS lexicon.

For SALAAM-lesk, the problem is distinctive glosses in WN1.7pre. Many of the glosses for the verb senses overlap very highly, rendering the choice among the different senses very hard, leading to low confidence by the VG algorithm.

All in all, we conclude with the following observations:

- SALAAM for verbs produces very good quality source sets.
- The IS-A taxonomy in WN1.7pre is not sufficient as a knowledge representation source since it only focuses on the manner aspect of the verb. Therefore, it should be interesting to apply these similarity measures to other relations in the taxonomy such as meronymy;
- The lack of a distinct measure that depends on an explicit syntactic model seems to play a distinct role in the regression coefficient value distributions.

6.6 Summary

In summary, we presented a novel design for obtaining human ratings on verb semantic similarity with *Context* and with *No-Context*. We use the human similarity judgments as a pivot for comparing different automatic similarity measures. Crucially, we conclude that combining evidence from different similarity measures that rely on different sources of information yields the highest correlation with human ratings in both *Context* and *No-Context* conditions for the human experiment.

We then present a framework for incorporating these observations in an actual WSD system, SALAAM, where the goal is to improve results for verb similarity. We have concluded that WordNet IS-A verb taxonomy is not the best source of semantic similarity especially if it were the only source of semantic relations between verb senses for all measures.

Chapter 7

Conclusions & Future Directions

7.1 Conclusions

The overall theme in this thesis is the search for non traditional sources of evidence and the combination of these sources in functional ways for the benefit of gaining insight into quantifiable aspects of word meaning. To that end, we investigate the characterization of an age old complex problem of functionally resolving word ambiguity, WSD, using evidence from translations into different languages. We address the problem of WSD from a multilingual perspective; we expand the notion of context to encompass multilingual evidence. We devise a new approach to resolve word sense ambiguity in natural language, using a source of information that was never exploited on a large scale for WSD before. We develop an algorithm that empirically investigates the feasibility and the validity of utilizing translations for WSD. The algorithm is an unsupervised approach, SALAAM, for word sense tagging large amounts of text given a parallel corpus and a sense inventory for one of the languages in the corpus. We evaluate the approach using machine translated parallel corpora, pseudo-translations. The performance for English nouns — as the source language in the parallel text — in a SENSEVAL 2 defined test set is rigorously evaluated using community-wide available tools and compared against state-of-the-art WSD systems. The results yielded are superior to all unsupervised methods evaluated on the same test set, moreover, SALAAM rivals some of the supervised methods and partially supervised methods. We observe that evidence from several languages aids SALAAM's performance. We conclude that language distance does have some impact on the quality of the results obtained. We quantifiably show the complementarity of a multilingual approach to monolingual approaches. We empirically establish that translation is a good source of sense distinction, thereby, lending solid support to the characterization of word meaning using translational correspondence.

Figure 7.1 summarizes SALAAM's performance against state of the art WSD systems which all rely on monolingual contexts.

Furthermore, Table 7.1 and Table 7.2 illustrate the significant departure of multilingual WSD from toy systems in terms of evaluation to large scale standardized

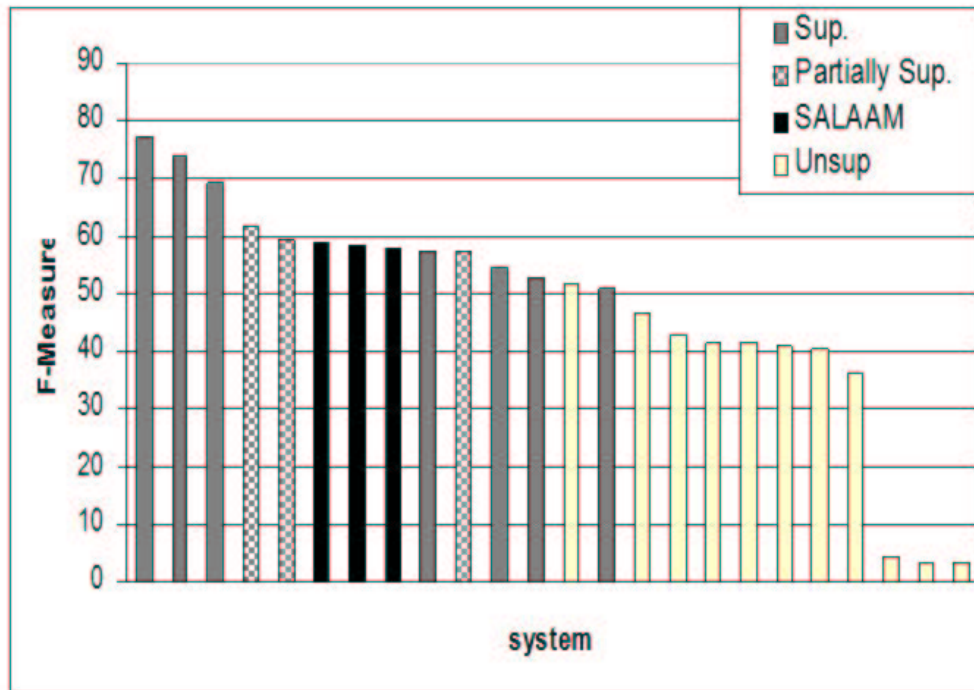


Figure 7.1: SALAAM F-Measure results depicted against state-of-art WSD systems

evaluation in the current thesis.

System	Method	Corpus Type	Inventory	# Label	Linguistic Tools
Brown et al	Sup.	Parallel		2 words	Tokenizers
Gale et al	Sup.	Parallel		2 words	Tokenizers
Dagan & Itai	Unsup.	Comparable	Biling. Dict.	n words	Tokenizers & Parsers
Kikui	Unsup.	Comparable	Biling. Dict.	n words	Tokenizers
SALAAM	Unsup.	Parallel	WordNet	n senses	Tokenizers

Table 7.1: Summary of Multilingual WSD Systems' Required Resources

SALAAM's robustness is tested with naturally occurring parallel corpora of genre types that are unrelated to the test set. The results obtained show no significant difference in performance precision for SALAAM using pseudo-translations of relevant corpora genre versus utilizing unrelated genre corpora for augmenting the test corpus.

Having established SALAAM as a good tagger for the source language, we investigate its tagging quality of the target language of the parallel corpus. We examine two target languages: Arabic and Spanish. The results obtained from Arabic demonstrate that 90.5% of the correct tags for English noun instances are appropriate tags for Ara-

System	Languages	Metric	Size	GS	Performance
Brown et al.	En-Fr	improv.	100 inst.	No	8% improv.
Gale et al	En-Fr	acc.	6 words, 140 inst./word	No	90% acc.
Dagan & Itai	Heb-En, Ger-En	prec., applic.	103 Heb, 54 Ger inst.	No	91% prec, 63%applic.
Kikui	En-Jap	acc.	120 inst.	Yes	79.1% acc.
SALAAM	En-Fr, En-Sp, En-Ar	prec., rec., FM	1071 noun inst	Yes	64.5% prec. 53% rec. ¹

Table 7.2: Summary of Multilingual WSD Systems' Evaluation

bic. SALAAM, as a tagger for the target language, zones in on the commonality of sense usage cross linguistically, in effect, quantifying meaning characterizations for a language with poor resources — such as Arabic — via its shared sense usages cross linguistically through a language such as English with rich resources. This usage of SALAAM is quite different from its application to the source language; in tagging a source language, SALAAM exploits divergences of meaning representation, but in tagging the target language, SALAAM exploits commonality. On the other hand, we perform a fully automated blind evaluation of the quality of projected tagging for Spanish data, despite severe lack of resources for this experiment. The results obtained are modest even though they significantly improve on a random baseline. The main reason for the modest performance is attributed to the use of source pseudo-translations accompanied by inconsistencies in alignments, therefore detrimentally affecting the quality of the tagging. But nonetheless, the technique presented is a new technique that is fully automated and, except for the parallel corpus, requires minimal resources.

Furthermore, SALAAM, as an algorithm, is explored as a method for seeding a WordNet style ontology for Arabic. By quantitative inspection the approach seems promising. We discuss different issues of representation for Arabic specifically. We conclude that stems as a first step are the appropriate level of representation for a taxonomic style ontology for a Arabic.

We view SALAAM yet from different angle, it is exploited as an inexpensive source for large amounts of acceptable quality sense annotated data. The produced annotated data enables us to investigate the trade-off between quantity and quality of annotated data for supervised learning WSD. We undertake a study to empirically explore the feasibility of bootstrapping a supervised learning WSD system using SALAAM tagged training data instead of human tagged data. In essence, we use SALAAM as an unsupervised learning approach for WSD. SALAAM produces several different tagged unsupervised data sets. Bootstrapping a machine learning WSD system using noisy data from SALAAM is shown to yield better performance than state-of-the-art bootstrapping performance, by Mihalcea [58], using clean tagged data on the same data set in a completely supervised learning experimental set up. We obtain PRs of > 0.90 for 12 of the data items tested compared to Mihalcea's system performance which yields the same PR as SALAAM on 6 data items only.

Figure 7.2 illustrates the results, specifically PR values, obtained using SALAAM training data versus PR values obtained by Mihalcea's system. The hashed bars in the graph are the PR values obtained by Mihalcea while the solid bars are those of

SALAAM.

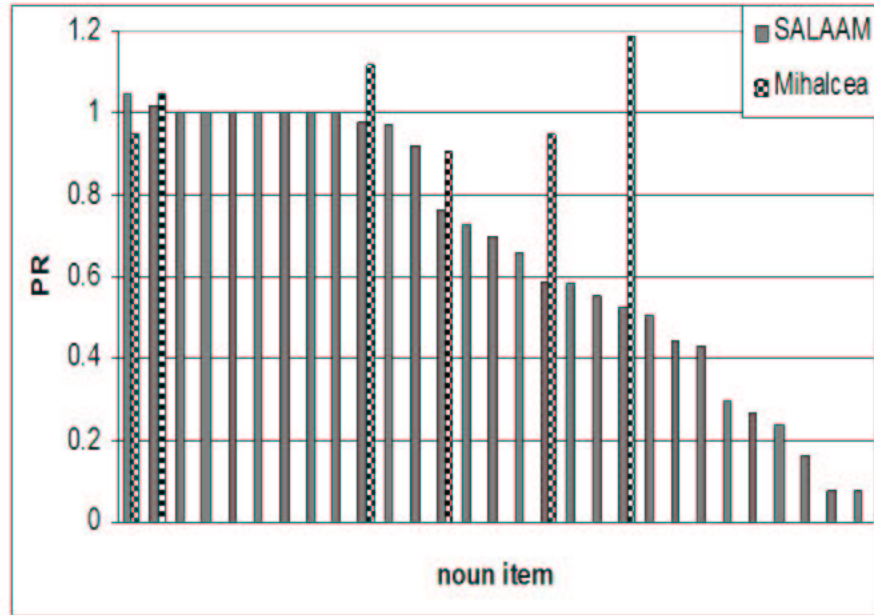


Figure 7.2: Comparison between Mihalcea’s results and SALAAM results on the same test set

Moreover, SALAAM rivals UMSST’s — a canonical supervised learning WSD system trained on human tagged data — performance on these 12 items. We analyze the different factors affecting the performance of the bootstrapping system with the intent of quantifying good predictors of good performance given noisy data. We conclude that different factors play some role, but the major contributors were Sense Context Confusability and Sense Distribution Contexts.

With the central role played by similarity in the SALAAM sense disambiguation method, we set out to investigate different aspects of semantic similarity. Driven by the fact that all WSD systems participating in the SENSEVAL 2 All-Words exercise did much worse on verbs than nouns and also by the fact that verbs are very interesting and complex entities in natural language in and of themselves, we explore dimensions of semantic similarity of verbs. We devise a novel experimental design for obtaining human judgments for verb similarity, where the subjects are given test items in **Context** and with **No Context** in an attempt at measuring the effect of context on human similarity ratings. We compare different automated similarity measures that crucially rely on various sources of information with the obtained human judgments. The intent is to provide a framework for quantifying the amount of contribution that should be attributed to these different sources of information based on insights derived from cog-

natively based studies of verb similarity. As expected, the combination of all automated measures examined in this investigation yields the highest correlation with the human ratings. Accordingly, we provide a cognitively based framework for combining evidence for multidimensional verb similarity measures that may be utilized to improve results obtained by WSD systems, in general, and SALAAM in particular, in the task of verb disambiguation.

7.2 Thesis Problems & Limitations

- In the current implementation of SALAAM, we note that it is limited by its dependence on the availability of parallel texts.
- SALAAM's performance is sensitive to the alignment quality and translation variability in text; it is sensitive to noise in the source type sets.
- The quality of the projected sense tagging of Arabic is tested with only one annotator, more annotators need to inspect the data with the results being judged taking inter-annotator agreements into consideration.
- SALAAM is tested with bad quality source for Spanish target data set which does not allow us to draw conclusive results regarding the quality of the projected Spanish sense annotations
- In the semantic similarity experiment in Chapter 6, the data is limited with the number of verb pairs 27 pairs tested on 10 participants only.

7.3 Research Contributions

This thesis contributed the following to the field of computational linguistics:

- A novel robust unsupervised approach to WSD which constitutes a significant departure from the traditional monolingual approaches. The approach is a validation of a sound linguistic assumption that meaning characterizations can be captured cross linguistically. We contribute a novel multilingual perspective on the notion of context for addressing the problem of WSD. Context scope is no longer confined monolingually.
- The thesis provides a detailed description of an end-to-end fully operational, modularly designed system for producing large amounts of good quality sense annotated data in both source and target languages for a parallel corpus. Given a token aligned parallel corpus, SALAAM can produce a fully annotated corpus in less than an hour.

- The thesis investigates the quality of automatic sense annotations for a language with few computerized linguistic resources such as Arabic.
- The thesis provides an operational end-to-end automatic framework for testing the quality of projected automatic sense annotations for Spanish.
- The thesis examines the feasibility of automatically bootstrapping a WordNet style ontology for Arabic via projected sense tags from English and concluded that is a tractable task given a large balanced parallel corpus.
- The thesis investigates the feasibility of bootstrapping WSD within a supervised learning paradigm using noisy data based on the results obtained using SALAAM data, thereby introducing a novel unsupervised approach for WSD since even in bootstrapping mode, SALAAM does not require any hand tagged data. SALAAM yields results that are superior to those obtained by the state-of-the-art bootstrapping method on the same test set; simultaneously, SALAAM rivals a canonical supervised system UMSST's performance on 12 out of 29 noun items of the SENSEVAL 2 test set.
- The thesis contributes a novel design for attaining human judgments on semantic similarity for verb pairs using contextual and non-contextual data. The thesis compared the results obtained by several automated semantic similarity measures against the human similarity ratings.
- The thesis utilizes insights derived from the human similarity judgment experiment to motivate an operational cognitively based framework for exploiting similarity in a novel way in order to improve WSD results obtained for verbs.

7.4 Future Directions

- Combining Monolingual Evidence from Monolingual Context with Multilingual Evidence: The source of evidence for word sense tagging using SALAAM is orthogonal to typical monolingual approaches that rely on the monolingual contexts of the polysemous words to resolve their ambiguity. In SALAAM, at this stage, monolingual evidence is disregarded. For instance, if the word *bank* occurs in a sentence such as **She walked by the river bank** the fact that *bank* is preceded by *river* does not play a role in the sense selection phase of SALAAM for *bank*. Moreover, given the encouraging quantitative results on complementarity of SALAAM with monolingual approaches, we can visualize explicitly exploiting the monolingual contextual information of the polysemous words as a means of constraining the sense inventory search space. So in this case of the given example, where the **river bank** sense is intended, only senses that

are related to the geographical sense of *bank* are to take part in the sense selection phase. A simple method of implementing this extension is to use the Lesk Algorithm in matching the overlapping context of words in the corpus and glosses in WordNet as a preliminary sense tagging step then applying SALAAM. This has the advantage of reducing the search space and introduces bias in the source type sets which could potentially aid in the sense selection process for the other words in the set. Such monolingual evidence can be obtained by bracketing a corpus or even parsing it in order to attain even more linguistically interesting biases for the appropriate sense of a word in question.

- Subclustering Source sets: We noted in the discussion section in Chapter 3 the detrimental effect of noise in the source sets on the performance of SALAAM. Many of the problems emerged from the presence of multiple clusters in the source sets. Therefore, it would be worthwhile to use quantitative clustering techniques on the source data token sets to split them into more coherent sub source sets.
- Application to comparable corpora: Comparable corpora are more widely available than parallel corpora. A corpus is considered comparable if the two corpora are of the same genre and the same time frame and size. Methods of finding translation equivalents in comparable corpora are very promising. One such method by Diab & Finch [17] introduced a novel unsupervised greedy algorithm that produces very reliable results for comparable corpora. Once we have the translation equivalents and we have a sense inventory for one of the languages of a comparable corpus, SALAAM may be directly applicable to the corpora at hand.
- Evaluating the quality of the projected Arabic sense annotations with more human annotators
- Evaluating the quality of the projected Spanish sense annotations using good quality source English text. This is achievable by human translating the Spanish Lexical Sample corpus into English.
- Implementing a system for predicting the conditions per item that would yield good enough examples with the right pedigree for narrowing the gap between training with noisy tagged data and hand tagged data
- Testing the bootstrapping approach for supervised WSD with the Spanish SEN-SEVAL 2 data
- More analysis of the verb data and operationally incorporating the results for improving the performance of SALAAM on verb sense tagging. This is achievable by adding more verb pairs as well as participants to the human judgments

experiment. Building resources of sense annotated corpora in order to render the *lin_dist* part of this verb similarity investigation (see Chapter 6).

- Using the SALAAM multilingual framework to distinguish homonymy and polysemy in the WordNet ontology: we believe that homonymy should be explicitly marked in the WordNet, thereby creating a multidimensional WordNet taxonomy.

BIBLIOGRAPHY

- [1] Eneko Agirre, Jordi Atserias, Luis Padró, and German Rigau. Combining Supervised and Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. *Computers and the Humanities*, 34:50–58, 2000.
- [2] Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. Statistical Machine Translation: Final Report. In *Summer Workshop on Language Engineering*. John Hopkins University Center for Language and Speech Processing, 1999.
- [3] Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Aqapulco, Mexico, August 2003.
- [4] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, California, 1984.
- [5] Eric Brill. Transformation-based tagger, version 1.14, 1995.
- [6] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.
- [7] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. A statistical approach to sense disambiguation in machine translation. In *Fourth DARPA Workshop on Speech and Natural Language*, Pacific Grove, CA, February 1991.
- [8] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 1993.
- [9] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.D. Lafferty, and R.L. Mercer. Analysis, Statistical Transfer, and Synthesis in Machine Translation. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pages 83–100, Montreal, Canada, 1992.
- [10] Rebecca Bruce and Janyce Wiebe. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, Las Cruces, New Mexico, June 1994.

- [11] Tim Buckwalter. Buckwalter Arabic Morphological Analyzer Version 1.0., LDC Catalog No.: LDC2002L49. Linguistic Data Consortium, University of Pennsylvania, 2000.
- [12] Clara Cabezas, Philip Resnik, and Jessica Stevens. Supervised sense tagging using support vector machines. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, Toulouse, France, July 2001.
- [13] Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer, editors. *SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, July 2001. ACL SIGLEX. <http://www.sle.sharp.co.uk/senseval2/>.
- [14] D. Cruse. *Lexical Semantics*. Cambridge University Press, 1986.
- [15] Ido Dagan and Alon Itai. Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*, 20(4):563–596, 1994.
- [16] Kareem Darwish. Building a shallow arabic morphological analyzer in one day. In *Proceedings of ACL Workshop on Semitic languages*, Pennsylvania, USA, 2002.
- [17] M. Diab and S. Finch. A Statistical Word-Level Translation Model for Comparable Corpora. In *Proceedings of RIAO 2000 Conference*, April 2000.
- [18] Mona Diab. An unsupervised method for multilingual word sense tagging using parallel corpora: A preliminary investigation. In *SIGLEX2000: Word Senses and Multi-linguality*, Hong Kong, October 2000.
- [19] Bonnie J. Dorr. *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge, MA, 1993.
- [20] Bonnie J. Dorr. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(4):271–322, 1997.
- [21] Bonnie J. Dorr. LCS Verb Database. Technical Report Online Software Database, University of Maryland, College Park, MD, 2001. http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html.
- [22] Bonnie J. Dorr, M. Antonia Martí, and Irene Castellón. Spanish EuroWordNet and LCS-Based Interlingual MT. In *Proceedings of the Workshop on Interlinguas in MT, MT Summit, New Mexico State University Technical Report MCCA-97-314*, pages 19–32, San Diego, CA, October 1997.
- [23] Helge Dyvik. Translations as semantic mirrors, 1998.
- [24] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998. <http://www.cogsci.princeton.edu/~wn> [2000, September 7].
- [25] Christiane Fellbaum, Martha Palmer, Hoa Trang Dang, Lauren Delfs, and Susanne Wolff. Manual and Automatic Semantic Annotation with WordNet. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations*, Carnegie Mellon University, Pittsburg, PA, 2001.

- [26] W. Francis and H. Kučera. *Frequency Analysis of English Usage*. Houghton Mifflin Co.: New York, 1982.
- [27] William A. Gale, Kenneth W. Church, and David Yarowsky. Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112, Montréal, Canada, June 1992.
- [28] Jane Grimshaw. Semantic Structure and Semantic Content in Lexical Representation. 1993.
- [29] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. MIT Press, Boston, 2001.
- [30] Philip Hayes. On Semantic Nets, Frames and Associations., 1977. Proceedings of the 5th International Joint Conference of Artificial Intelligence.
- [31] Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. Evaluating Translational Correspondence using Annotation Projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, 2002.
- [32] Nancy Ide. Cross-lingual sense discrimination: Can it work? *Computers and the Humanities*, 34:223–34, 2000.
- [33] Nancy Ide and Jean Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. In *Computational Linguistics 24(1)*: 1-40, 1998.
- [34] Julio Gonzalo Irina Chugur and Felisa Verdejo. Polysemy and sense proximity in the senseval-2 test suite. In *Proceedings of Word Sense Disambiguation: Recent Successes and Future Directions*, University of Pennsylvania, Pennsylvania, July 2002.
- [35] Ray Jackendoff. *Semantics and Cognition*. The MIT Press, Cambridge, MA, 1983.
- [36] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*. Springer, 1998.
- [37] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, New Jersey, USA, 2000.
- [38] Edward Kelly and Philip Stone. Computer Recognition of English Word Senses, 1975.
- [39] Genichiro Kikui. Resolving translation ambiguity using non-parallel bilingual corpora. In *Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language Processing.*, College Park, Maryland, 1999.
- [40] A. Kilgariff and J. Rosenzweig. Framework and Results for English SENSEVAL. *Computers and the Humanities*, 34:15–48, 2000.
- [41] Adam Kilgariff. SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs, 1977.

- [42] Adam Kilgarriff. Inheriting Polysemy. In P. Saint-Dizier and E. Viegas, editors, *Computational Lexical Semantics*, pages 319–335. Cambridge University Press, England, 1995.
- [43] Michael E. Lesk. Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the SIGDOC Conference*, 1986.
- [44] Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL, 1993.
- [45] Beth Levin and Malka Rappaport Hovav. The Elasticity of Verb Meaning. In *Proceedings of the Tenth Annual Conference of the Israel Association for Theoretical Linguistics and the Workshop on the Syntax-Semantics Interface*, University of Haifa, Israel/Ben Gurion University of the Negev, Be’er Sheva, Israel, June 12–13 1994.
- [46] Beth Levin and Malka Rappaport Hovav. From Lexical Semantics to Argument Realization. Technical report, Northwestern University, October 1996. <http://www.ling.nwu.edu/~beth/pubs.html>.
- [47] Beth Levin and Malka Rappaport Hovav. Building Verb Meanings. In M. Butt and W. Geuder, editors, *The Projection of Arguments: Lexical and Compositional Factors*, pages 97–134. CSLI Publications, Stanford, CA, 1998.
- [48] Dekang Lin. Government-Binding Theory and Principle-Based Parsing. Technical report, University of Maryland, 1995. Submitted to Computational Linguistics.
- [49] Dekang Lin. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, July 1997.
- [50] Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL98*, Montreal, Canada, 1998.
- [51] Dekang Lin. Dependency-Based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, Granada, Spain, May 1998.
- [52] Adam Lopez, Michael Nossal, Rebecca Hwa, and Philip Resnik. Word-level alignment for multilingual resource acquisition. In *Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data, at the Third International Conference on Language Resources and Evaluation (LREC-2000)*, Las Palmas, Canary Islands, Spain, June 2002.
- [53] M. McCord. Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars. In *In R. Studer (Ed.), Natural Language and Logic*, pages 118–145, Berlin, Heidelberg, 1990.
- [54] Dan I. Melamed. Measuring semantic entropy. In *SIGLEX Workshop on Tagging Text with Lexical Semantics*. ACL, 1997.

- [55] I. Dan Melamed. Models of Translational Equivalence among Words. *Computational Linguistics*, 26(2):221–249, June 2000.
- [56] I. Dan Melamed and Philip Resnik. Evaluation of sense disambiguation given hierarchical tag sets. *Computers and the Humanities*, (1–2), 2000.
- [57] R. Mihalcea and D. Moldovan. A method for word sense disambiguation of unrestricted text, 1999.
- [58] Rada Mihalcea. Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd International Conference on Languages Resources and Evaluations (LREC-2000)*, Las Palmas, Canary Islands, Spain, June 2002.
- [59] Rada Mihalcea and Dan I. Moldovan. Word sense disambiguation based on semantic density. In Sanda Harabagiu, editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 16–22. Association for Computational Linguistics, Somerset, New Jersey, 1998.
- [60] Arthur Nadas. A decision-theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-31(4):814–817, August 1983.
- [61] Hwee Tou Ng and Hian Beng Lee. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In *Proceedings of the 34th Annual Conference of the Association for Computational Linguistics*, pages 40–47, Santa Cruz, CA, June 1996.
- [62] Franz J. Och and Hermann Ney. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, pages 440–447, Hongkong, China, October 2000.
- [63] Ted Pedersen. Machine learning with lexical features: The duluth approach to senseval 2. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, July 2001.
- [64] Steven Pinker. *Learnability and Cognition: The Acquisition of Argument Structure*. The MIT Press, Cambridge, MA, 1989.
- [65] M.R. Quillian. Semantic Memory. In M. Minsky, editor, *Semantic Information Processing*. The MIT Press, Cambridge, MA, 1968.
- [66] Malka Rappaport Hovav and et al. Levels of Lexical Representation. *Semantics and the Lexicon*, pages 37–54, 1983.
- [67] Philip Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, December 1993.
- [68] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI-95*, pages 448–453, Montreal, Canada, August 20–25 1995.

- [69] Philip Resnik. Selectional Preference and Sense Disambiguation. Technical report, University of Maryland, 1997.
- [70] Philip Resnik. Disambiguating Noun Groupings with Respect to WordNet Senses. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 77–98. Kluwer Academic, Dordrecht, 1999.
- [71] Philip Resnik. Mining the Web for Bilingual Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, University of Maryland, College Park, Maryland, June 1999.
- [72] Philip Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Artificial Intelligence Research*, (11):95–130, 1999.
- [73] Philip Resnik, Mari Olsen, and Mona Diab. The Bible as a Parallel Corpus: Annotating the Book of 2000 Tongues. *Computers and the Humanities*, (33):129–153, 1999.
- [74] Philip Resnik and David Yarowsky. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering*, 1(1):1–25, 1998.
- [75] Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124, 1998.
- [76] John Sinclair, editor. *Collins Cobuild English Dictionary*. Collins, 1995. Patrick Hanks, managing editor.
- [77] Steven Small. Word Expert Parsing: A Theory of Distributed Word Based Natural Language Understanding, September 1980. Doctoral Dissertation. Computer Science Department, University of Maryland.
- [78] P. Vossen, W. Peters, and J. Gonzalo. Towards a Universal Index of Meaning. pages 1–24, 1999.
- [79] Piek Vossen, Pedro Diez-Orzas, and Wim Peters. The Multilingual Design of EuroWordNet. In *Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Application*, Madrid, Spain, 1997.
- [80] W. Weaver. Translation(1949). In *Machine Translation of Languages*. MIT Press, Cambridge, MA, 1955.
- [81] Yorick Wilks. Preference Semantics. In E.L. Keenan, editor, *Formal Semantics of Natural Language*. Cambridge University Press, Cambridge, MA, 1975.
- [82] Louise Guthrie Wim Peters and Yorick Wilks. Cross-linguistic discovery of semantic regularity, 2001.

- [83] D. Yarowsky. Word-Sense Disambiguation: Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 454–460, Nantes, France, 1992.
- [84] David Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *In Proceedings of COLING-92*, Nantes, France, 1992.
- [85] David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 189–196, Cambridge, MA, 1995.