

WORD SENSE DISCOVERY AND DISAMBIGUATION

Krister LINDÉN

*Academic dissertation to be publicly discussed,
by due permission of the Faculty of Arts at the University of Helsinki,
in auditorium 107 at the Department of General Linguistics,
on the 11th of June, 2005, at 10 o'clock.*

University of Helsinki
Department of General Linguistics
P.O. Box 9
FIN-00014 University of Helsinki
Finland

PUBLICATIONS
No. 37
2005

©Krister Lindén

ISSN 0355-7170
ISBN 952-10-2471-2 (bound)
ISBN 952-10-2472-0 (PDF)

<http://ethesis.helsinki.fi>

Helsinki 2005
Helsinki University Press

*Suit the action to the word,
the word to the action.*

— William Shakespeare (c. 1600-01)
Hamlet

Abstract

The work is based on the assumption that words with similar syntactic usage have similar meaning, which was proposed by Zellig S. Harris (1954,1968). We study his assumption from two aspects: firstly, different meanings (word senses) of a word should manifest themselves in different usages (contexts), and secondly, similar usages (contexts) should lead to similar meanings (word senses).

If we start with the different meanings of a word, we should be able to find distinct contexts for the meanings in text corpora. We separate the meanings by grouping and labeling contexts in an unsupervised or weakly supervised manner (Publication 1, 2 and 3). We are confronted with the question of how best to represent contexts in order to induce effective classifiers of contexts, because differences in context are the only means we have to separate word senses.

If we start with words in similar contexts, we should be able to discover similarities in meaning. We can do this monolingually or multilingually. In the monolingual material, we find synonyms and other related words in an unsupervised way (Publication 4). In the multilingual material, we find translations by supervised learning of transliterations (Publication 5). In both the monolingual and multilingual case, we first discover words with similar contexts, i.e., synonym or translation lists. In the monolingual case we also aim at finding structure in the lists by discovering groups of similar words, e.g., synonym sets.

In this introduction to the publications of the thesis, we consider the larger background issues of how meaning arises, how it is quantized into word senses, and how it is modeled. We also consider how to define, collect and represent contexts. We discuss how to evaluate the trained context classifiers and discovered word sense classifications, and finally we present the word sense discovery and disambiguation methods of the publications.

This work supports Harris' hypothesis by implementing three new methods modeled on his hypothesis. The methods have practical consequences for creating thesauruses and translation dictionaries, e.g., for information retrieval and machine translation purposes.

Preface and Acknowledgements

One could say that the work on this thesis has its roots in my childhood when I was 9 years old. I had already made two significant decisions in my life. I wanted to be a scientist and I intended to invent a speech translation device so that people wouldn't have to learn foreign languages. At that time PCs and mobile phones were still unknown. I already spoke two languages and was learning English at school, but it was the prospect of having to move to Finland and learn a fourth and radically different language, Finnish, that begot these lofty ideas.

Language Technology as a subject was not yet invented, when I began my studies at the Computer Science Department at Helsinki University. After graduation I joined the Department of Linguistics, where I was involved in an English-to-Finnish Machine Translation project under the supervision of Dr. Lauri Carlson. His vast knowledge of both applied and formal linguistics combined with his down-to-earth remarks paved my way to linguistics. During that period Dr. Kimmo Koskenniemi became the first professor of Computational Linguistics in Finland, and Prof. Fred Karlsson was Head of the Linguistics Department leading a project on constraint grammar. This environment was tremendously inspiring and their ideas and views on morphology, surface syntax, constraint grammar, and translation I will forever carry with me.

I was, however, thrown into the business world of Language Technology, where I had the opportunity to participate in the start-up of a company called Lingsoft, which I headed for a number of years, before I went on to be its Chief Technology Officer. At Lingsoft I took part in a range of interesting projects. When doing language technology for information retrieval, I was fortunate to meet Prof. Kalervo Järvelin, at Tampere University (UTA). During a project on the Finnish dictionary, Perussanakirja, I met Dr. Krista Lagus, now at the Helsinki University of Technology (HUT). Both were later to become supervisors of this Ph.D. thesis. As one of my projects at Lingsoft I also designed and supervised the implementation of a Finnish speech recognition system.

By that time I had taken part in the development of all the necessary components for the speech translation device I had set out to create in childhood. Why then a dissertation on word senses? Well, after having seen all the components, I

was also acutely aware of how much work remained for a practical large-scale solution. I needed a new angle. Applied lexical semantics and a statistical approach seemed necessary. Prof. Lauri Carlson, at Helsinki University (HU), accepted to be the main supervisor of my thesis, and the project was made financially possible during 2001-2005 by the Graduate School of Language Technology in Finland and the Department of General Linguistics at the University of Helsinki, which I gratefully acknowledge.

During my time at the Graduate School of Language Technology, I had the privilege to cooperate with graduate students from three different universities. Foremost among those have been Mathias Creutz at HUT, Jussi Piitulainen at HU, and Heikki Keskustalo at UTA. Together with them I was able to make some of the ideas materialize into publications.

As a complement to the intellectual work, I have enjoyed folk dancing several times a week. So much so that I now also hold a degree as a folk dancing instructor. I am grateful to the folk dancers at Arbetets Vänner and Brage for providing me with relaxing, playful, but also challenging and stimulating environments for folk dancing. Merry motions always bring about a good mood.

I am also deeply indebted to my parents, Stig and Eva, without whose unfailing belief in my capabilities, and without whose decision long ago to move back to Finland, this thesis may never have happened, and to my sister, Lilian, for many discussions on the meaning of everything, and finally, to my partner in life, Juhani, without whose delicious food and good-natured support I would have felt much lonelier.

Contents

Abstract	iv
Preface and Acknowledgements	v
1 Introduction	1
1.1 Word Sense Disambiguation	2
1.2 Word Sense Discovery	3
1.3 Open problems	4
1.3.1 Word Senses	4
1.3.2 Context	5
1.3.3 Evaluation	5
1.4 Duality of Disambiguation and Discovery	6
1.5 Organization of the Thesis	6
1.6 Publications	7
1.6.1 List of Publications	7
1.6.2 Summaries and Contributions	7
2 Word Senses	10
2.1 Language Philosophy	10
2.2 Enumeration vs. Generation	11
2.3 The Origin of Features	12
2.4 Recording Word Senses	13
2.4.1 Frequency Distribution	14
2.4.2 Acceptability in Context	15
2.5 Word Sense Dictionary Specification	17
2.6 Existing Sense Inventories	18
2.6.1 WordNet	18
2.6.2 Finnish Synonym Dictionary	19
2.6.3 Cross-lingual Terminology	20
2.7 Summing up	20

3	Context	22
3.1	Dimensions of Context	22
3.1.1	Context Size	23
3.1.2	Context Modality	24
3.1.3	Depth of Preprocessing	25
3.2	Related Work	27
3.3	Collecting Contexts	28
3.3.1	Corpus Collection	28
3.3.2	Annotated Corpora	29
3.4	Representing Context	30
3.5	Summing up	32
4	Evaluation	33
4.1	General Performance	34
4.1.1	Evaluation Measures	34
4.1.2	Division of Training and Evaluation Data	36
4.1.3	Training Data and Parameter Tuning	36
4.2	Using Annotated Contexts	37
4.2.1	Real Word Contexts	37
4.2.2	Pseudo-Word Contexts	38
4.2.3	Application-Specific Contexts	38
4.2.4	Context Preprocessing	39
4.3	Using Word Sense Inventories	40
4.3.1	Obstacles	40
4.3.2	Real Word Inventories	40
4.3.3	Pseudo-Word Inventories	41
4.3.4	Application-Specific Word Inventories	42
4.4	Summing up	42
5	Disambiguation	43
5.1	Manually Developed Resources	43
5.1.1	Early Ideas	44
5.1.2	Artificial Intelligence Ideas	45
5.1.3	Knowledge-based Ideas	45
5.2	Machine Learning Methods	46
5.2.1	Supervised	46
5.2.2	Semi-Supervised	46
5.2.3	Unsupervised	47
5.2.4	Combining classifiers	47
5.2.5	Degrees of Supervision	47
5.3	Word Sense Discrimination	48

5.4	THESSOM	49
5.4.1	Background	49
5.4.2	Creating Document Maps	50
5.4.3	Calibration and Disambiguation	50
5.4.4	Test results	51
5.5	Summing up	51
6	Discovery	53
6.1	Early Ideas	54
6.1.1	Word Sense Similarity	54
6.1.2	Word Sense Clustering	55
6.2	Monolingual Discovery	56
6.2.1	Corpus Data	57
6.2.2	Feature Extraction	57
6.2.3	Similarity Calculations	57
6.2.4	Clustering	57
6.2.5	Evaluation	58
6.2.6	Test results	59
6.3	Multilingual Discovery	59
6.3.1	Probabilistic Framework	61
6.3.2	Data	61
6.3.3	Test results	62
6.3.4	Extending Context	62
6.4	Summing up	62
7	Conclusion	64
7.1	Conclusion	64
7.2	Future Directions	65
	Bibliography	67

Chapter 1

Introduction

*“When I use a word,” Humpty Dumpty said, . . .
“it means just what I choose it to mean – neither more nor less.”
“The question is,” said Alice,
“whether you CAN make words mean so many different things.”*

—Lewis Carroll (1875)

Through The Looking-Glass: And What Alice Found There

Word sense discovery and disambiguation are the essence of communication in a natural language. Discovery corresponds to growing or acquiring a vocabulary. Disambiguation is the basis for understanding. These processes are also key components of language evolution and development. In this work we will restrict ourselves to the core processes of word sense discovery and disambiguation in text-based computer applications.

We will try to demonstrate that word sense discovery and disambiguation are two sides of the same coin: you cannot have one without first having the other. The resolution of this paradox requires some form of external reference. For humans the reference is provided by the world and the language community we live in. Since we are dealing with computer programs analyzing text, we will refer to written representations of language communities, i.e., text corpora and machine-readable dictionaries.

In the introduction we outline the processes involved in word sense discovery and disambiguation and briefly touch on some of the main problems common to both. We then outline the organization of the work and give an account of the author’s contributions.

1.1 Word Sense Disambiguation

Word sense disambiguation is the task of selecting the appropriate senses of a word in a given context. An excellent survey of the history of ideas used in word sense disambiguation is provided by Ide and Veronis (1998). Word sense disambiguation is an intermediate task which is necessary in order to accomplish some other natural language processing task, e.g.,

- translation selection in machine translation,
- eliminating irrelevant hits in information retrieval,
- analyzing the distribution of predefined categories in thematic analysis,
- part-of-speech tagging, prepositional phrase attachment and parsing space restriction in grammatical analysis,
- phonetization of words in speech synthesis and homophone discrimination in speech recognition, and
- spelling correction, case changes and lexical access in text processing.

Word sense disambiguation (WSD) involves the association of a given word in a text or discourse with a definition or meaning which is distinguishable from other meanings potentially attributable to that word. The task therefore necessarily involves two steps according to Ide and Veronis (1998). The *first* step is to determine all the different senses for every word relevant to the text or discourse under consideration, i.e., to choose a sense inventory, e.g., from the lists of senses in everyday dictionaries, from the synonyms in a thesaurus, or from the translations in a translation dictionary.

The *second* step involves a means to assign the appropriate sense to each occurrence of a word in context. All disambiguation work involves matching the context of an instance of the word to be disambiguated either with information from external knowledge sources or with contexts of previously disambiguated instances of the word. For both of these sources we need preprocessing or knowledge-extraction procedures representing the information as context features. For some disambiguation tasks, there are already well-known procedures such as morpho-syntactic disambiguation and therefore WSD has largely focused on distinguishing senses among homographs belonging to the same syntactic category.

However, it is useful to recognize that a *third* step is also involved: the computer needs to learn how to associate a word sense with a word in context using either machine learning or manual creation of rules or metrics.

It is the third step which is the focus of this work and especially the machine learning aspect. Unless the associations between word senses and context features are given explicitly in the form of rules by a human being, the computer will need to use machine learning techniques to infer the associations from some training material. In order to avoid confusion, we will speak of manually¹ created disambiguation techniques as a separate category and only divide the machine learning techniques into the subcategories of supervised, semi-supervised and unsupervised.

1.2 Word Sense Discovery

Word sense discovery is defined as the task of learning what senses a word may have in different contexts. Word sense discovery is what lexicographers do by profession. Automated word sense discovery on a large scale in order to build a thesaurus has a much shorter history. Some of the first attempts were made in the 1960s by Karen Spärck Jones (1986). As sufficiently large corpora and efficient computers have become available, several attempts to automate the process have been undertaken.

In lexicography, when building mono- and multilingual dictionaries as well as thesauruses and ontologies, word sense discovery is regarded as a preprocessing stage (Kilgarriff et al., 2004; Kilgarriff and Tugwell, 2001). In various applications, it is seen as a part of the lexical acquisition and adaptation process, e.g., in

- translation discovery when training statistical machine translation systems,
- synonym discovery for information retrieval,
- document clustering providing a domain analysis,
- detecting neologisms² or rare uses of words in part-of-speech tagging and grammatical analysis,
- discovering ontological relations³ for terminologies,

¹In the word sense disambiguation literature, notably SENSEVAL-2 (2001), manually created metrics or disambiguation rules are referred to as unsupervised. From a machine learning point of view, this is perhaps technically correct because no final automated training was used to improve the performance with a training corpus. According to the same reasoning, e.g., a manually designed wide-coverage parser would be an unsupervised method from a machine learning point of view.

²Basic lexical acquisition is done all the time in most natural language applications. Often it is simply dismissed as part of the preprocessing heuristics for neologisms, i.e., new words or out-of-vocabulary items.

³Ontological relations are: type and subtype (isa), part-of and whole, etc.

- named entity recognition, and
- automated discovery of morphology and syntax.

Word sense discovery involves the grouping of words by their contexts into labeled sets of related words. Also this task can be seen as consisting of three steps. The *first* step is to determine the groups of related words in context, i.e., create a context clustering. It involves calculating the similarity of the word contexts to be clustered, or to use similarity information from external knowledge sources.

The *second* step is to determine a suitable inventory of word sense labels. There is no well-established convention for labeling the context-clustered word groups. The predefined labels are typically taken from sense descriptors in everyday dictionaries, labels in thesauruses and ontologies, or translations in a translation dictionary. The labeling varies according to purpose: in terminology mining the ontological relations are frequently used, in thesaurus discovery thesaurus relations are often used, and in statistical machine translation the translations are suitable labels of word clusters.

In word sense discovery, the *third* step involves a way to learn how to associate a word sense label with a word cluster using either machine learning or manually created rules or metrics.

In word sense discovery it is mainly the first step which is the focus of our interest in this work.

1.3 Open problems

The broad problems in both the field of word sense disambiguation and that of word sense discovery are the sense divisions, the role of context and system evaluation. This stems from the fact that the meaning of words is an abstract notion, which we can observe only when someone produces an utterance in some context. This utterance in context then influences how we understand the meaning of the used words, which may influence our own use of the words in our next utterance and so on. The meanings of words as we can observe them are constantly evolving.

1.3.1 Word Senses

The meaning of a word can be divided into word senses in several ways. Although there is some psychological validity to the notion of sense, lexicographers are well aware of the lack of agreement on word senses and sense divisions. The sense distinctions made in many dictionaries are sometimes beyond those which human readers themselves are capable of making, consider senses number 1 and 5 of

interest in WordNet (Miller et al., 2003): (1) interest, involvement as “an interest in music”, and (5) pastime, interest, pursuit as “his main interest is gambling”. Combining dictionary senses does not solve the problem, because the degree of granularity is task dependent. In many cases, meaning is best considered as a continuum of shades of meaning. The points at which senses are combined or split can vary.

It is important to distinguish between word senses and word usages (contexts). Word senses are defined by lexicographers in dictionaries and word usages are what we observe. However, as Wittgenstein formulated it in his later work (Wittgenstein, 1953):

For a large class of cases—though not for all—in which we employ the word ‘meaning’ it can be defined thus: the meaning of a word is its use in the language.

In that light, this work is all about how to associate word senses with word usages.

1.3.2 Context

Context is the environment in which a word is used, and context, viz. word usage, provides the only information we have for figuring out the meaning of a new or a polysemous word. In a broader perspective, we can look at word context from three different aspects:

- the *modality* of the context, i.e., what we see, hear, feel, etc., and
- the *size* of the context: zero (i.e., the word itself), local (i.e., clause or phrase), and global (i.e., sentence or more),
- the *depth of processing*: morphological, syntactic, semantic, pragmatic, etc.

For this work we are mainly interested in the efforts to collect and preprocess written contexts of various sizes, but we need to keep in mind a broader view of context if some day we are to model real natural language acquisition and understanding.

1.3.3 Evaluation

In evaluation, two main approaches are generally used: *in vitro*, where systems are tested independently of an application using specially constructed benchmarks, and *in vivo*, where results are evaluated in terms of their overall contribution to the performance of a system. The *in vitro* evaluation makes it possible to focus separately on different aspects like the importance of word sense distribution,

context descriptions or algorithm differences with regard to word sense disambiguation and discovery. The *in vivo* evaluation can use existing test suites for various embedding systems.

The comparison of evaluation results has in the past been hampered by the lack of common resources, when researchers used different corpora and different sense inventories. The state of word sense disambiguation evaluation has improved with the SENSEVAL initiative (Kilgarriff and Palmer, 2000; Edmonds and Kilgarriff, 2002; Mihalcea and Edmonds, 2004), where a standardized set of test suites have been made available to the research community. For word sense discovery, the situation is still rather diverse, but the WordNet thesaurus is the most-cited common reference – in those more than 35 languages for which WordNet exists.

1.4 Duality of Disambiguation and Discovery

As we have seen in the previous sections of the introduction, word sense discovery and word sense disambiguation have many similarities. The assumption that words with similar usage have similar meaning was proposed by Harris (1954, 1968). In our work the assumption is studied from two aspects: firstly, similar word usages should give rise to similar meanings, and secondly, different meanings should manifest themselves in different word usages.

We define *word sense discovery* as determining which contexts are similar and therefore represent the same word sense, i.e., starting with a set of contexts for a word we cluster the contexts and try to label them with word senses. We define *word sense disambiguation* as determining how different word senses are associated with different contexts, i.e., starting with a set of word senses we try to identify and emphasize the essential differences in their context descriptions. Learning to disambiguate word senses can thus be regarded as the dual task of discovering word senses.

1.5 Organization of the Thesis

Sense divisions and related problems will be elaborated in Chapter 2 *Word Senses*. The nature of context is considered in detail in Chapter 3 *Context*. Basic issues related to system evaluation are presented in Chapter 4 *Evaluation*. An overview of ideas and systems for word sense disambiguation is presented in Chapter 5 *Disambiguation* and for word sense discovery in Chapter 6 *Discovery*. Conclusions and future work are outlined in Chapter 7 *Conclusion*.

1.6 Publications

The following is a short summary of the publications for the thesis and the author's contributions to each of them.

1.6.1 List of Publications

- 1 Lindén, Krister and Krista Lagus. 2002. Word Sense Disambiguation in Document Space. In *Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics*. Hammamet, Tunisia.
- 2 Lindén, Krister. 2003. Word Sense Disambiguation with THESSOM. In *Proceedings of the WSOM'03 – Intelligent Systems and Innovational Computing*. Kitakyushu, Japan.
- 3 Lindén, Krister. 2004. Evaluation of Linguistic Features for Word Sense Disambiguation with Self-Organized Document Maps. *Computers and the Humanities*, 38(4):417–435. Kluwer Academic Publishers.
- 4 Lindén, Krister and Jussi Piitulainen. 2004. Discovering Synonyms and Other Related Words. In *Proceedings of CompuTerm 2004, 3rd International Workshop on Computational Terminology*. Geneva, Switzerland.
- 5 Lindén, Krister. 2005. Modeling Cross-lingual Spelling Variants. *Information Retrieval*. Kluwer Academic Publishers. Accepted 2005.

1.6.2 Summaries and Contributions

- **Publication 1** The idea that explicit word sense disambiguation can be done in document space by labeling and reading a sufficiently large domain-independent context representation was proposed by the first author. The idea was initially tested via access to the patent document map (Kohonen et al., 2000) over the world-wide web, which provided the idea for how to label a map using explicit word senses in context. The results of my initial test on the patent map convinced Krista Lagus, the second author and supervisor, that the idea was worth pursuing. With her guidance and accurate advice I was able to plan the experiment and effectively carry out the large-scale testing. This was my first paper in the area of machine learning and

my co-author was in many ways instrumental in shaping my writing habits, for which I will be forever grateful.

The author's contribution was the idea to *label a very large WEBSOM map (Kohonen et al., 2000) using a set of prelabeled words-in-context independent of the map training material, developing a method for calculating the word sense label of a new instance*, as well as writing most of the article.

- **Publication 2** In the previous article we used a preexisting map and only the base forms of the words in the context. From a computational linguistic point of view, only using base forms seems like ignoring relevant linguistic information, e.g., syntactic information, which is known to highly influence the word sense disambiguation task (Lee and Ng, 2002; Yarowsky and Florian, 2002). There are also other linguistically motivated cues which improve the capacity to resolve word sense ambiguities.

The contribution in this article was a *first attempt at mathematically formulating the disambiguation method*, to use *separate WEBSOM maps for each word in context exploring the use of linguistic features in addition to the base forms*. In this article the method is named THESSOM, because it reads the map as a thesaurus-like structure. The word THESSOM also means “among these” in old Norse, which seems appropriate for an algorithm that chooses a sense from a given set of labels.

- **Publication 3** For the previous article the author tested linguistic features and made a number of experiments with the parameters for the WEBSOM map. The separate impacts of all of these changes had not been evaluated, although they all contributed to the performance improvement.

In this journal article the contribution is a *sensitivity analysis of the impact of the various linguistic features and WEBSOM parameters* as well as a *more elegant mathematical formulation of the THESSOM method*. The analysis shows that the impact of advanced linguistic features is more significant than changing the parameters for creating the maps.

- **Publication 4** The idea of this paper was to use the linguistic features which had been shown to be the most efficient for distinguishing word senses in context in the previous experiments in order to make explicit the synonym groups which are formed by clustering words according to their context, i.e., unsupervised learning of synonyms and other related words.

Our contributions in this article are four-fold. The first contribution is to *apply the information radius in a full dependency syntactic feature space* when calculating the similarities between words. Previously, only a restricted set

of dependency relations has been used with the information radius, whereas other similarity measures have been applied with the full feature set. The second contribution is a *similarity recalculation during clustering*, which we introduce as a fast approximation of high-dimensional feature space and study its effect on some standard clustering algorithms. The third contribution is a simple but efficient way to *evaluate the synonym content of clusters by using translation dictionaries for several languages*. Finally, we show that *69-79% of the words in the discovered clusters are useful* for thesaurus construction. The second author made the first contribution. The rest were the contributions of the first author including most of the writing of the article.

- **Publication 5** The idea of this journal article is that a term and its translation have similar meaning in similar cross-lingual contexts. Under certain conditions this context can be as narrow as the sound pattern of the term implying that we can look at probable transliterations in order to find an appropriate translation.

The first contribution of this article is to show that *a distance measure which explicitly accounts for the order of the letter or sound n-grams, significantly outperforms models based on unordered bags of n-grams*. The second contribution is to *efficiently implement an instance of the general edit distance* with weighted finite-state transducers using context sensitive transliterations. The costs for the edit distance are learned from a training sample of term pairs. The third contribution of this work is to demonstrate that the model needs *little or no adaptation for covering new language pairs* and that the *multilingual model is robust*, i.e., adding a new language does not adversely affect the performance of the model for the already trained languages.

Chapter 2

Word Senses

*How many angels can dance on the point of a very fine needle,
without jostling one another?*

— Isaac D’Israeli (1766-1848)

What is the meaning of a word? Unless one believes that we are born with an innate set of meanings waiting to find their corresponding expression in language, another option is that we learn the meaning of a word by observing how it is used by the language community we are born in. Some usages find their way into dictionaries and become established word senses.

In order to understand what constitutes a word sense, we can look at the criteria lexicographers use when they decide that a word usage is a word sense and record it in a dictionary for future generations. Finally, we will also describe the dictionaries and lexical resources that were used for the research in this work.

2.1 Language Philosophy

From a machine learning point of view Wittgenstein’s suggestion (Wittgenstein, 1953) that “*the meaning of a word is its use in the language*” sounds plausible, because there is nothing else for a machine to observe. This view of meaning was made more specific by Harris, when he proposed that words with similar syntactic usage have similar meaning (Harris, 1954, 1968).

Even if we accept that the *potential* usage of words is unlimited, we are mainly interested in *real* usage when we learn to identify similarities or differences of word meaning. The real usage is prone to fluctuations and idiosyncrasies, viz. usage preferences, of different language communities. A language community is any group of individuals who communicate. Some usage preferences become recognized by most communities of a language, a process known as lexicalization.

The lexicalization progresses differently in different communities of a language giving rise to, e.g., synonyms.

The usage preferences as they manifest themselves in real usages characterize similarity or difference of word meaning. If someone says “Shoot!” when a bear is attacking, it is emotionally quite different from the same command when a small bird is flying by, although both require some weaponry. However, a reporter can shoot a question without extra equipment. For most usages of a written word, we do not have access to the full usage context, so there may be essential differences in other aspects than those in the text presented to a computer. Indirectly, by observing other usages of words in the context, it may still be possible for a computer to group the usages of *shoot* in ‘shoot a bear’, ‘shoot a bird’, and ‘shoot a question’ into two main groups of shooting with and without weapons. Then we present the machine with ‘shoot a bullet’ and expect the *bullet* to be more like a *question* than a *bear*, because in fact the main division does not depend on the presumed weapon, but whether the object of *shoot* is animate or inanimate. We call this distinction a semantic feature. A multiple-inheritance taxonomy of such features is a feature structure. The animate and inanimate distinction is not fixed for every word, but may lend itself to modification or underspecification as in ‘shooting stars’. A machine making observations based on a limited number of samples of the real usage of a word in written text will end up with a piecewise approximation of features such as animate and inanimate.

2.2 Enumeration vs. Generation

The simplest way to create a dictionary of word senses is to enumerate each sense separately. If no further information is provided about how the senses are related, this representation requires each new sense to be manually added. A more flexible representation is presented by Pustejovsky (1998), a generative lexicon (GL), where the word senses are generated through the unification of feature structures guided by an inheritance system for the argument, event and qualia structures.

The GL is sometimes seen as a fundamentally different approach from the idea of dictionaries or lexicons as a simple enumeration of word senses, because the theory on generative lexicons claims that the GL also accounts for novel uses of words. Kilgarriff (2001) tested this claim on a set of corpus words and found that most of the novel or non-standard usages were unlikely to be accounted for by any GL, i.e., those usages that were not accounted for in a regular dictionary. The main benefit of a large-scale dictionary based on the GL theory would be that similar distinctions would consistently be made throughout the dictionary for all words with similar or related usages.

From a computer programming point of view, it is not particularly surprising

that a lexicon program, i.e., a GL, is more flexible than a list of word descriptions, more consistent and more compact, but equally unimaginative. In addition, as the GL grows, it is likely to be more unpredictable and more difficult to maintain. A GL comes with all the benefits and drawbacks of a large computer program and as such it covers only the words and senses it has been either intentionally or unintentionally programmed to cover.

2.3 The Origin of Features

A more fundamental problem related to language learning and child language acquisition is how we learn to associate meaning with sound sequences or words. We do not get closer to a solution for this problem by dividing a word into semantic features, because then we have to ask where the features come from or how they become primitives of the lexicon.

Interesting research on how meaning is associated with sound sequences has been done by Kaplan (2001) in his simulation of a robot society communicating about positions of several colored figures, i.e., circles, triangles and squares, on a white board using a Wittgensteinian language game. He was able to demonstrate that, when several stable language communities had evolved, synonymy arose. When the communities were in sporadic interaction, the communities kept their own words for the concepts but were able to understand other variants. By inspecting the robots he could determine that they had words for colors, shapes and relative positions. The robot simulations indicate that with suitable and not too complicated models, language can be learned from scratch in a language community interacting with the external world.

Research by (one of Harris' students) Gleitman (1990, 2002) and Gleitman et al. (2005) on child language acquisition indicates that children learn nouns with external references before they learn verbs and then start distinguishing between different argument structures of the verbs. Her research supports the assumption that the meaning of verbs is tightly tied to their argument structure. The child language research gives some psychological relevance to the GL approach indicating that a GL is not merely a way of compressing the lexicon description.

If we accept that features and the meaning of features can be induced through language usage in a language community, a full-scale GL for some application would be an interesting effort both as a collection of linguistic knowledge and as a benchmark for future automatically induced vocabularies. It is quite likely that for some time to come high-performing computational lexicons will be partly hand-made with a generative component and a trainable preference mechanism¹. A well-designed linguistically motivated GL with a trainable preference learning

¹On a parallel note, we quote Kohonen's personal comment on his self-organizing maps: "Once

mechanism might be a good candidate for how to organize a word sense lexicon. There is no need for a computer to always learn the lexicon from scratch, despite the fact that this seems to be the way nature does it.

2.4 Recording Word Senses

New words and concepts arise at a steady pace and old words become associated with new meanings especially in technology and biotechnology which are currently the focus of intense research efforts. In these areas, specialized efforts like named entity recognition aim at identifying the meaning of new terms in the form of abbreviations, nouns and compound nouns by looking at their context. These entities are typically classified into semantic types like names, dates, places, organizations, etc. Named entities and word senses represent two extremes of the same problem². Named entities are usually new previously unseen items that acquire their first word sense, whereas word sense discovery and disambiguation typically have assumed that words have at least two word senses in order to be interesting. It is, however, likely that the mechanism or process that attaches the first word sense to a string is the same as the one that later attaches additional meanings or word senses to the same string either by coincidence, i.e., homonymy, or by modifying some existing meaning, i.e., polysemy.

Another aspect of word senses is when a word gets different translations (Resnik and Yarowsky, 2000) and the sense identification problem is restricted to finding the appropriate translation in context. The translation analogy can be taken further, because finding the first word sense is in some ways equivalent to finding the first translation, which is especially important for cross-lingual information retrieval in the same areas where named entity recognition is important. A method which significantly outperforms previously known comparable methods for finding translations of named entities in a cross-lingual setting has been proposed by Lindén (2004) and is more fully elaborated in (Publication 5).

Automatically identifying a word's senses has been a goal since the early days of computational linguistics, but is not one where there has been resounding success. An overview of methods that have used artificial intelligence, machine-readable dictionaries (knowledge-based methods) or corpora (knowledge-poor methods) can be found in Ide and Veronis (1998) and Grefenstette (1994). An

it has been shown that a map always organizes regardless of how random the initial state is, there is no need to show this every time. It is quite acceptable to speed things up by starting from an educated guess."

²Other work on this theme distinguishes different semantic roles and build semantic frames (Gildea and Jurasfky, 2002), which can be seen as a practical step in between named entity recognition and full word sense disambiguation.

excellent bibliography of the research related to word sense research is provided by Rapaport (2005).

Kilgarriff (1997) suggests that the lack of success in word sense disambiguation may be unclarity as to what a word sense is. A word might not have been seen in a context because it is not acceptable there, or it might not have been seen there simply because the corpus was not large enough (Kilgarriff, 2003c). In the following, we will first look at the frequency aspect and then at the acceptability aspect.

2.4.1 Frequency Distribution

Where a lexicographer is confronted with a large quantity of corpus data for a word, then, even if all of the examples are in the same area of meaning, it becomes tempting to allocate the word more column inches and more meanings, the lexicographer Kilgarriff admits (Kilgarriff, 2004) and considers the words *generous* and *pike* as examples:

Generous is a common word with meanings ranging from generous people (who give lots of money) to generous helpings (large) to generous dispositions (inclinations to be kind and helpful). There are no sharp edges between the meanings, and they vary across a range. Given the frequency of the word, it seems appropriate to allocate more than one meaning, as do all of the range of dictionaries inspected. *Pike* is less common (190 BNC occurrences, as against 1144) but it must be assigned distinct meanings for fish and weapon (and possibly also for Northern English hill, and turnpike, depending on dictionary size), however rare any of these meanings might be, since they cannot be assimilated as minor variants. Pike-style polysemy, with unassimilable meanings, is the kind that is modeled in this paper. Where there is generous-style ambiguity, one might expect less skewed distributions, since the lexicographer will only create a distinct sense for the 'generous disposition' reading if it is fairly common; if the lexicographer encounters only one or two instances, they will not. Polysemy and frequency are entangled.

In the same article, Kilgarriff (2004) observes that the dominance of the most common sense increases with n , the frequency of the word. In additional corpus data, we find additional senses for words. Since a majority of the words are monosemous³, finding additional senses for them dominates the statistic. On the

³E.g. WordNet, an online lexical reference system, contains approximately 126 000 monosemous and 26 000 polysemous words (Miller et al., 2003), cf. Section 2.6.1.

average, the proportion of the dominant sense therefore increases with n simply because the proportion of the first sense, $(n - 1)/n$, compared with that of the additional sense, $1/n$, increases with n . He proceeds to demonstrate that the distribution of word senses roughly follows a Zipfian power-law similar to the well-known type/token distribution (Baayen, 2001; Zipf, 1935). Kilgarriff uses the sense-tagged SemCor database (Mihalcea, 2004b) for empirical figures on the proportion of the most common sense for words at various frequencies, and compares the empirical figures with the figures his model predicts when initialized with the word frequency distribution from the British National Corpus (BNC) (Burnard, 1995). The fit between the SemCor and the predicted figures makes it believable that word frequencies and word sense frequencies have roughly similar distributions and that we can expect the skew to become more pronounced for higher values of n .

The conclusions we can draw from Kilgarriff (2004) are that a large-scale domain-independent word sense disambiguation system, which always chooses the most common sense out of two or more senses, will over time perform accurately in 66–77 % of the ambiguous cases based on the weighted average of the SemCor figures, or even in 66–86 % of the cases according to the figures predicted by the larger BNC corpus model. For high-frequency words, the ambition of a lexicographer to account for all the source material rather than for all the senses is a partial explanation for why some word senses are difficult to disambiguate even for humans.⁴ If such senses were disregarded, the higher predicted proportions of the dominant sense may in fact be more valid for the high-frequency words. Another implication of the Zipfian distribution is that over time all words are likely to appear in most contexts with a very low probability, and in practice most word senses will never have been seen more than once in any specific context.

2.4.2 Acceptability in Context

As soon as we start limiting acceptability of words in certain contexts, we begin losing creative language use. One possibility is to relate the contents of a sentence to the world we live in, in order to estimate the plausibility of the sentence. However, this will complicate matters, because we then also have to model the plausibility of events in the world. An approximation of how objects and events of the world relate to one another is provided by an ontology. Unfortunately, there is yet no world-wide ontology around, but we have fairly large thesauruses.

The difference between a thesaurus and an ontology is that the former deals

⁴As Atkins (1991) points out, however, lexicographers are only allotted a limited amount of space for any word in a dictionary counterbalancing the urge to create very fine-grained sense distinctions.

with words and their relations observable in language use and the latter deals with objects and their relations in the world we live in. To highlight the distinction, we can consider the famous quote “Colorless green ideas sleep furiously” by Chomsky (1957). From a purely language use perspective this full sentence is unexpectedly likely occurring more than 5700 times on the world-wide web. It is so common that it can be regarded as idiomatic. From an ontological perspective, the fact that it has been repeated into idiomhood by the world’s linguists does not make its content more plausible. Compositionally it still means little, but contextually it is a very pregnant construction. However, people tend to speak and write more often about things they have or would like to have experienced than they spend time producing and repeating random sequences of words, so the natural language we can observe, e.g., on the web, is a noisy reflection of the relations between objects in the world. As a consequence, the difference is not so wide between a thesaurus constructed from observations of language use and an ontology constructed from observations of the world.

A bigger practical problem is that thesauruses usually do not contain well-defined word senses that we could use for plausibility judgments. In an effort to clarify the relation between words and their multiple meanings Kilgarriff (2003b) tries to explain why thesauruses do not really contain word senses. The first priority of authors of thesauruses is to give coherent meaning-clusters, which results in quite different analyses from those in dictionaries, where the first priority is to give a coherent analysis of a word in its different senses (Kilgarriff and Yallop, 2000). From a practical point of view, if we wish to use a thesaurus for a natural language processing (NLP) task, then, if we view the thesaurus as a classification of word senses, we have introduced a large measure of hard-to-resolve ambiguity to our task (Kilgarriff, 2003b). For this reason Kilgarriff claims that, even though Roget may have considered his thesaurus (Roget, 1987) a simple taxonomy of senses, it is better viewed as a multiple-inheritance taxonomy of words.

The direct consequence of Kilgarriff’s argument is that a thesaurus is perhaps useful as a backbone for a generative lexicon, but as such the words in a thesaurus are ambiguous. Kilgarriff’s argument is easier to understand if we keep in mind that the meaning of a word is defined by the contexts in which it occurs. The real problem is that a meaning-cluster in a thesaurus seldom includes the common contexts in which the words of the meaning-cluster occur. So what can we use a thesaurus for? Systems which try to discover word senses, also classify words based on their context into maximally coherent meaning-clusters, i.e., thesauruses can serve as test beds for automatic word sense discovery systems. The somber consequence of Kilgarriff’s argument is that for NLP systems the words in a meaning-cluster are in fact an epiphenomenon⁵. The valuable part is the con-

⁵This is not to say that word sense and thesaurus discovery efforts are futile. Word lists are

text description by which the words were grouped. The context description is a compact definition of the meaning of the word cluster and this is the part that is usually made explicit in a regular dictionary analyzing the senses of a word. It is the context description that can be used for determining the acceptability of the word sense in various contexts.

2.5 Word Sense Dictionary Specification

If we use a generative lexicon to determine the acceptability of a word sense in context and the lexicon provides hard constraints, we will end up not covering creative language use after all. We could, however, account for creative language use by basing plausibility judgments⁶ on observable language. Ideally, a lexicon provides structure and soft constraints based on context descriptions giving more plausibility to more likely objects and events.

To summarize the discussion of the previous sections, we can set up a general wish list of what a context description of a word sense in an ideal lexicon should contain, loosely based on the idea of a generative lexicon (Pustejovsky, 1998): *part of speech* categories, *argument structure* of arguments and adjuncts, *event structure* for the argument structure, *qualia structure* describing an object, its parts, the purpose and the origin of the object, *interlexical relations*, e.g., synonymy, antonymy, hyponymy lexical inheritance, entailment, translation, *plausibility estimate* by providing all of the above with frequency or probability information⁷

An example of the plausibility information that the lexical model needs to incorporate is given by Lapata and Brew (2004), where they highlight the importance of a good prior for lexical semantic tagging. They use verb classes based on Levin (1993) and obtain their priors directly from subcategorization evidence in a parsed but semantically untagged corpus.

Another example is the prevalence ranking for word senses according to domain, which should be included in the generative lexical look-up procedure. The sense distributions of many words depend on the domain. Giving low probability to senses that are rare in a specific domain permits a generic resource such as

primarily intended for consumption by systems that are capable of filling in the appropriate context descriptions themselves, e.g., human beings. A central issue in information retrieval (IR) research is to devise strategies which cope with missing context. This may partially explain why IR often seems to have more to offer thesaurus makers than the other way around, see Sanderson (2000).

⁶A plausibility judgment is at least a weak partial ordering of the relative plausibility of statements.

⁷From a Bayesian statistics point of view we would have prior linguistic information combined with the information provided by corpus data expressed in a probability distribution over the feature structures that define the linguistic relations.

WordNet to be tailored to the domain. McCarthy et al. (2004b) present a method which calculates such prior distributions over word senses from parsed but semantically untagged corpora.

2.6 Existing Sense Inventories

For the research conducted in this work we needed existing lexical resources as evaluation material or for the purpose of creating evaluation material. For the English experiments there was WordNet (Fellbaum, 1998)⁸ and the ongoing effort to provide samples for the sense inventory of WordNet by tagging corpora for the SENSEVAL (Mihalcea, 2004a) evaluation tasks. Even though WordNet has been implemented in many languages (Vossen, 2001; Vossen and Fellbaum, 2004), no WordNet exists for Finnish, so for Finnish we needed to invent a procedure for approximating a thesaurus. For the cross-lingual experiments we used medical terminology available on the web.

2.6.1 WordNet

WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. Now version 2.0 of WordNet is available (Miller et al., 2003).

WordNet currently contains approximately 152 000 unique strings (nouns 115 000, verbs 11 000, adjectives 21 000, and adverbs 5 000) divided into 115 424 synsets with approximately 203 000 word sense pairs⁹. There are approximately 126 000 monosemous words with as many word senses, and 26 000 polysemous words with 78 000 word senses (Miller et al., 2003).

WordNet has links between derivationally and semantically related noun/verb pairs. A topical organization is also included. Related word senses, such as military terms, are assigned a domain. When disambiguating the terms in the synset glosses, links will be inserted that indicate the context-appropriate WordNet sense of each open-class term in the definitional gloss. The goal for WordNet is to further increase the connectivity in the three areas mentioned above, i.e., derivational

⁸Other existing resources for English are Roget's Thesaurus (Roget, 1987) and the Macquarie Dictionary (Delbridge et al., 1987) and the General Inquirer (Stone et al., 1966), whose merits have been explored by Inkpen and Hirst (2005)

⁹The total of all unique noun, verb, adjective, and adverb strings is actually 144309, i.e., many strings are unique within a syntactic category, but can be found in more than one syntactic category.

Language	Target word	Back translation
<i>English</i>	deficit	<i>vaje, vajaus, alijäämä; tilivajaus</i>
	shortfall	<i>vaje, alijäämä</i>
<i>German</i>	Defizit	<i>vajaus, vaje, alijäämä; kassavajaus, tappio; tilivajaus; puutos, puute</i>
	Unterbilanz	<i>alijäämä, vajaus, vaje, kauppavaje</i>
	Fehlbetrag	<i>vajaus, alijäämä, tappio, virhemaksu</i>
<i>French</i>	déficit	<i>alijäämä, miinus, tilivajaus; vajaus, vaje; tappio</i>

Table 2.1: Translations of the Finnish source word *alijäämä* into English, German and French with the back translations into Finnish. The shared back translations *vaje, vajaus, alijäämä, tilivajaus* are highlighted.

morphology, topical clustering, and disambiguation of terms in glosses (Miller et al., 2003).

2.6.2 Finnish Synonym Dictionary

There is no wide-coverage synonym dictionary online publicly available for Finnish. In order to mechanically create an approximation for a synonym dictionary in Finnish, we recall that synonyms are words that mean roughly the same thing. We note that when translating a word from a source language the meaning of the word is rendered in a target language. Such meaning preserving relations are available in translation dictionaries. If we translate into the target language and back, we end up, inter alia, with the synonyms of the original source language word. In addition, we may also get some spurious words that are related to other meanings of the target language words. If we assume that the other words represent spurious cases of polysemy and homonymy in the target language, we can reduce the impact of these spurious words by considering several target languages and for each source word we use only the back-translated source words that are common to all the target languages (Publication 4). We call such a group of words a source word synonym set. For an example, cf. Table 2.1.

We extracted translations for a sample of approximately 1800 Finnish words in the Finnish-English, Finnish-German and Finnish-French MOT dictionaries (Kielikone, 2004) available in electronic form. We then translated each target language word back into Finnish using the same resources. The dictionaries are based on extensive hand-made dictionaries. The choice of words may be slightly different in each of them, which means that the words in common for all the dictionaries after the back translation tend to be only the core synonyms.

2.6.3 Cross-lingual Terminology

As a cross-lingual terminology resource we used a technical medical terminology in eight different languages. We chose English as the target language. The terminology was extracted from the web pages of the EU project *Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages* by Stichele (1995). Only eight languages were available on the web server: Danish, Dutch, English, French, German, Italian, Portuguese and Spanish. We collected 1617 words which had at least one translation in all eight languages.

Based on the 1617 English terms we created the corresponding data for Finnish (Publication 5) by consulting several online resources. The most significant resource was the online medical language database *Tohtori.fi – Lääkärikirja* (Nienstedt, 2003). We found Finnish equivalents for 1480 of the medical terms.

For the cross-lingual task we also had a separate test set of 271 English terms with translations into six languages: Finnish, French, German, Italian, Spanish and Swedish. The terms did not occur in a standard translation dictionary. The terms can be grouped into domains. The number of terms in English in each domain is indicated in parenthesis: medicine or biology (90), geographical place names (31), economics (55), technology (36) and others (59). The test set had been created at the University of Tampere (Keskustalo et al., 2003).

2.7 Summing up

In text we can observe word forms which through morphological analysis get a base form. A base form may have several meanings that are generally understood by a language community. Some *base form–meaning* pairs, i.e., word senses, are recorded in dictionary definitions and glosses by lexicographers, but for many word senses we only have indirect evidence through word usage.

In this chapter, we have briefly described the criteria lexicographers use when they decide what word usage constitutes a word sense. The fact that the bulk of all language use is a reflection of the world we live in, makes some word senses of a word dominant. Most previously unseen word usages are creative simply because they are unexpected or surprising. A natural language processing (NLP) system needs to recognize that a usage is unexpected. However, the context in which the usage appears is what the word means and should be recorded for future reference, e.g., telephones used to be stationary until the advent of mobile phones, so a sentence like “He walked down the street talking on the phone” was implausible 30 years ago, but is now highly likely and the walking-talking context has become part of the meaning of the word *telephone*. Finally, we also described the lexical resources that are used as sense inventories for our research.

We have argued that word meaning is not discrete. However, the meaning of words is quantized into word senses in dictionaries. If we need a common world view, we can refer to a sense inventory of an agreed upon dictionary, otherwise we may as well compare word contexts directly. Next we turn to the context and different dimensions of context.

Chapter 3

Context

R2D2 beeps: -

C-3PO: I beg your pardon, but what do you mean, "naked?"

R2D2 beeps: -

C-3PO: My parts are showing? Oh, my goodness, oh!

— Star Wars: Episode I – The Phantom Menace (1999)

If we leave out the original source of a message, i.e., the speaker, and study only the message and the environment it was produced in, we no longer have access to any underlying mental or even metaphysical representation of the message. As human beings it is possible to believe and argue that we have some innate capacity which provides this access. In this thesis we study machine learning, and it should be easy to accept that the only way for a machine to learn the meaning of a specific word in a message, is by studying co-occurring words in one or several messages and whatever recording we have of the environment in which these messages were produced, i.e., the context.

First we present different aspects of context, then we describe some context collections, and finally we specify the context descriptions used in this work and their implementations.

3.1 Dimensions of Context

The context of a word is everything that occurs before, during and after a word is uttered including the word itself. Since we cannot take everything into consideration, it is useful to regard previous and future events as lying along a continuum and to consider the role and importance of contextual information as a function of distance from the target word. Some explicit attempts at defining the importance

of the context as a function of the distance have been made (Yarowsky and Florian, 2002), but the most common approach is to define the importance in terms of linguistic units surrounding the word, e.g., phrase, clause, sentence, discourse, topic, domain.

The context has at least three different dimensions whose impact on word sense discovery and disambiguation are seldom investigated independently: *context size*, *modality of context*, and *depth of context processing*. The easiest one to investigate is the context size. For word senses the size and shape of the context has been found to influence different word categories differently, e.g., nouns often benefit more from a longer context than verbs which often depend on a local context (Agirre and Martinez, 2001; Martinez and Agirre, 2000; Ide and Veronis, 1998). In this work we have focused on evaluating the impact of the depth of linguistic preprocessing (Publication 3).

3.1.1 Context Size

When looking at the size of the context, it is useful to divide it into three separate size categories:

- zero context – the word itself
- local context – phrase and clause
- global context – sentence, discourse, topic and domain

The size categories are assumed to be independent of the modalities and the depth of processing in each of the modalities. The zero context is sometimes disregarded, but it contains important information about the internal structure of the word, e.g., capitalization, sounds or morphs, which often relate the word to other words with similar meaning. The local context is sometimes defined as a narrow window of 3 – 5 words centered on the word itself. A narrow window is an approximation of the linguistic concepts, such as phrase and clause, which may not be available without linguistic software. The global context is defined as a window of 25 – 1000 words centered on the word itself, which fairly well approximates contexts starting from the immediately surrounding sentence to the whole document or domain.

It is sometimes useful to define syntactic context separately as the context provided by the heads and immediate constituents, because the syntactic context may carry information crossing the borders of all three of the context sizes. The syntactic relations which are most important for characterizing the word sense of a word can often be found within the same clause boundaries as the word itself, so in practice they can be regarded as part of the local context.

We wish to point out that the division of context into different sizes is purely a tool for linguistically motivating the experiments, where we extracted different sizes of contexts in order to show their relative influence on word sense disambiguation. One could argue that there are no distinct boundaries between local and global context. There are only more or less influential context features, whose general tendency is that their influence diminishes with increasing distance from the word itself.

3.1.2 Context Modality

Human language technology systems have typically focused on the “factual” aspect of content analysis. Other aspects, including pragmatics, point of view, and style, have received much less attention. However, to achieve an adequate understanding of a text, these aspects cannot be ignored. (Qu et al., 2004)

The two primary modalities for perceiving language are hearing (audition) and seeing (vision). In addition to these two, we learn the meaning of words in context by taste (gustation)¹, smell (olfaction) and several types of physical feeling (tactition, thermoception, nociception, equilibrioception and proprioception²). The main categories make up the five Aristotelian senses. Some animals have at least three more: electroreception, magnetoception and echolocation, which to humans may seem as instances of the “sixth” sense. Two of these additional senses, i.e., electroreception and magnetoception are used by psychologists in the form of EEG (Electro Encephalo Gram) and MRI (Magnetic Resonance Imaging) to study brain activity.

As humans we grow up learning to interpret written information in relation to our senses. If the words in a machine-readable dictionary could be given appropriate and relevant initial values for all of these senses, it would perhaps be possible for a computer to make better generalizations of basic semantic features such as animate and inanimate. In the future, robots may be conditioned on this kind of life experience. In the meantime, we may have to encode some of the basic lexical semantic features separately for each word in order to bolster computer inferences based on natural language text input. The concept hierarchies of machine-readable thesauruses and ontologies are potential sources for such world-knowledge. Even if WordNet (Fellbaum, 1998) does not give explicit lexical domain relations like “sports” for the words “racket” and “court”, the information can be extracted with some additional processing. Similarly, other more fundamental lexical semantic features could be precomputed by inference. Another similar source on a larger

¹E.g., the vocabulary of the smell or taste of wine: nutty, sturdy, foxy, etc.

²These are the sensations of touch, warmth, pain, balance and muscle movement, which are needed for understanding the terminology of dancing and many other physical exercises.

scale is the Internet, especially if the effort to create a Semantic Web is successful, cf. Section 5.1.2.

3.1.3 Depth of Preprocessing

As a separate dimension of context we have the depth of preprocessing. There are two radically different ideological points of view for context preprocessing. Some aim at encoding underlying principles of language and communication and expect language models to emerge as a by-product of applying such principles to a corpus. If this is the goal, it is useful to get by with as little preprocessing as possible relying mainly on machine learning and statistical language models. Some aim at drawing on existing linguistic knowledge and use as much preprocessing as is efficiently available and necessary to solve new problems. We will now outline the existing levels of computational linguistic preprocessing.

Looking at the surface form of a word, we find that the original capitalization of a word form is an annotation entered by the author of a document. If the word forms are normalized so that capital letters are turned into lower case, at least one prominent distinction is lost between, e.g., *Church* as an institution and *church* as a building.

Traditionally, a base form is the form found in a dictionary. Some word forms may have several base forms depending on context. In English the base form is often unique. The ambiguity is mainly between parts-of-speech with the same base form. One notable exception is the analysis of participles, e.g., “a *drunk* driver/has *drunk* a lot” with base forms *drunk/drink* or “was *heading* south/the newspaper *heading* is” with base forms *head/heading* etc. The correct base form can be determined in context as a side-effect of part-of-speech tagging.

An intermediate level before full dependency parsing is head syntax, which only indicates in which direction the head word is and what part of speech the head word is. The main advantage of head syntax is that it avoids attachment ambiguities, e.g., in “the man on the hill *with* the telescope” the preposition *with* is tagged as a dependent of some noun to the left, e.g., *with* $N<$. Full dependency syntax builds parse trees with one head word for each word. Each head word may have several dependents. For a rule-based approach to dependency syntax, see Tapanainen and Järvinen (1997), and for a statistical approach, see Samuelsson (2000).

It is now widely accepted that high-quality morphological analysis by TWOL (Two-Level Model) (Koskenniemi, 1983; Voutilainen et al., 1995), part-of-speech tagging by CG (Constraint Grammar) (Karlsson, 1990; Voutilainen, 1995) and syntactic parsing by FDG (Functional Dependency Grammar) (Voutilainen, 1997; Tapanainen and Järvinen, 1997; Connexor, 2002a,b) of written text can be achieved in a fairly domain and topic independent way by computer programs

(Lindén, 1993, 1997). The knowledge of natural language grammar encoded in these programs is the potential syntactic interaction between words and groups of words.

An alternative route would be feature structure-based context descriptions, e.g., using unification, which were popular in the beginning of the 1990s (Carlson and Lindén, 1987; Gazdar and Mellish, 1989) and which was adopted in the LFG (Kaplan and Bresnan, 1982) and HPSG (Pollard and Sag, 1994) frameworks. However, the wide-coverage FDG parsers (Connexor, 2002b) are now also available with attribute-value feature structure representation of the output. The attribute-value grammars may at the time have seemed too static for capturing usage preferences, such as connotations, and extended usages, such as metaphors. There should at least have been a mechanism in them for learning usage preferences based on observations, but it was not until Abney (1997) that the attribute-value graphs received a proper stochastic framework with an algorithm to estimate the parameters. The algorithm was rather heavy to compute.

Recently interesting attempts at creating morphological (Creutz and Lagus, 2004) and syntactic (Klein, 2005) parsers using unsupervised learning have been made. The idea is to incorporate general language independent principles in a natural language morphology or syntax discovery algorithm and then try to find a grammar and a lexicon that embody the training material as succinctly as possible. It is likely that the quality of the output will improve as the encoded discovery principles become more specific, even if the goal is to find as general principles as possible.

We are aiming at semantic similarity defined as substitutability in context, so we need to study how far we can get with a piecewise representation of the context and the linguistic structures in the context.³

As a summary, computational linguistics currently offers the following high-quality domain-independent preprocessing:

- token analysis, i.e., word forms are separated from punctuation marks and fixed expressions are identified. In some languages this phase may also include identifying potential word boundaries.
- morphological analysis, i.e., dictionary look-up of inflectional tags for word forms. Part-of-speech tagging is morpho-syntactic analysis disambiguating

³Named entity recognition (NER) provides general semantic typing of certain items in text. NER can be seen as a practical preprocessing step providing additional features, i.e., piecewise representations of linguistic structures, for word sense disambiguation or word sense discovery. The same thing can also be said about tagging semantic roles in text (Gildea and Jurafsky, 2002) or term identification for translation. Such methods are either part of the solution or the problem depending on what granularity and type of word sense discovery or disambiguation we are aiming at.

inflectional tags.

- syntactic analysis, i.e., immediate constituents and a syntactic head-word are identified for each word in a clause. The nature of the identification may vary from a simple indication of the direction of the head-word to a precise co-indexing, which can serve to build a tree structure or a dependency graph.

3.2 Related Work

As pointed out by Ide and Veronis (1998), information from the microcontext, topical context and domain contributes to the sense selection, but their interrelations are not well understood. The microcontext generally means a context of a few words up to an entire sentence. It was an early finding by Kaplan (1950) that ± 2 word contexts are highly reliable and even ± 1 contexts are reliable in 8 out of 10 cases. In the microcontext (Leacock et al., 1998; Martinez and Agirre, 2000; Pedersen, 2001; Agirre and Martinez, 2001; Lee and Ng, 2002; Yarowsky and Florian, 2002) it is also recognized that the distance to the keyword, the collocations as well as the syntactic relations are significant for local word sense disambiguation.

Topical context means a window of several sentences and has been discussed in information retrieval for years (Salton, 1968). Local context can account for most of the ambiguities, but topical context can improve the result (Gale et al., 1993; Yarowsky, 1995; Voorhees et al., 1995; Towell and Voorhees, 1998). Although a distinction is made between microcontext and topical context in current WSD, it is not clear whether this distinction is meaningful. It may be more useful to regard the two as lying along a continuum and to consider the role and importance of contextual information as a function of distance from the target (Ide and Veronis, 1998).

A lexical domain is the words commonly used in a specific domain, e.g., the medical or legal domain. The lexical domain is often represented by a domain-specific glossary called a microglossary. A microglossary reduces ambiguity, but it does not eliminate it. The claim that there is in general only one sense per discourse (Gale et al., 1992b) can be disputed, as the influence of the domain depends on the relation among the senses like common vs. specialized usage. The claim was made a bit stronger by Yarowsky (1995), when he noted that there seems to be only one sense per collocation and that words tend to keep the same sense during a discourse. Leacock et al. (1998) pointed out that some words have senses which may occur in almost any discourse and Ide and Veronis (1998) mention that in the Encyclopaedia Universalis the French word *intérêt* occurs 62 times only in its financial sense in an article on interest on money, 139 times only in its human sense

in an article on philosophy, and 2 times for each sense in an article on the third world. Despite such counterexamples, the tendency of one sense per discourse was slightly refined to one sense per topic and text genre by Martinez and Agirre (2000). Magnini et al. (2002) manually grouped the word senses for WordNet belonging to the same domain and were able to show that one domain per discourse is a better prediction than one sense per discourse.

In terms of our previously defined context dimensions, the microcontext can be characterized as the written local context with syntactic preprocessing. The topical context can be characterized as the written global context with morphological tagging (often discarding other word classes than nouns). Domain is different, even if it is a derivative of the global written context. A domain is characterized by the words and the word senses that tend to occur in it, i.e., a domain is an aggregate over all the contexts of a word sense when this aggregate is contrasted with similar aggregates for other word senses.

3.3 Collecting Contexts

We now have an idea of what contexts could be useful and what contexts typically have been used. For word sense discovery we can collect unlabeled contexts, but for word sense disambiguation we need labeled or annotated contexts. Next we provide a brief overview of the corpus collection and annotation efforts.

3.3.1 Corpus Collection

The first words in context were collected by lexicographers on index cards. In the early 1960s the one-million word, machine-readable Brown corpus opened a new way of thinking in linguistics. However, soon more data was needed and in the late 1970s Sinclair and Atkins started the COBUILD project, which raised the expectations on corpus size from one million to ten million words (Sinclair, 1987). Ten years later, Atkins again initiated the development of the British National Corpus (Burnard, 1995), which raised expectations for sizable corpora to 100 M words. The BNC corpus even contained 10 % transcribed spoken English, but at the time no one realized that the original sound files could also be of general interest, so for them no distribution rights were acquired (Burnard, 2002). At the end of the 1990s the BNC World Edition was finally released making the corpus widely available at low cost covering a wide variety of British English from the mid-1990s.

However, for some purposes the BNC is not large enough. This is an outcome of the Zipfian nature of word frequencies. While 100 M words is a vast number, and the BNC contains ample information on the dominant meanings and

usage-patterns for the 10,000 words that make up the core of English, the bulk of the lexical stock occurs less than 50 times in it, which is not enough to draw statistically stable conclusions about a word. For rarer words, rare meanings of common words, and combinations of words, we frequently find no evidence at all. (Kilgarriff, 2003a)

The development of corpus size has closely followed the development of computer capacity which has become orders of magnitude larger, faster and cheaper. However, in the last decade, the greatest development in computing has been the web. According to a conservative estimate, it contains 100,000 M words of English as well as lesser, but still very large, quantities of many other languages (Grefenstette and Nioche, 2000). This collection of text may not always be the clean and tidy corpus we are used to, but is a corpus nonetheless (Kilgarriff and Grefenstette, 2003).

3.3.2 Annotated Corpora

Unfortunately, the speed of semantically annotating corpora has not kept up with the corpus collection development, because such corpora need to be manually annotated. One of the first large sense-annotated corpora was the DSO corpus by Ng and Lee (1996), which was based on the Brown corpus and the Wall Street Journal. Its most frequent nouns and verbs were annotated with WordNet 1.5 senses. Another large annotated corpus is SemCor (Fellbaum, 1998) which is based mainly on the Brown corpus. Some of its texts had all open words semantically annotated with WordNet senses, but some texts had only the verbs annotated. Ng and Lee (1996) achieved only 57 % interannotator agreement, when they compared the overlapping portions of the DSO and SemCor corpus. This made it impossible to regard these taggings as a gold standard (Kilgarriff, 1998) for word sense disambiguation. What would be the interpretation if an automatic disambiguation procedure achieved more than 57 % accuracy⁴ according to either corpus?

In order to create a gold standard, SENSEVAL-1 (Mihalcea, 2004a), for comparing automated word sense annotation systems, a set of sample sentences were taken from the BNC and tagged with HECTOR senses. The word sense analyses had been developed using the sample. For a gold standard, it was critical that the interannotator agreement was high. To this end, a similar procedure as for creating the EngCG test corpus (Samuelsson and Voutilainen, 1997) was used. The individuals to do the tagging were carefully chosen: whereas other tagging exercises had mostly used students, SENSEVAL-1 used professional lexicographers. The material was multiply tagged, and an arbitration phase was introduced: first, two

⁴In this work accuracy means number of correctly labeled answers. It is the same as precision and recall, cf. Section 4.1.1, when one label is given and expected for each instance.

or three lexicographers provided taggings. Then, any instances where these taggings were not identical were forwarded to a third lexicographer for arbitration. The scores ranged between 88 % to 100 %, with just five out of 122 results falling below 95 %. (Kilgarriff and Rosenzweig, 2000)

A similar evaluation was done for the SENSEVAL-2 corpora, which were manually annotated with WordNet senses. The interannotator agreement was found to be 85.5 % (SENSEVAL-2, 2001). The lower agreement was expected and is probably due to the fact that WordNet senses have been created by providing synonym sets for words instead of word senses based on corpus analysis, cf. Section 2.4.2. To overcome the lack of sense-tagged data and the limitations imposed by the creation of such data using trained lexicographers, the Open Mind Word Expert system enabled the collection of semantically annotated corpora over the Web for some of the SENSEVAL-3 corpora (Chklovski and Mihalcea, 2002).

Within the SENSEVAL-1 framework corpora for French, Italian, Portuguese, and Spanish were also annotated. The SENSEVAL-2 exercise provided corpora for languages such as Basque, Chinese, Czech, Danish, Dutch, English, Estonian, Italian, Japanese, Korean, Spanish, and Swedish annotated with WordNet senses. For SENSEVAL-3 Basque, Catalan, Chinese, English, and Italian corpora were created as well as some multilingual corpora. (Mihalcea, 2004a)

Outside the SENSEVAL framework there are a host of efforts for annotating corpora with the intended meaning relevant to a special application domain, e.g., bi- and multilingual text alignment for machine translation, aligning speech and text corpora for speech technology, relevance judgments for information retrieval.

3.4 Representing Context

When representing the context, all systems extract characteristics from the context and make them into features with values. In this work we only use written text input with the following context sizes and depth of linguistic preprocessing:

- *ZER* – zero context with capitalization and morpheme structure
- *LOC* – local context with bigrams and head syntax
- *SYN* – local context with dependency syntax
- *GLOB* – global context with base form of all words in surrounding sentence

A dependency syntax parser by Connexor (Connexor, 2002b) provided all the linguistic analysis used in the feature extraction for the features that require linguistic preprocessing.

The global features were the morphologically disambiguated base forms in context (GLOB) which were extracted for all the words in the sentence surrounding the keyword. The sentence was found to be the most useful context for global features in the experiments in (Publication 1).

We used two types of local features: (LOC) and (SYN). The local features (LOC) were extracted from a window of ± 3 words around the keyword. We used the bigram collocations of the keyword and the base forms, the unnormalized word form of the keyword, and the head syntax and part of speech labels of the base forms. If a ± 3 -window extended over the sentence border, it was padded with empty words.

The syntactic features (SYN) consisted of the dependency information in the n-tuples $\langle W_1, M_1, R, W_2, M_2 \rangle$, where W_1 and W_2 are base forms in a dependency relation R , and M_1 and M_2 are the word class features of W_1 and W_2 , respectively. If M_2 is a preposition or a coordinator, the n-tuple $\langle W_1, W_2, R, W_3, M_3 \rangle$ was also extracted, where W_3 is in a dependency relation to W_2 , and M_3 is the morphological tagging.

The zero context feature (ZER) is the unnormalized form of the keyword, i.e., capitalization and inflection is preserved.

For the word sense discovery task, cf. Section 6.2, we only used a simplified set of syntactic features of the shape: $\langle R, W_2 \rangle$, and for discovering cross-lingual transliterations, cf. Section 6.3, we only used the zero context feature with normalized capitalization.

In Table 3.1⁵, we present a summary of the linguistic features for a sample sentence.

Often complex sentence structures, e.g., dependency or phrase structure relations, are represented as graphs. These structures can be given a piecewise representation in a bag-of-words or bag-of-features approach. The bag is implemented as a feature vector with each feature or feature combination as a separate dimension in order to allow for efficient matrix calculations. Such vector implementations lend themselves to kernel methods or kernel-based methods⁶ in supervised machine learning techniques, e.g., support vector machines (SVM) (Christianini and Shawe-Taylor, 2000), and in unsupervised machine learning techniques, e.g., self-organizing maps (Ritter and Kohonen, 1989; Kohonen, 1997).

Recently, implementations using weighted finite-state representations have also become available for kernel methods. Rational kernels define a general kernel

⁵The abbreviations of the morpho-syntactic tags in the table are: NH noun head, VA auxiliary verb, >N noun determiner, CC coordinator, PRON-SUP-PL superlative plural pronoun, V-PRES present tense verb, A-ABS absolute adjective, N-PL plural noun, DET determiner

⁶Originally, a kernel is a mathematical object which gives some features or feature combinations zero weight. For practical purposes, the interesting features and feature combinations are usually those which get weights different from zero.

Table 3.1: The linguistic features of a test sample for the keyword *church*. The bag of base forms is abbreviated as BoBF, head syntax label as HS, and part of speech label as PoS.

<i>Sample</i>	... most have quaint churches and other features of interest.		
<i>GLOB</i>	<i>correct base forms in sentence context</i>		
BoBF	... many have quaint church and other feature of interest		
<i>LOC & ZER</i>	<i>collection of features in window context</i>		
-3 bigram	< -3, many >	-3 HS, PoS	< -3, NH, PRON-SUP-PL >
-2 bigram	< -2, have >	-2 HS, PoS	< -2, VA, V-PRES >
-1 bigram	< -1, quaint >	-1 HS, PoS	< -1, >N, A-ABS >
0 string	< 0, churches >	0 HS, PoS	< 0, NH, N-PL >
+1 bigram	< +1, and >	+1 HS, PoS	< +1, CC, CC >
+2 bigram	< +2, other >	+2 HS, PoS	< +2, >N, DET >
+3 bigram	< +3, feature >	+3 HS, PoS	< +3, NH, N-PL >
<i>SYN</i>	<i>collection of local dependency relation features</i>		
With attr	< church, N, attr, A, quaint >		
Coord by	< church, N, cc, CC, and >		
Coord with	< church, and, cc, N, feature >		
Object of	< church, N, obj, V, have >		

framework based on weighted finite-state transducers to extend kernel methods to the analysis of variable-length sequences or, more generally, weighted automata (Mohri et al., 2003; Mohri, 2003; Cortes et al., 2004).

3.5 Summing up

We have studied the context and different dimensions of context. Then we had an overview of the ongoing efforts to collect contexts relevant for making word sense discovery and evaluating word sense disambiguation. Finally, we have specified the kind of context descriptions we use in this work. Next we turn to how lexical resources and context descriptions are used in evaluation.

Chapter 4

Evaluation

*Alice laughed. “There’s no use trying,” she said,
“one CAN’T believe impossible things.”
“I daresay you haven’t had much practice,” said the Queen.
“When I was your age, I always did it for half-an-hour a day.
Why, sometimes I’ve believed as many as
six impossible things before breakfast.”*

— Lewis Carroll (1875)

Through The Looking-Glass: And What Alice Found There

Is there a single best approach to word sense discovery or disambiguation? The answer is probably that different systems are best-suited for different applications and domains, but until we are able to evaluate the systems in a comparable way, we cannot even tell which system is best suited for what purpose. We now begin to have an idea of what the characteristics of a word sense are, and we have a notion of how a context may determine a word sense.

In the evaluation of language processing technologies, two main approaches are used: *in vivo*, where results are evaluated in terms of their overall contribution to the performance of an application, and *in vitro*, where systems are tested independently of an application by using specially constructed benchmarks. The *in vitro* evaluation is further divided into two strategies by using either pseudo-words or real words.

In this work we study the relation between word senses and their contexts from two aspects: firstly, similar usage should give rise to similar meaning, and, secondly, different meanings should manifest themselves in different usages. This means that depending on the aspect, discovery or disambiguation, we need to evaluate either a discovered word sense inventory or labeled words in context. First we will recapitulate some basic concepts related to the evaluation of machine learning systems, and then we will describe approaches for evaluating word sense

disambiguation and discovery systems.

4.1 General Performance

In order to evaluate the performance of a system we can compare either with the difficulty of the problem, with a reference solution, or with other similar systems. We define some of the generally used measures for evaluating system performance comparing with the difficulty of the problem or with a reference solution. We then describe the impact of the training procedure and the training data when comparing similar systems.

4.1.1 Evaluation Measures

Baselines

Baselines with both lower and upper bounds serve to characterize the range within which any system is expected to perform. If the system performs outside of this range, it requires additional explanations. Baselines give an initial estimate of how difficult the problem is and they are useful when there is no other similar system to compare with. Upper and lower bounds are always given with reference to the training or test material because they say something about the bias in the data.

As a lower bound we can use random choice, i.e., a uniform prior, if the algorithm is expected to pick at least one answer out of several. A uniform prior is a good baseline only if we know nothing about the prior distribution of the potential answers. Usually, there is some known inherent bias in the data, which makes the uniform prior an artificially low baseline. In word sense disambiguation the sense distribution is often very skewed, cf. Section 2.4.1, so the result of always picking the most frequent sense, i.e., the predominant sense, is used as a lower bound. It has recently been demonstrated (McCarthy et al., 2004a) that the predominant sense can be estimated from an unannotated domain specific corpus and a thesaurus. This means that in theory the lower bound is always known for any domain and sense inventory.

To estimate the upper bound for semantic tasks in natural language processing systems, we often use humans as a reference. The interannotator agreement¹ is considered to be the best that a reliable system can be expected to perform. If a system performs better, it is obviously biased toward a particular view of the test material.

¹The term interannotator agreement is sometimes used to refer to agreement between human annotators, whereas the term intertagger agreement is reserved for agreement between tagging systems, but some refer to both humans and machines with the same term.

In word sense discovery upper and lower bounds are difficult to construct, especially as everyone is using different training corpora.

Precision and Recall

Let K be the set of intended keys and L be the set of answers found by the system. The recall R , i.e., the number of true positives, is the number of found keys among the intended keys $|K \cap L|/|K|$. The precision P is the number of correct answers among the found answers $|K \cap L|/|L|$. Trivially, the precision is high with a low recall if only one correct answer out of many is provided, and the recall is high with a low precision if everything including all possible answers is always provided. Often a harmonic mean of precision and recall is calculated in order to promote systems providing all and only the correct keys. The harmonic mean is defined as $F = 2 * P * R / (P + R)$.

Average Precision at Maximum Recall

A fairly good picture of a system is given by the average precision at maximum² recall, i.e., we compute the precision when we receive the last correct answer for a test instance. At the end of the testing, we take the average of all the precisions. If we optimize the system with regard to this measure, the main effort is directed toward removing noise or intervening unwanted answers.

In an application we are often satisfied with even one correct solution. For this purpose the ideal method gives the correct answer as the first choice, i.e., the precision is 100 % at the level of 100 % recall. Occasionally, non-ideal methods provide x incorrect answers before the correct one, i.e., the precision is $1/(1+x)$ at 100 % recall. If the correct answer is not among the candidates, the precision is set to 0 % at 100 % recall for this test instance. The overall performance is again the average of the precisions at 100 % recall of all test instances.

Confidence Intervals and Significance Tests

When we have a long series of observations, we get statistically more confident in the average as we collect more observations. Based on the average and the standard deviation of the observations, we can give an estimate for our confidence that a new observation will be in a certain interval, a confidence interval. If the observation is outside the interval, we can reject the observation at a known relatively small risk p . When we test if an observation is outside our confidence interval, we

²The maximum recall is conventionally called “100 % recall”, because the maximum is the most we will get from a method in practice, although theoretically there may be additional correct answers.

use a two-tailed t-test. Then p is the risk we take (or the significance level) if we choose to believe that the new observation is generated by a different process, i.e., the new observation is statistically significantly different at the p level.

If we have two matching series of observations, we can use a t-test on the difference. This is a very sensitive test and for more convincing results we can use a slightly less sensitive test, the McNemar test (Somes, 1983). McNemar is a non-parametric test using matched pairs of labels. It is essentially a sign test on nominal data, which uses matched-pairs of labels (correct, incorrect) (Kanji, 1999). It determines whether the proportion of changes from incorrect to correct and vice versa is significant.

4.1.2 Division of Training and Evaluation Data

The credibility of the research requires that there is a separate set of training data on which the parameters are fine-tuned either manually or by some automatic procedure. When the experimenting or training is done, a separate test data set is used for evaluating the parameter selection.

During the training phase, it is useful to split the training data into an estimation and a validation data set. The estimation data set is used for learning characteristics about the data and the validation data set is used for testing the system on new data. If the data is split into a single estimation and validation subset, we call this the split-sample validation method. Since data is often scarce, we need to use the estimation data maximally. In k -fold cross-validation, the data is divided into k non-overlapping subsets of approximately equal size. The system parameters are estimated k times, each time leaving out one of the subsets from training and using it for validation. Kearns (1996) shows that the optimal split of the data into estimation and validation subsets for k -fold cross-validation is generally found, when $k = 5 \dots 10$. Cross-validation is always more effective than split-sample validation except in some pathological cases as demonstrated by Goutte (1997).

4.1.3 Training Data and Parameter Tuning

When comparing systems, it would in fact be more important to know how they improve with increased amounts of data than comparing them on a fixed data set. The argument is made vividly by Banko and Brill (2001). They explore the performance of a number of machine learning algorithms as the size of the training corpus grows from a million to a billion words. All the algorithms steadily improve in performance, but different systems are best with the largest data size than with the smallest data size. In addition, they show that getting more data will generally have a bigger impact than fine-tuning the parameters of an algorithm.

From the study we can conclude that important characteristics of a machine learning system are the training speed, the learning rate, and the capacity to improve with quality and quantity of data. If the training algorithm is too cumbersome, the learning rate is too slow, or there is a limit to how much the system can improve despite more data of better quality, the system is likely to be outperformed with increased amounts of training data.

A study by Daelemans and Hoste (2002) demonstrates that the impact of parameter settings and feature selection accounts for a higher variation in performance than differences between the algorithms. The consequence is that the suitability of an algorithm for a particular task cannot really be decided in a meaningful way without parameter setting and feature extraction optimization.

4.2 Using Annotated Contexts

The capability to evaluate systems for word sense disambiguation has improved with the SENSEVAL initiative (Kilgarriff and Palmer, 2000; Edmonds and Kilgarriff, 2002; Mihalcea and Edmonds, 2004) providing generally available annotated context collections as test material. The debate is mainly focusing on how well this material represents the word sense disambiguation tasks encountered in applications.

4.2.1 Real Word Contexts

If we evaluate real words in real contexts, the SENSEVAL initiative (Mihalcea, 2004a) offers the best available corpora. The corpora have been collected both for the all-words task, i.e., word sense tagging of all the open words in a sentence, as well as the lexical sample task, i.e., word sense tagging of specially chosen words in context. The test suites in SENSEVAL have been divided into training material and test material, with approximately 2/3 of the material allocated as training material.

The choice of word senses in the SENSEVAL material is WordNet for SENSEVAL-2 and SENSEVAL-3. As we saw earlier, cf. Section 3.3.2, this had the impact of lowering the interannotator agreement from SENSEVAL-1, which had the HECTOR senses. In addition, the best system for the English lexical sample in SENSEVAL-1 performed with 77.1 % accuracy (Kilgarriff and Rosenzweig, 2000), whereas for SENSEVAL-2 the best system performed with 64.2 % accuracy (SENSEVAL-2, 2001). In SENSEVAL-3 the best system for the English lexical sample improved to 72.9 % accuracy (Mihalcea et al., 2004).

A model which always chooses the predominant sense performs remarkably well. In the English lexical sample task the most frequent sense baseline was

69 %, 47.6 %, and 55.2 %, in SENSEVAL-1, SENSEVAL-2, and SENSEVAL-3, respectively. A system that concentrates on identifying the predominant sense could well outperform one that concentrates on disambiguation. The corresponding interannotator agreement scores were 96 % and 85.5 %, for SENSEVAL-1 and SENSEVAL-2, respectively³.

4.2.2 Pseudo-Word Contexts

The use of pseudo-words in context means that the ambiguity is artificially constructed by taking all the instances of two or more words and replacing them with an artificial word. This artificial word is then regarded as ambiguous and can be disambiguated to any of the original words.

The pseudo-word method was especially popular, when no large-scale manually annotated corpora existed (Gale et al., 1992a; Schütze, 1992). Initially, only two-way ambiguities were used with the general misconception that pseudo-words with a 50 : 50 distribution for both senses were best suited for evaluation purposes. Generally, the disambiguation was reported to have 90 – 95 % accuracy for words chosen by the evaluators themselves. Such results were probably too optimistic with regard to real words, which has been verified by Gaustad (2001).

As we have seen, cf. Section 2.6.1, the words in WordNet are on the average three-way ambiguous⁴. In addition, Kilgarriff shows that an even distribution is quite rare, cf. Section 2.4.1. A very skewed distribution of the predominant sense of a word as well as a higher number of senses for frequent words is much more likely. If we control for these parameters, the pseudo-words in context are still a viable, efficient, and low-cost approach to constructing training and test corpora.

4.2.3 Application-Specific Contexts

If we evaluate word sense disambiguation routines with applications, e.g., for machine translation, information retrieval or speech technology, the applications introduce additional problems and techniques for coping with these problems that may partially solve some of the disambiguation problems as a by-product. As a consequence, solving the word sense ambiguity problem does not necessarily have a direct influence on the final performance.

Word sense disambiguation is a central problem for machine translation, but it is usually sidestepped, by using different lexicons for different domains. While

³No interannotator agreement scores have been reported for SENSEVAL-3, only the intertagger agreement of the best two systems, which was reported at 67.3 %. The corresponding figure was 66.5 % for SENSEVAL-2 (Mihalcea et al., 2004)

⁴Polysemous verbs are on the average 4-way ambiguous and the other polysemous words are 2–3-way ambiguous (Miller et al., 2003).

this approach may have been born out of pragmatism rather than theory, a model always selecting the most frequent sense tends to support it. If a machine translation (or some other NLP) system is operating within a domain, it may be more effective to customize the lexicon for the domain than to attempt to resolve ambiguity at runtime (Kilgarriff, 2004).

A well-known example of the influence of context in applications is found in information retrieval where query disambiguation seems to help significantly only for one-word queries, i.e., queries where there is no context (Sanderson, 2000). The conclusion is that long queries contain a sufficient amount of context for automatically selecting the correct word sense of a keyword.

Presently, the most likely development seems to be precompiled word sense disambiguation by tailoring a general lexicon to the domain using the methods suggested by McCarthy et al. (2004a,b) for finding the domain-specific predominant sense and for filtering out word senses that are rare or non-existing in a given domain. The ideal would be to have the lexicon dynamically adapt to shifting domains. In the future, this might make separately maintained domain-specific lexicons redundant.

4.2.4 Context Preprocessing

It has been shown that the quality of the context representation is often more important than the difference in system design. In Lee and Ng (2002), it was shown that the disambiguation effect of local linguistic features was considerable regardless of which learning method was chosen achieving results between 57.2–65.4% accuracy on the fine-grained lexical task of the English SENSEVAL-2 data. Their analysis showed that adding more complex linguistic features to the base forms, e.g. syntax and part-of-speech labels, accounted for an absolute improvement of 8–9% of the disambiguation result of the best algorithm. Yarowsky and Florian (2002) and Voorhees et al. (1995) compare several linguistic features and algorithms arriving at the conclusion that major differences in the feature space was a more dominant factor than differences in algorithm architecture. (Publication 3)

In order to control for the linguistic preprocessing of the context, it could be useful to have test corpora that are preprocessed in the same way for all systems, which the systems could use as a baseline in order to compare with their own preprocessing strategies. The need to optimize the feature extraction for each algorithm as pointed out in Section 4.1.3, may make it impossible in practice to decide what kind of preprocessing to provide for the corpus. Instead, test reports should contain some sensitivity analysis for different types of features and their interaction.

4.3 Using Word Sense Inventories

The capability to compare systems for word sense discovery is still rather poor, because there is no agreed-upon standard word sense inventory.

4.3.1 Obstacles

Partly the difficulty is due to the fact that word sense discovery is open ended. However, this is a problem shared with information retrieval (IR). In IR, relevant documents for a query are identified in a database and used as an evaluation standard (Voorhees, 2000). In a similar fashion, relevant word senses (with lists of synonyms and words in other thesaurus relations) should be identified for words with regard to a given training material in order to form an evaluation data set. To some extent the available training and test data sets for SENSEVAL (Mihalcea, 2004a) can be used for this purpose, but the amount of data is rather small for discovery purposes. We would need large corpora for which a sense inventory for certain words has been agreed on without explicitly tagging each instance.

In addition, we face the problem that distributional similarity contains a partial ordering, i.e., words are similar only to the extent in which they appear in similar contexts, e.g., a *dog* is similar to an *animal* only when they both appear in a *dog* context, whereas *animal* also has many other meanings. We may eventually need to evaluate the similarities with reference to some generalized context description. This has not been explicitly addressed, although context samples for WordNet synonym sets are presently added.

At this time researchers are feeding very different corpora to their discovery procedures and the preprocessing of these corpora is widely different, which complicates the comparison. The preprocessing that has been found important for discriminating annotated contexts is even more important when trying to establish similarities and differences between contexts in order to create word sense clusters, i.e., the quality and depth of the linguistic preprocessing is crucial, because the extracted context features are the only available information to base the similarity judgments and clustering decisions on.

Despite the obstacles a few evaluation methods have been proposed as we will see in the next three sections.

4.3.2 Real Word Inventories

Researchers have been comparing their discovered word senses with WordNet and its sense inventory using WordNet as a gold standard. For English, some have also used Roget's or some other available thesaurus. There is no standardized way for using them and like Lin (1998) many pick a target word in a thesaurus, from which

they extract all the synonyms (or other related words) of the target word, and create a list of the words and their similarity ratings within the thesaurus relative to the target word. This method provides rather long similarity lists. Such similarity lists can be represented as vectors and compared with, e.g., by the cosine of two vectors.

Pantel (2003) has a more elaborate approach using an edit distance on sets in order to compare the existing synonym sets in WordNet with the word clusters he discovers. For Finnish (Publication 4), we compared the clustered synonym lists with a sense inventory created from translation dictionaries. The intersection of back translations in the translation dictionaries provided fairly concise synonym sets with an average of only 3.5 synonyms per synonym set.

Evaluation against human plausibility judgments is related to evaluation against a gold standard and is often used when a gold standard is unavailable or incomplete. We tried this for testing the method in (Publication 4) by manually grouping the words in the similarity lists extracted from a corpus.⁵ We later manually evaluated the quality of the clusters produced by the best clustering algorithm.

4.3.3 Pseudo-Word Inventories

Creating pseudo-synonym sets could be a possible road ahead for evaluating word sense discovery algorithms. The pseudo-word disambiguation technique was used in word sense disambiguation, as mentioned above, even though it was done largely on false assumptions about the word sense distributions.

A similar study as Kilgarriff did for word senses should be done for the distribution of synonyms in synonym sets. Our conjecture is that similar Zipfian power-law behavior will be discovered. Based on this study pseudo-synonym sets with n synonyms can be created for a monosemous word w by taking the instances of w and renaming them with w_1, \dots, w_n according to a power law distribution. With this pseudo-synonym corpus one can compare word sense discovery algorithms based on their ability to cluster pseudo-synonyms. Grefenstette (1994) used this technique and split the word into two synonym sets at different percentages finding that the recall of the smaller synonym portion drops considerably when the portion shrinks below 20 %. As for pseudo-words the figures for pseudo-synonyms give upper bounds. If we had a better estimate of the real distribution of synonyms, we could get a more realistic figure for the recall of low frequency synonyms.

⁵Both authors of the article agreed that we would never ever force anyone to look at such lists again. After a few words with each of them having 100 synonyms and related words we gave up. This was definitely a task for machines.

4.3.4 Application-Specific Word Inventories

As is pointed out by Weeds (2003), evaluation against a semantic gold standard like WordNet is an evaluation task in the application of automatic thesaurus generation. Grefenstette (1994) evaluated the automatically generated synonyms in an information retrieval task and in linking words to a WordNet hierarchy. For other applications, see Curran (2003).

A task such as word alignment for machine translation is an instance of cross-lingual word sense discovery, where an online translation dictionary can serve as an evaluation standard for the translation sense inventory we wish to discover. A related problem can be found in information retrieval, where proper nouns and technical terms often are prime keys in queries (Pirkola and Järvelin, 2001). In rapidly evolving domains like biotechnology new translations are often missing in online translation dictionaries, but we can use the available technical terms in multilingual technical or medical terminologies as evaluation standards.

4.4 Summing up

In this chapter, we have looked at how to evaluate systems for word sense disambiguation using context collections, and how to evaluate word sense discovery systems using word sense inventories. For disambiguation systems the situation is acceptable and improving due to the SENSEVAL initiative, but for discovery systems we are still only beginning to think about what the evaluation would require by exploring various ideas.

Next we turn to an overview of the broad ideas motivating word sense disambiguation systems. Then we look at a specific model for how the context features can be selected and emphasized in order learn to identify a given word sense, i.e., how to “suit the action to the word”⁶.

⁶Shakespeare, “Hamlet”

Chapter 5

Disambiguation

*Arthur Weasley*¹: “Now, Harry you must know all about Muggles², tell me, what exactly is the function of a rubber duck?”

— JK Rowling (2002)

Harry Potter And The Chamber Of Secrets

Word sense disambiguation is the task of selecting the appropriate sense of a word in a given context. In order to do so we need to train a disambiguator to weight the context features to accurately distinguish the given word senses.

Disambiguators have been constructed by two approaches: manually or by machine learning. The machine learning approach can be divided into supervised, semi-supervised and unsupervised approaches. All of them are faced with the same task: given two or more different meanings or word senses, they need to find distinctive contexts or usages for each sense.

First we give a brief overview of the ideas that were introduced already in the 50’s and are still used in word sense disambiguation. For a more in-depth historical overview of how the ideas developed, we refer the reader to Ide and Veronis (1998) and Rapaport (2005). Then we look at how the different approaches to machine learning are applied to word sense disambiguation. Finally, we give an outline of the ideas and the contribution of this work to word sense disambiguation.

5.1 Manually Developed Resources

The main ideas for manually developed resources in word sense disambiguation has evolved into two broad categories, i.e., artificial intelligence-based and

¹A wizard employed at the Ministry of Magic. His hobby is to collect all sorts of facts about humans and their behavior.

²Humans with no magical powers.

knowledge-based ideas. First we will look at some of their common early ideas following the outline of Ide and Veronis (1998). We will then turn to machine learning methods³, which are the bulk of the current systems.

5.1.1 Early Ideas

Many of the ideas which are reiterated and refined today in WSD were put forward already in the early days of computational linguistics during the 1950s. Kaplan (1950) observed that two words on either side of a word was not significantly better or worse than giving the entire sentence to human translators in order for them to do word sense resolution. Reifler (1955) observed that the syntactic relations between words are often a decisive component when determining a word sense. At the same time, Weaver (1955) pointed out that a given word most often only has one meaning in a particular domain. (Ide and Veronis, 1998)

Following the observation that a word generally has only one meaning in one domain, many natural language processing systems still today use domain-specific dictionaries or microglossaries as the main solution for tailoring the process to a specific domain.

As Weaver emphasized in his Memorandum (Weaver, 1955):

This approach brings into the foreground an aspect of the matter which is absolutely basic—namely, the statistical character of the problem. . . . And it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken as a necessary primary step.

This was also pursued by various researchers, and Pimsleur (1957) introduced the notion of levels of depth for a translation corresponding to what today is known as the most frequent sense or baseline tagging. Level 1 uses the most frequent equivalent, producing 80 % correct translations, level 2 contains additional meanings, producing 90 % correct translations, etc. (Ide and Veronis, 1998)

As Ide and Veronis (1998) point out in their survey, many of the fundamental ideas of word sense disambiguation that have resurfaced several decades later, were originally tested only on a small scale due to severe constraints on the available resources, i.e., computers, corpora and machine-readable dictionaries. Even though the fundamental ideas themselves are not new, it is still interesting to revisit them in light of recent developments in computational linguistics and soft computing, see Baayen (2001) or Manning and Schütze (1999).

³Ide and Veronis (1998) used the term corpus-based methods, but currently everything in computational linguistics is corpus-based – even the manually created resources are at least corpus-informed and many are also evaluated against corpora.

5.1.2 Artificial Intelligence Ideas

The idea of interlingua was also proposed in the 1950s leading to semantic networks by Richens (1958) and Masterman (1962) using a language-independent semantic network of concepts, onto which the words could be mapped. AI methods such as semantic networks were used for disambiguation by trying to find the shortest path through a common concept for two words. One of the problems with the AI methods was the knowledge acquisition bottleneck. (Ide and Veronis, 1998)

The knowledge acquisition bottleneck is now being approached in a distributed effort all over the World Wide Web under the title Semantic Web with the aim to create a compatible set of ontologies or knowledge repositories which can be used for various language understanding and information processing tasks by man or machine. “The Semantic Web is a web of data, in some ways like a global database” as the effort is characterized by its inventor Tim Berners-Lee (1998, 2000). What seemed like a bottle-neck in the 1950s, because the information needed to be manually encoded in a structured way, now seems like it could be achievable due to a world-wide distributed encoding effort and exploited as sample contexts for machine learning methods in a not too distant future.

5.1.3 Knowledge-based Ideas

In order to remedy the knowledge acquisition bottleneck for natural language processing, several knowledge-based methods using machine-readable dictionaries and thesauruses have been investigated. The main problem with machine-readable dictionaries is that they are designed for humans. Often they are too detailed and not formal enough for machines. The most common way to disambiguate with machine-readable dictionaries is to select the word sense which according to some given criteria maximizes the relatedness among the word senses of the words co-occurring in a passage of text.

One of the most frequently used resources is WordNet (Fellbaum, 1998), an electronic online thesaurus with more than 152,000 words of English. WordNet has served as a basis for creating corresponding WordNets in more than 35 languages (Vossen, 2001; Vossen and Fellbaum, 2004). WordNet may not have all the necessary information, but currently there are few other publicly available resources which could compete with it. Since it is the focus of such interest, it will hold its position and develop further for some time to come.

5.2 Machine Learning Methods

Statistical methods have become a standard paradigm in computational linguistics. They can be grouped roughly into descriptive statistics, generative models (i.e., stochastic finite-state, context-free, and attribute-value grammars), and machine learning methods (i.e., supervised, semi-supervised, and unsupervised methods). (Abney, 2005)

Most current methods for word sense disambiguation use machine learning. Supervised methods require annotated corpora. Semi-supervised methods use mainly unannotated corpora combined with some annotated data. Unsupervised methods get by on unannotated corpora combined with a thesaurus for a sense inventory. The trend is to use all the resources available (Agirre et al., 2000), and consequently some hybrid methods use unannotated corpora, thesauruses and annotated corpora when available. Annotated corpora are costly to develop, so the trend is also toward semi-supervised or unsupervised methods.

5.2.1 Supervised

Supervised machine learning of classifiers uses a set of annotated training data from which a classifier is induced by an algorithm. The training algorithm is called supervised if it uses the annotated training data for improving the capacity of an classifier to reproduce the annotation. For an introduction to supervised machine learning, see Manning and Schütze (1999) or Christianini and Shawe-Taylor (2000). For a description and evaluation of a number of supervised methods applied to word sense disambiguation, see Kilgarriff and Palmer (2000), Edmonds and Kilgarriff (2002) and Mihalcea and Edmonds (2004).

5.2.2 Semi-Supervised

The semi-supervised or minimally supervised methods are gaining popularity because of their ability to get by with only a small amount of annotated reference data while often outperforming totally unsupervised methods on large data sets. There are a host of diverse methods and approaches, which learn important characteristics from auxiliary data and cluster or annotate data in light of the acquired information. The auxiliary data may provide seeds for labeling the primary data (Yarowsky, 1995; Blum and Mitchell, 1998; Nigam et al., 2000; Banko and Brill, 2001; Abney, 2002), or it may provide some structure of its own, which guides the clustering of the primary data (Tishby et al., 1999; Kaski et al., 2005b,a).

5.2.3 Unsupervised

Typically, unsupervised methods use unannotated data sets, which mainly enable clustering of the data. In order to perform annotation they need to learn some annotation for the clusters. As long as they adopt the annotation without modifying the clustering, the algorithms are considered unsupervised. For an account of the unsupervised systems used in word sense disambiguation, see Kilgarriff and Palmer (2000), Edmonds and Kilgarriff (2002) and Mihalcea and Edmonds (2004).

Following the observations of Kilgarriff that the most dominant sense is usually very dominant, cf. Section 2.4.1, it may be interesting to determine only the predominant sense of a word and then tag each word with this word sense. McCarthy et al. (2004b) create a corpus-based similarity list for a target word. They use the list for ranking the WordNet senses of the target word, and identify the predominant sense. For some words the predominant word sense may vary according to domain as Ide and Veronis demonstrated, cf. Section 3.2, so the predominant word sense should be learned from a domain-specific corpus.

5.2.4 Combining classifiers

As classifiers usually are strong at different aspects of a task, it is often beneficial to combine them into ensembles. For different methods of combining classifiers for word sense disambiguation, see Florian et al. (2002) and Florian and Yarowsky (2002). Classifiers in an ensemble need to be different from each other, otherwise there is no gain in combining them. Unfortunately, there is no single best way for measuring classifier diversity (Brown et al., 2004).

5.2.5 Degrees of Supervision

We need to dwell on the difference between supervised and unsupervised training algorithms in the case where some classifier output is needed.

If we transfer information from unannotated data to an existing classifier, this counts as unsupervised training. However, if we transfer information in the other direction using a classifier to improve the clustering of the data, the situation needs clarification. Many would agree that preprocessing for an unsupervised system is allowed if it does not add the annotation we are later trying to discover. Postprocessing is also allowed, as long as it does not modify the clusters provided by the unsupervised learning. However, it is at least a mild form of supervision if an external classifier is used for determining or improving the clustering in order to better reproduce the annotation.

An external classifier may appear in many disguises: it may arrive in the form of a hand-made precompiled data structure with an elaborate access function, or it may be a list of annotated samples with a similarity function. In word sense disambiguation we have examples of both: a hand-made data structure with an elaborate access function is, e.g., the WordNet with the Lesk algorithm for semantic distance, whereas lists of annotated samples are provided by the SENSEVAL data.

SENSEVAL mainly uses the word senses provided by WordNet. All classifiers involved in word sense disambiguation must learn how to annotate an instance of a word in its context. The association can be learned from the WordNet hierarchies, from annotated WordNet synset glosses, or from a set of annotated samples. An unsupervised method may learn the annotation of a word belonging to a context cluster from any of these three sources.

The SENSEVAL organizing committee divided the algorithms of the SENSEVAL-1 and SENSEVAL-2 Lexical Sample task into supervised and unsupervised, solely based on whether they use SENSEVAL training data or not. This may have been a clear-cut and easy-to-implement rule, but it put the unsupervised classifiers in an awkward position, because they did not even have access to unannotated data from the same domain as they are being evaluated on. From what we now know about domain influence on word sense distribution, this was an overly harsh constraint giving unfair advantage to the supervised methods. However, this changed with SENSEVAL-3, where the best unsupervised system (Ramakrishnan et al., 2004) in SENSEVAL-3 was allowed to use the full training data set by defining it as an extended set of WordNet glosses.

5.3 Word Sense Discrimination

Schütze (1998) proposed a technique for word sense discrimination as an unsupervised learning method, which avoids the word sense labeling. He provided the ability to decide whether two senses are the same or different, based on the context in which they occur, but the link between the context clusters and the lexicographic word senses was not apparent.

The link between word sense and context was made explicit by the additional work of the current author in (Publication 1) as well as demonstrating that a sufficiently large and varied independent representation of context can assist in word sense disambiguation. The idea was to create a self-organized map of the contexts with WEBSOM (Kohonen et al., 2000), label the map with known seed contexts, and with the labeled map determine the sense of a word in context. Effectively, we use a word context to associate the word with its word sense. The method is named THESSOM and can produce the n-best scored word senses, which pro-

vides a preference method for determining the acceptability of a word sense in context.

5.4 THESSOM

First we present the philosophy behind selecting WEBSOM as the means to represent semantic space, then we outline the WEBSOM method, which is a method for creating large two-dimensional self-organized document maps using the SOM algorithm. Then we present THESSOM, which is a method for reading WEBSOM maps for the purpose of word sense disambiguation.

5.4.1 Background

In this work, we want to substantiate the claim that lexical semantics does not need a full metric between words, a local metric is sufficient. For word sense disambiguation it may be useful to know that *house* and *residence* are related and it may also be of interest whether in some context they are more closely related than *house* and *building*. However, it would be sufficient to know that *house* and *zinc mine* are unrelated in most contexts. It is unlikely that we need an accurate measure of whether they are more unrelated than e.g. *house* and *leg*.

The idea that most concepts are closely related to only a few others is supported by the research by Steyvers and Tenenbaum (2005) demonstrating that, when e.g. thesauruses grow, new concepts are likely to be added to clusters of already locally tightly linked concepts. Conversely, most concepts and their concept clusters are only indirectly related occupying distant parts of the semantic space.

The concepts that are closely related in semantic space can be captured by a low-dimensional local metric. This idea is supported by Lowe (1997) when he shows that most co-occurrence data is inherently very low-dimensional suggesting that in many cases two dimensions may be sufficient. Using single-value decomposition (SVD), he shows that 80 % of the co-occurrence information could be encoded in only a few SVD components (Lowe, 2001).

SOM is a non-linear projection of high-dimensional space onto a low dimensional display. SOM tends to preserve a fairly accurate image of the local high-dimensional neighborhood, although similar guarantees cannot be given for distant parts of the high-dimensional space (Martinetz and Schulten, 1994). We call this the local neighborhood preservation property of SOM.

5.4.2 Creating Document Maps

Assume that we have a collection of documents of various lengths from different domains. We also have a domain classification of the documents. All the words occurring in a document are thereby related to the domain of the document in the context of the other words in the document. If we consider each word or combination of words to be a dimension in semantic space, we wish to create a low-dimensional projection of the high dimensional semantic space such that documents with similar content end up near each other.

The WEBSOM method (Kohonen et al., 2000; Honkela et al., 1996) uses the Self-Organizing Map (SOM) algorithm (Kohonen, 1997; Ritter and Kohonen, 1989) to organize a large document collection in an unsupervised way onto a two-dimensional display called the map. The SOM map consists of a set of map units ordered on a two-dimensional lattice. By virtue of a model vector stored with each map unit, searches can be performed on the map in order to locate the most similar model vector for a new document or short context (Kohonen, 1997). The map unit of the most similar model vector is called the best-matching unit.

In WEBSOM, documents are encoded by using the bag-of-words vector space model. The features in the document vectors are weighted before the vectors are normalized. The cosine measure (dot product for normalized vectors) is used for measuring similarities between document vectors. Documents similar in content are located near each other on the ordered map display (Kohonen et al., 2000).

5.4.3 Calibration and Disambiguation

Assume that we have a word in context. The word has a number of possible word senses. We wish to determine which word sense is used in the current context. We also have a collection of sense-tagged samples of the word in various contexts. Assume that we have a representation of semantic space in the form of a document map. The document map decides which of the sense-tagged samples in our sample collection are relevant in a particular context by displaying similar samples near each other on the map. By also displaying the untagged sample on the document map and looking at the nearby sense-tags we may determine which sense is appropriate for the word in the current context.

In (Publication 1), a method is presented which calibrates a self-organized document map and uses it for word sense disambiguation. The method is called THESSOM. In (Publication 2), the method is presented in detail. In (Publication 3), the impact of several linguistic features on large-scale clustering is evaluated. The evaluated features are various qualitative features, e.g. part-of-speech and syntactic labels, and quantitative features, e.g. cut-off levels for word frequency. The best combination of parameters and linguistic features were separate

keyword maps⁴ using all features (SYN+LOC+GLOB) without frequency cut-off. This produces a representation of semantic space achieving 62.9 %±0.73 % correct results on the fine-grained lexical task of the English SENSEVAL-2 data. On the 96.7 % of the test cases which need no back-off to the most frequent sense we achieve 65.7 % correct results.

Our experiments in (Publication 3) show that extracting features guided by the syntactic structure (SYN) gives more precision to the word sense contexts than merely using the linear context encoded in (GLOB) or (LOC), but (GLOB) and (LOC) provide more coverage, so in practice the (GLOB) and (LOC) feature sets provide fall-back for the (SYN) feature set. As a consequence, the word sense disambiguation works best in total when using all three contexts together.

5.4.4 Test results

Lee and Ng (2002) evaluated the impact of different feature combinations extracted from the SENSEVAL-2 material using several supervised learning systems and compared the results with the three best systems in the SENSEVAL-2 exercise. The best reported performance without combining classifiers on the English SENSEVAL-2 data for a fine-grained lexical task is 65.4 % with the best results being in the range 62.9–65.4 %, i.e. 66.8–73.2 % for adjectives, 66.8–69.5 % for nouns and 56.3–61.1 % for verbs (Lee and Ng, 2002; SENSEVAL-2, 2001). Only by combining classifiers has a better overall result of 66.5 % been achieved by Florian and Yarowsky (2002).

It should be emphasized that all of the comparison systems are supervised, whereas the current system in light of the SENSEVAL-3 system classification would have been unsupervised, cf. Section 5.2.5. The best unsupervised system on the SENSEVAL-2 data performed with approximately 40 % accuracy.

5.5 Summing up

We recapitulated some of the initial ideas for word sense disambiguation. Then we looked at the different types of machine learning applied to word sense disambiguation, and finally we presented our own contribution to word sense disambiguation.

⁴Later I learned that Pulkki (1995) tested the idea of a separate SOM map for each word by resolving the verb/preposition ambiguity of the word form *like* in the context of ±1 word forms in the domain of Grimm fairy tales. My experiment aimed at more fine-grained distinctions of word senses at a larger domain-independent scale, but our separate findings are consistent with regard to the number of maps per word.

Next we turn to an overview of the broad ideas motivating word sense discovery systems. Then we look at two specific models for how a suitable word sense can be discovered for a given set of contexts, i.e., how to “suit the word to the action”⁵.

⁵Shakespeare, “Hamlet”

Chapter 6

Discovery

The Red Queen shook her head. “You may call it ‘nonsense’ if you like” she said, “but I’ve heard nonsense, compared to which that would be as sensible as a dictionary!”

— Lewis Carroll (1875)

Through The Looking-Glass: And What Alice Found There

Word sense discovery is the task of learning what senses a word may have in different contexts. Word sense discovery is regarded as a preprocessing stage when building mono- and multilingual dictionaries as well as thesauruses and ontologies, but it can also be seen as part of a larger effort in information extraction to adapt lexical resources to a new domain or a new language by acquiring new lexical items or new meanings of existing lexical items (Wilks and Catizone, 2000).

In the same way as word sense disambiguation systems, word sense discovery systems can also be constructed by two approaches: manually or by machine learning. The manually constructed discovery systems go back to the early systems for information extraction (Wilks and Catizone, 2000) and manually constructed extraction techniques have recently been applied in lexicography for dictionary construction (Kilgarriff et al., 2004; Kilgarriff and Tugwell, 2001). The aim is toward automated procedures, and machine learning is gaining space with an emphasis on unsupervised and semi-supervised methods, but for special tasks supervised methods are used.

First we give a brief overview of the ideas that were introduced already in the 50’s and 60’s and are still used in word sense discovery. For a more in-depth historical overview of how the ideas developed, we refer the reader to Grefenstette (1994) and Rapaport (2005). Finally, we give an outline of the ideas and the contribution of our work to word sense discovery.

6.1 Early Ideas

One of the earliest applications for the hierarchical structure of Roget's thesaurus was Masterman's work in the 1950s (Masterman, 1956) on creating an interlingua and meaning representation for early machine translation work. Masterman was convinced that the strong mathematical structure underlying, e.g., Roget's thesaurus, was a lattice of sets that could be exploited with a set theoretic interpretation of the thesaurus heads as sense categories (Wilks, 1998). Spärck Jones (1986) pioneered work in statistical semantic similarity analysis of words by using methods like factor analysis on the synonym lists of Roget's thesaurus to discover synonym sets. The sets could be identified with semantic primitives or semantic features like animate or human. (Wilks, 1998)

For language applications there is universal agreement that lexicons need to be adapted or tuned to new user domains. The disagreement is about what tuning implies and whether there is real benefit in terms of recall and precision. Lexical tuning has meant a number of things: the notion (as far back as Wilks 1972) has meant adding a new sense to a lexicon on corpus evidence because the text could not be accommodated to an existing lexicon. In 1990, Pustejovsky used the term to mean adding a new subcategorization pattern to an existing sense entry from corpus evidence. In the information extraction (IE) tradition there have been a number of pioneering efforts to add new words and new subcategorization/preference patterns to a lexicon from a corpus as a prolegomenon to IE. (Wilks and Catizone, 2000)

6.1.1 Word Sense Similarity

Since the first attempts were made in the 1960s by Spärck Jones (1986), several other methods have been tried for automating the process of synonym set and thesaurus construction from text corpora (Hindle, 1990; Grefenstette, 1994; Lin, 1998; Pantel, 2003; Curran, 2003; Weeds, 2003) during the last ten to fifteen years, when sufficiently large corpora and efficient computers have become available. In his paper Hindle (1990) refers to Zellig Harris' hypothesis about syntactically similar usage conveying similar meaning, and his similarity measure is based on that idea.

Grefenstette (1994) calculates similarity lists of a target word with other words in a corpus. He uses his system (SEXTANT)¹ to extract the 10 most similar words for each target word found in a corpus. His evaluation shows that the target word and its 10 most similar words appeared under the same head in Roget's thesaurus

¹The similarity calculation is based on a full set of dependency features using a variety of similarity measures, but only the results of the Jaccard and weighted Jaccard distance measures are reported.

significantly more often than theoretically could be expected. Some words did not co-occur because they were missing from the thesaurus. Among these, a small number of words were synonyms in some domain-specific context outside the coverage of a general English thesaurus. Others were semantically related but not synonyms and some were totally unrelated. Grefenstette demonstrates the stability of the result. From his experiments we can draw the conclusion that adding 20 % new context instances to the corpus material of words that already have more than 50 context instances does not greatly influence the list of the top 10 most similar words. As noted in Section 4.3.3, this stability means that non-dominant senses represented by less than 20 % of the contexts tend not to be among the top 10 similar words.

For an exhaustive overview and formal comparison of various similarity measures, see Mohammad and Hirst (2005). They propose that it may be useful to distinguish between semantic similarity and semantic relatedness. Semantic similarity is conveyed by similar syntactic usage as originally proposed by Harris, whereas semantic relatedness is conveyed by other co-occurring words. This seems plausible, but since all words in a sentence are somehow syntactically related, the interpretation must be that immediate constituents and head words determine semantic similarity as substitutability, whereas the remaining words in the context influence only the semantic relatedness. This is supported by our findings in (Publication 3), where we demonstrate that words related syntactically through immediate constituents or head words provide better accuracy for semantic similarity judgments than other words in a sentence.

6.1.2 Word Sense Clustering

The recent word sense discovery research has focused on evaluating similarity measures. Of the methods referred to above only that of Pantel (2003) actually takes the full step to synonym set discovery in addition to the method we propose. However, there are others who have focused more directly on synonym sets as word clusters.

In a seminal paper Pereira et al. (1993) using the Kullback-Leibler (KL) distance hierarchically cluster English nouns based on the probability distributions of the nouns as objects of verbs. This was generalized into a weakly supervised clustering method called the Information Bottleneck method by Tishby et al. (1999). Lee (1999) investigated various similarity measures for judging substitutability of nouns as objects of verbs and found that the Jensen-Shannon measure, the Jaccard measures and the Skew divergence were good candidates. The Jensen-Shannon measure is the basis for our research on word sense clustering in (Publication 4).

Schütze (1998) proposed a technique for word sense discrimination, which essentially clusters word contexts into senses or concepts based on the cosine

distance between contexts represented as vectors. This idea was used in the Information Mapping project² for concept-based information retrieval. Similar ideas of clustering documents in order to create a concept space were explored in the WEBSOM project³. One of the essential differences between the two projects is that WEBSOM also creates a topological ordering or map of the clusters using the SOM algorithm, which makes it possible to exploit the map surface directly for similarity judgments as demonstrated in (Publications 1, 2 and 3).

The naming of the clusters remains problematic for unsupervised methods, except perhaps in the case of translation discovery, where a word cluster in one language can be named according to its common target word in the other language. The labeling of clusters requires that the labels are provided either with sample contexts by which the labels can be matched with the clusters (Publication 1), cf. Section 5.4 THESSOM, or with principles for labeling clusters according to their properties. In the Information Mapping project Dominic Widdows (2002) used the two principal components of the word contexts of a set of similar words for producing a 2-dimensional display of how the words are related⁴. In the WEBSOM project, there is already a 2-dimensional display of the whole document collection due to the SOM algorithm, so Lagus and Kaski (1999) is able to select salient words of the WEBSOM map regions for visualizing the contents of the map regions.

6.2 Monolingual Discovery

Finding related words among the words in a document collection can be seen as a clustering problem where we expect the words in a cluster to be closely related to the same sense or to be distributional substitutes or proxies for one another. A number of language-technology tasks can benefit from such word clusters, e.g. document classification applications, language modeling, resolving prepositional phrase attachment, conjunction scope identification, word sense disambiguation, word sense separation, automatic thesaurus generation, information retrieval, anaphor resolution, text simplification, topic identification, spelling correction (Weeds, 2003).

In the following we will give an abridged overview of our method for monolingual word sense discovery. For the full version, see (Publication 4).

²<http://infomap.stanford.edu/>

³<http://websom.hut.fi/websom/>

⁴Dorow and Widdows (2003) also used lists of syntactically coordinated nouns to create a graph for visualizing the links between words and word clusters.

6.2.1 Corpus Data

Our corpus consists of nouns in a sentence context. We used all the nouns (in base form) that occurred more than 100 times (in any inflected form) in a corpus of Finnish newspaper text. The corpus contained 245 000 documents totaling 48 million words of the Finnish newspaper *Helsingin Sanomat* from 1995–1997. Excluding TV and radio listings, there were 196 000 documents with 42 million words. As corpus data we selected all the 17 835 nouns occurring more than 100 times comprising 14 million words of the corpus.

6.2.2 Feature Extraction

The present experiments aim at discovering the nouns that are most similar in meaning to a given noun. In order to determine the similarity of the syntactic contexts, we represent a word w as a probability distribution over a set of features a occurring in the context of w : $P(a|w)$. The context features a are the major class words w' (nouns, adjectives and verbs) with direct dependency links to the word w . The context feature is the word w' in base form labeled with the dependency relation r . For example, the noun might occur as an object of a verb and with an adjective modifier; both the verb and the adjective including their dependency relations are context features. We used Connexor's dependency parser FDG for Finnish (Connexor, 2002b) for parsing the corpus.

6.2.3 Similarity Calculations

If we assume that a word w is represented as a sum of its contexts and that we can calculate the similarities between such word representations, we get a list L_w of words with quantifications of how similar they are to w . Each similarity list L_w contains a mix of words related to the senses of the word w .

6.2.4 Clustering

When we wish to discover the potential senses of w by clustering, we are in this work only interested in the 100 most similar words in L_w with a similarity estimate for w . The other words are deemed to be too dissimilar to w to be relevant. We cluster the words related to w with standard algorithms, e.g., average-link clustering (Manning and Schütze, 1999). We also compare with a number of other clustering algorithms.

6.2.5 Evaluation

For evaluation purposes we collected 453 synonym sets of back translations from an online translation dictionary. For an example, see Table 2.1. Only synonym sets that had more than one word were interesting because the source word was automatically included. The average number of synonyms or back translations for these test words was 3.53 including the source word. In addition to the mechanical rating of the synonym content, we also manually classified the words of some cluster samples into synonymy, antonymy, hyperonymy, hyponymy, complementarity and other relations. For manual classification we used a sample of 50 of the evaluation words and another 50 words from all words. We evaluated the mechanically best-rated cluster for each word.

The evaluation was a simple overlap calculation with the gold standard generated from the intersection of back translations of three translation dictionaries. By counting the number of cluster words in a source word synonym set and dividing by the synonym set size, we get the recall R . By counting the number of source word synonyms in a cluster and dividing by the cluster size, we get the precision P . The best performance was achieved with average link clustering, which provided $R = 47$ and $P = 42$ with a cluster size of approximately 6 words.

In Table 6.1, we see a few sample clusters whose words we rated during manual evaluation. The results are shown in Table 6.2.

Word	alijäämä/deficit	maatalous/agriculture	tuki/aid
Synonymy	vaje/deficiency		avustus/subsidy
	vajaus/shortfall		apu/help
Antonymy	ylijäämä/surplus		
Complementarity		teollisuus/industry	
		vientiteollisuus/export industry	
		elintarviketeollisuus/food industry	
Hyperonymy		elinkeinoelämä/business	
		talouselämä/economy	
Hyponymy			rahoitus/financing

Table 6.1: Semantic relations of the cluster content of some sample words (English glosses added)

The manual evaluation agrees with the mechanical evaluation: the manual evaluation found a synonym content of 52 %, compared to the minimum synonym content of 42 % found by the mechanical evaluation. This means that the clusters actually contain a few more synonyms than those conservatively agreed on by the three translation dictionaries.

If we evaluate the sample of key clusters drawn from all the words in the test sample, we get a synonym content of 38 %. This figure is rather low, but can be explained by the fact that many of the words were compound nouns that had no synonyms, which is why the translation dictionaries either did not have them

Content Relations	Dictionary sample	All words sample
Synonymy	52 %	38 %
Antonymy	1 %	1 %
Complementarity	12 %	34 %
Hyperonymy	2 %	4 %
Hyponymy	1 %	3 %
Other	31 %	21 %
Total	100 %	100 %

Table 6.2: Manual evaluation of the percentage of different semantic relations in the cluster content in two different samples of 50 clusters each.

listed or contained no additional source word synonyms for them.

6.2.6 Test results

By using translation dictionaries for several languages we introduce a simple but efficient way to evaluate the synonym content of clusters when in our case a WordNet for Finnish is not available. The evaluation shows that 69 – 79 % of the words in the discovered clusters are useful for thesaurus construction.

6.3 Multilingual Discovery

Multilingual discovery of word senses is most commonly pursued in machine translation via text and sentence alignment of parallel texts, i.e., an alignment of the global context of word senses. Aligned texts serve as the basis for phrase and word alignment. Word alignment is the core of training statistical machine translation systems. However, parallel texts are scarce and in many cases non-existent, so there are on-going efforts to discover cross-lingual synonyms, i.e., word translations, from non-parallel corpora in the same domain in different languages.

If we have reason to believe that the meaning of a word is attached to the sound pattern of the word stem, we can simply try to find a word with a matching sound pattern, a spelling variant, in the other language, i.e., we can use the zero context of the word. Finding term translations as cross-lingual spelling variants *on the fly* is an important problem for CLIR as it circumvents the need for explicit translation dictionary updating.

CLIR is typically approached by automatically translating a query into the target language. For an overview of approaches to cross-lingual information retrieval, see Oard and Diekema (1998) and Grefenstette (1998). When automat-

ically translating the query, specialized terminology is often missing from the translation dictionary. The analysis of query properties by Pirkola and Järvelin (2001) has shown that proper nouns and technical terms often are prime keys in queries, and if not properly translated or transliterated, query performance may deteriorate significantly. As proper nouns often need no translation between languages with the same writing system, a trivial solution is to include the untranslatable keys as such into the target language query. However, technical terms in European languages often have common Greek or Latin roots, which allows for a more advanced solution using approximate string matching to find the word or words most similar to the source keys in the index of the target language text database (Pirkola et al., 2001), i.e., we use the zero context to discover the transliteration of the source word in a target language database index.

Between European languages, the loan words are often borrowed with minor language-specific modifications, which allows for initial testing of some straightforward approximate string matching methods. A comparison of methods applied to cross-lingual spelling variants in CLIR for a number of European languages is provided in Keskustalo et al. (2003). They compare exact match, simple edit distance, longest common subsequence, digrams, trigrams, tetragrams, as well as skipgrams, i.e., digrams with gaps. Such methods were initially tested in cross-lingual information retrieval by, e.g., Davis (1998) and Buckley et al. (1997), who both used the simple edit distance, for Spanish and French, respectively.

Between languages with different writing systems, foreign words are often borrowed based on phonetic rather than orthographic transliterations suggesting a phoneme-based rather than a grapheme-based approach. In Knight and Graehl (1998), a phoneme-based generative model is introduced which transliterates words from Japanese to English using weighted finite-state transducers. In Qu et al. (2003), this approach is successfully evaluated in an information retrieval context. This model uses context-free transliterations, which produces heavily overgenerating systems. Context-sensitivity requires more training data, but training data is less readily available for phoneme-based approaches, which lately have been rivaled by grapheme-based approaches, e.g., to Arabic (Al-Onaizan and Knight, 2002), Japanese (Ohtake et al., 2004; Bilac and Tanaka, 2004), and Chinese (Zhang et al., 2004). Until now, such models have included only one language pair.

In the following we will give an abridged overview of our method for multilingual word sense discovery based on string matching proposed in Lindén (2004). The method represents supervised learning requiring some training material, but the amount of training material is modest when adapting the method to a new language. For the full version, see (Publication 5).

6.3.1 Probabilistic Framework

Assume that we have a word in a foreign language. We call this the source word S . The word looks familiar and we want to know the possible meanings of this word in a language known to us, but we do not have a translation dictionary, only a word list L of the words in the known language. We take the word and compare it with all the words in L in order to determine which word is most similar to the unknown word. We call the most similar word the target word T . In the beginning we can only compare how many letters are similar or different, but having done so a few times, we start to learn the regularities involved where some letters are likely to be inserted, deleted or replaced with others. We also observe that the likelihood for insertion, deletion and replacement for each letter is different in different contexts.

A model for this procedure is to find, in the target word list L , the target word T which is most likely to convey the meaning of the source word S by accounting for all the sounds encoded by the source word letter sequence. To find the most likely target word for any given source word, we need to maximize the probability $P(T|S)$, i.e.

$$\arg \max_{T \in L} P(T|S) . \quad (6.1)$$

6.3.2 Data

The problem of finding terms for unknown words crops up in many different areas with quickly developing terminologies, e.g. medicine, economics, technology. As training data for our model we chose technical medical terminology in seven different languages as source words. Based on the English training data terms we found Finnish equivalents for 1480 of the medical terms as adaptation data for Finnish. To be able to compare the test results we used the test data created by Keskustalo et al. (2003) at the University of Tampere, cf. Section 2.6.3.

The target words consisted of a list containing all words of an English full-text database index of 84 000 documents (LAT, Los Angeles Times used in the CLEF 2000 experiments) (Peters, 2000). It contained around 189 000 unique word forms. The words were either in base form if recognized by the morphological analyzer ENGTWOL (Voutilainen et al., 1995) used in indexing, or in case of unrecognized word forms, the original words as such. All words were written in monospace.

6.3.3 Test results

Against the index of the LAT newspaper database we achieve 80-91 % precision at the point of 100 % recall for a set of medical terms in Danish, Dutch, French, German, Italian, Portuguese and Spanish. On the average, this is a relative improvement of 26 % on the simple edit distance baseline. Using the medical terms as training data, we achieve 64-78 % precision at the point of 100 % recall for a set of terms from varied domains in French, German, Italian, Spanish, Swedish and Finnish. On the average, this is a relative improvement of 23 % on the simple edit distance baseline. For Swedish there is no training data and for Finnish we need only a small amount of training data, approximately 10 % of the adaptation data, for adapting the model.

6.3.4 Extending Context

By only considering the transliteration candidates for which we have a matching translation of some general words in the local context, we are likely to further improve the precision of the method we suggested. In a Finnish text one might see *Helsingin yliopisto* several times and in an English text *Helsinki University*. In a Finnish-English dictionary one finds only *university* ↔ *yliopisto*. From the repeated partially overlapping translation contexts it is possible to conclude that *Helsingin* is a Finnish variant for the proper name *Helsinki*, even if it is not listed in a translation dictionary. Another way to use local context is suggested by Zhang and Vines (2004), who extract explicit translations or transliterations mentioned in related corpora, e.g., source language terms mentioned in brackets after the terms in the target language.

In order to retrieve corresponding contexts in non-parallel corpora it is to some extent also possible to use partial translations and then fill in the missing translations or significant related terms using pseudo-relevance feed-back. This has successfully been applied as pre- and post-translation pseudo-relevance feed-back to queries Ballesteros and Croft (1997) and to documents Levow (2003) in cross-lingual information retrieval (CLIR). This presupposes that the partial translation contains sufficiently specific information, otherwise the query may deteriorate making the pseudo-relevance feed-back find too general documents as implied by the findings of Pirkola and Järvelin (2001).

6.4 Summing up

We recapitulated some of the historical ideas in word sense discovery and lexical acquisition and then we presented our own contribution to mono- and multilingual

word sense discovery. We close this introduction to the publications with the conclusions and future work.

Chapter 7

Conclusion

“Would you tell me, please, which way I ought to go from here?”

“That depends a good deal on where you want to get to,” said the Cat.

“I don’t much care where—” said Alice.

“Then it doesn’t matter which way you go,” said the Cat.

“As long as I get somewhere,” Alice added as an explanation.

“Oh, you’re sure to do that,” said the Cat, “if you only walk long enough.”

— Lewis Carroll (1865)

Alice’s Adventures in Wonderland

7.1 Conclusion

This work supports Harris’ hypothesis about similar syntactic contexts conveying similar meaning. We have demonstrated that other kinds of context also convey similarity of meaning. The work proposes and tests three ways to exploit the hypothesis with practical consequences for creating thesauruses and translation dictionaries for, e.g., information retrieval and machine translation purposes.

We created a word sense disambiguation method based on a general context space (Publication 1). The method compared favorably with the best supervised methods in the SENSEVAL-2 exercise and was far better than the unsupervised methods (Publication 2). The sensitivity analysis (Publication 3) of the method with regard to various context features shows that the syntactic context is the most important factor for determining a good clustering of word senses. Syntactic context features gave sparse data, so the other context features served to smooth the model. This research provided a theoretical motivation for selecting the features that are best suited to create meaningful word clusters.

Using Finnish monolingual material, we found synonyms and other related words in an unsupervised way (Publication 4), based on dependency syntactic features. This is the first time the foundations of a thesaurus, i.e., the synonym

sets, have been automatically generated from raw text in Finnish. Further work is needed to relate the synonym sets to one another into a hierarchical thesaurus structure. Synonyms and other thesaurus relations are used in information retrieval for enhancing recall. The automatically generated word clusters can also be used as preprocessing for thesaurus and ontology building efforts.

Using multilingual material, we automatically created a multilingual transliterator (Publication 5), whose precision and recall by far exceeds any similar algorithm proposed previously. We have reason to believe that additional context would further enhance the precision of the transliterator. The transliterator was motivated by research in information retrieval showing that new terminology is seldom found in online dictionaries, technical terms are crucial for the search result, and they are often transliterated loan-words.

In this overview of the field we have tried to bring out the duality between the word sense disambiguation training and the word sense discovery procedures. Disambiguation training starts from the word senses and identifies the important context features, whereas discovery starts from the contexts and aims at identifying the word senses.

7.2 Future Directions

1. Word sense disambiguation has served as a test-bed for numerous approaches to machine learning in natural language processing. The disambiguation task has shown that syntactic preprocessing of the context is necessary for this task. Recently methods have been proposed for morphology (Creutz and Lagus, 2004; Yarowsky and Wicentowski, 2000) and syntax discovery (Klein, 2005; Klein and Manning, 2004). Such automatically generated morphologies and grammars may only partially resemble their hand-made counterparts, although they use linguistically annotated corpora as the gold standards for morphology (Creutz and Lindén, 2004) and for syntax (Marcus et al., 1994). It remains to be seen whether they are effective in word sense disambiguation and discovery.
2. Structured thesauruses are finding their place in natural language processing as general word sense repositories which draw their preference information from unannotated domain-specific corpora. Lexicons and the relations between words should no longer be seen as static entities, but as dynamic entities which adapt to the domain. A version of this idea has been explored by Lagus and Kurimo (2002) and Kurimo and Lagus (2002), where the language model of a speech recognition system dynamically adapts to the domain.

3. If lexicons become dynamic adaptive entities, it is quite possible that the right place for word sense disambiguation in natural language processing is to provide constraints on the lexicon with regard to a domain leaving the remaining ambiguities for taggers or parsers to resolve in a local context.
4. The references of proper nouns is an important practical problem in information retrieval (Pirkola, 1999). In search engines and question answering systems, word sense disambiguation would be useful if it could determine whether two terms or expressions refer to the same entity (Edmonds et al., 2004). Often part of one name is another name, e.g., New York City contains the names of a state, an airport, a baseball team, and a basketball team. Other practical sources of ambiguity are date and number references as well as abbreviations. These have not been the core issues of word sense disambiguation, even if their solution would be useful. For such purposes a static word sense inventory is hardly adequate: there are either too many or not enough sense distinctions.
5. Machine translation (MT) is often mentioned as the prime application of word sense disambiguation and discovery, but there is a fair amount of skepticism in the MT community (Edmonds et al., 2004). Commercial MT systems like Systran use rules, in which word senses are interlaced with everything else. However, word sense annotated data has not really been tested in MT, because no large scale annotators are available. It might be useful to look at errors of MT systems and use them as a basis for evaluating whether word sense disambiguation could have prevented them.
6. Applications in general often need information beyond the word level. They need the ability to identify and generate a complex string of words which carries more or less the same meaning as another string (Edmonds et al., 2004). This problem is related to machine translation. It means that we need to identify entities which are substitutable in a sentence with only minor changes to the meaning. If this is seen as compositionality of meaning, it presupposes that we know the meaning of the words or phrases involved.
7. The goal has often been to pursue a general purpose word sense disambiguator. If the long-term goal is natural language understanding, meanings have to be represented, but not necessarily with a given sense inventory (Edmonds et al., 2004). We may even prefer an indirect representation through contexts for which we can define similarity measures.
8. If the general trend in word sense disambiguation is toward indirect representation of word senses by word contexts, we will encounter many of the

important issues in word sense discovery, e.g., defining measures for the distributional similarity of words in context and defining criteria for semantic relatedness among word clusters.

Bibliography

- Abney, Steven P. 1997. Stochastic attribute-value grammars. *Computational Linguistics* 23(4):597–618.
- Abney, Steven P. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 360–367. Philadelphia, PA.
- Abney, Steven P. 2005. Statistical methods. In *Encyclopedia of Cognitive Science*. Nature Publishing Group, Macmillian.
- Agirre, Eneko and David Martinez. 2001. Knowledge sources for word sense disambiguation. In V. M. et al., ed., *TSD 2001, Proceedings of the International Conference on Text, Speech and Dialogue*, LNAI 2166, pages 1–10. Springer-Verlag, Berlin.
- Agirre, Eneko, German Rigau, Jordi Atserias, and Lluís Padró. 2000. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and Humanities* 34(1–2).
- Al-Onaizan, Yaser and Kevin Knight. 2002. Machine transliterations of names in arabic text. In *Proceedings of ACL Workshop on Computational Approaches to Semitic Languages*.
- Atkins, Beryl T. 1991. Admitting impediments. In U. Zernik, ed., *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 233–262. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Baayen, Harald R. 2001. *Word Frequency Distributions*, vol. 18 of *Text, Speech and Language Technology*. Dordrecht: Kluwer Academic Publishers.
- Ballesteros, Lisa and W. Bruce Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 84–91. New York, NY, USA: ACM Press. ISBN 0-89791-836-3.

- Banko, Michelle and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Toulouse, France.
- Berners-Lee, Tim. 1998. Semantic web road map. [<http://www.w3.org/DesignIssues/Semantic.html>].
- Berners-Lee, Tim. 2000. *Weaving the Web, The Original Design and Ultimate Destiny of the World Wide Web*. HarperCollins Publishers.
- Bilac, Slaven and Hozumi Tanaka. 2004. A hybrid back-transliteration system for Japanese. In *Proceedings of the 20th International Conference on Computational Linguistics, Coling 2004*, pages 597–603. Geneva, Switzerland.
- Blum, Avrim and Tom Mitchell. 1998. Combining labeled and unlabeled data with cotraining. In *COLT: Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers.
- Brown, Gavin, Jeremy Wyatt, Rachel Harris, and Xin Yao. 2004. Diversity creation methods: A survey and categorization. *Information Fusion Journal Special issue on Diversity in Multiple Classifier Systems*.
- Buckley, Chris, Mandar Mitra, Janet A. Walz, and Claire Cardie. 1997. Using clustering and superconcepts within smart: Trec 6. In *TREC*, pages 107–124.
- Burnard, Lou. 1995. *Users' Reference Guide for the British National Corpus, version 1.0*. Oxford University Computing Services, Oxford, UK.
- Burnard, Lou. 2002. Where did we go wrong? A retrospective look at the British National Corpus. In Ketterman, Bernhard, and Marko, eds., *Proceedings of the Fourth International TALC, Teaching and Learning by doing Corpus Analysis*, pages 51–70.
- Carlson, Lauri and Krister Lindén. 1987. Unification as a grammatical tool. *Nordic Journal of Linguistics* 10:111–136.
- Chklovski, Timothy and Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. In *Proceedings of the ACL 2002 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*. Philadelphia.
- Chomsky, Noam. 1957. *Syntactic structures*. Janua Linguarum Series Minor, Volume 4. The Hague, The Netherlands: Mouton de Gruyter.

- Christianini, Nello and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and other Kernel-based methods*. Cambridge University Press.
- Connexor. 2002a. Machine phrase tagger. [<http://www.connexor.com/>].
- Connexor. 2002b. Machine syntax. [<http://www.connexor.com/>].
- Cortes, Corinna, Patrick Haffner, and Mehryar Mohri. 2004. Rational kernels: Theory and algorithms. *Journal of Machine Learning Research* 5:1035–1062.
- Creutz, Mathias and Krista Lagus. 2004. Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51. Barcelona.
- Creutz, Mathias and Krister Lindén. 2004. Hutmegs – Helsinki University of Technology Morphology Evaluation Gold Standard. Neural Networks Research Center, Helsinki University of Technology, Publications in Computer and Information Science, Report A77, Espoo, Finland.
- Curran, James Richard. 2003. *From Distributional to Semantic Similarity*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.
- Daelemans, Walter and Véronique Hoste. 2002. Evaluation of machine learning methods for natural language processing tasks. In *Proceedings of the LREC-2002, the third International Conference on Language Resources and Evaluation*, pages 755–760. Las Palmas, Spain.
- Davis, Mark. 1998. On the effective use of large parallel corpora in cross-language text retrieval. In G. Grefenstette, ed., *Cross-Language Information Retrieval*, The Kluwer International Series on Information Retrieval, pages 12–22. Kluwer Academic Publishers.
- Delbridge, A., J. R. L. Bernard, D. Blair, W. S. Ramson, and Susan Butler, eds. 1987. *The Macquarie Dictionary*. Sydney, Australia: Macquarie Library.
- Dominic Widdows, Beate Dorow, Scott Cederberg. 2002. Visualisation techniques for analysing meaning. In *Appeared in Fifth International Conference on Text, Speech and Dialogue (TSD 5)*, pages 107–115. Brno, Czech Republic.
- Dorow, Beate and Dominic Widdows. 2003. Discovering corpus-specific word senses. In *EACL 2003, Conference Companion (research notes and demos)*, pages 79–82. Budapest, Hungary.

- Edmonds, Philip, Bill Dolan, Dekang Lin, German Rigau, and Carlos Strappavara. 2004. Senseval 3 panel on applications of wsd. SENSEVAL-3 Workshop, Barcelona, Spain.
- Edmonds, Philip and Adam Kilgarriff, eds. 2002. *Special Issue on Evaluating Word Sense Disambiguation Systems*, vol. 8(4) of *Natural Language Engineering*. Cambridge University Press.
- Fellbaum, Christiane, ed. 1998. *WordNet, An Electronic Lexical Database*. The MIT Press.
- Florian, Radu, Silviu Cucerzan, Charles Schafer, and David Yarowsky. 2002. Combining classifiers for word sense disambiguation. *Natural Language Engineering* 8(4):327–341.
- Florian, Radu and David Yarowsky. 2002. Modeling consensus: Classifier combination for word sense disambiguation. In *Proceedings of EMNLP-2002*, pages 25–32.
- Gale, Bill, Kenneth Church, and David Yarowsky. 1992a. Work on statistical methods for word sense disambiguation. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54–60. Cambridge, MA: AAAI Press.
- Gale, William A., Kenneth W. Church, and David Yarowsky. 1992b. One sense per discourse. In *Proceedings of the Speech and Natural Language Workshop*, pages 233–237. San Francisco: Morgan Kaufmann.
- Gale, William A., Kenneth W. Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 25(5):415–439.
- Gaustad, Tanja. 2001. Statistical corpus-based word sense disambiguation: Pseudowords vs. real ambiguous words. In *Companion Volume to the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001) – Proceedings of the Student Workshop*, pages 61–66. Toulouse, France.
- Gazdar, Gerald and Christopher Mellish. 1989. *Natural Language Processing in PROLOG: An Introduction to Computational Linguistics*. Great Britain, Kent: Addison Wesley.
- Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28(3):496–530.

- Gleitman, Lila R. 1990. The structural sources of verb meaning. *Language Acquisition* 1:3–55.
- Gleitman, Lila R. 2002. Verbs of a feather flock together II: The child's discovery of words and their meanings. In B. E. Nevin and S. B. Johnson, eds., *The Legacy of Zellig Harris: Language and information into the 21st century*, vol. 1: Philosophy of science, syntax and semantics of *Current Issues in Linguistic Theory*, pages 209–229. John Benjamins Publishing Company.
- Gleitman, Lila R., Kimberly Cassidy, Rebecca Nappa, Anna Papafragou, and John C. Trueswell. 2005. Hard words. *Language Learning and Development* 1(1):23–64.
- Goutte, Cyril. 1997. Note on free lunches and cross-validation. *Neural Computation* 9(6):1245–1249.
- Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers. ISBN 0792394682.
- Grefenstette, Gregory, ed. 1998. *Cross-Language Information Retrieval*. The Kluwer International Series on Information Retrieval. Kluwer Academic Publishers.
- Grefenstette, Gregory and Julien Nioche. 2000. Estimation of English and Non-English language use on the www. In *Proceedings of the RIAO (Recherche d'Informations Assistée par Ordinateur)*, pages 237–246. Paris, France.
- Harris, Zellig S. 1954. Distributional structure. *Word* 10:146–162.
- Harris, Zellig S. 1968. Mathematical structures of language. *Interscience Tracts in Pure and Applied Mathematics* 21(ix):230 pp.
- Hindle, Donald. 1990. Noun classification from predicate-argument structures. In *Meeting of the Association for Computational Linguistics*, pages 268–275.
- Honkela, Timo, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. 1996. News-group exploration with websom method and browsing interface. Tech. Rep. A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.
- Ide, Nancy and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics* 24(1):1–40. Special Issue on Word Sense Disambiguation.

- Inkpen, Diana Z. and Graeme Hirst. 2005. Building and using a lexical knowledge-base of near-synonym differences. [<http://ftp.cs.toronto.edu/pub/gh/Inkpen+Hirst-2005.pdf>] (submitted).
- Kanji, Gopal K. 1999. *100 Statistical Tests*. Sage Publications, new edition edn.
- Kaplan, Abraham. 1950. An experimental study of ambiguity and context. *Mimeographed* (Published 1955 in *Mechanical Translation*, 2(2):39–46.).
- Kaplan, Frédéric. 2001. *La Naissance d'une Langue chez les Robots*. Collection Technologies et Culture. Paris, France: Hermes Science Publications.
- Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In J. Bresnan, ed., *The Mental Representation of Grammatical Relations*, pages 173–281. Cambridge, MA: MIT Press.
- Karlsson, Fred. 1990. Constraint grammar as a framework for parsing unrestricted text. In H. Karlgren, ed., *Proceedings of the 13th International Conference of Computational Linguistics*, vol. 3, pages 168–173. Helsinki, Finland.
- Kaski, Samuel, Janne Nikkil, Janne Sinkkonen, Leo Lahti, Juha Knuuttila, and Christophe Roos. 2005a. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* .
- Kaski, Samuel, Janne Sinkkonen, and Arto Klami. 2005b. Discriminative clustering. *Neurocomputing* .
- Kearns, Michael. 1996. A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds., *Advances in Neural Information Processing Systems*, vol. 8, pages 183–189. The MIT Press.
- Keskustalo, Heikki, Ari Pirkola, Kari Visala, Erkkä Leppänen, and Kalervo Järvelin. 2003. Non-adjacent digrams improve matching of cross-lingual spelling variants. In *SPIRE 2003 - 10th International Symposium on String Processing and Information Retrieval*. Manaus, Brazil.
- Kielikone. 2004. Dictionary service mot - dictionaries and terminologies. [<http://www.kielikone.fi/en/>].
- Kilgarrieff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* 31(2):91–113.

- Kilgarriff, Adam. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language* 12(4):453–472. Special Issue on Evaluation edited by R. Gaizauskas.
- Kilgarriff, Adam. 2001. Generative lexicon meets corpus data: the case of non-standard word uses. In P. Bouillon and F. Busa, eds., *The Language of Word Meaning*, pages 312–328. Cambridge: Cambridge University Press.
- Kilgarriff, Adam. 2003a. Linguistic search engine. In K. Simov, ed., *Shallow Processing of Large Corpora: Workshop Held in Association with Corpus Linguistics 2003*. Lancaster, England.
- Kilgarriff, Adam. 2003b. Thesauruses for natural language processing. In *Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering*. Beijing: Beijing Media Center.
- Kilgarriff, Adam. 2003c. What computers can and cannot do for lexicography, or Us precision, them recall. Tech. Rep. ITRI-03-16, Information Technology Research Institute, University of Brighton. Also published in Proceedings of ASIALEX.
- Kilgarriff, Adam. 2004. How dominant is the commonest sense of a word? In P. Sojka, I. Kopeček, and K. Pala, eds., *Proceedings of TSD 2004, Text, Speech and Dialogue 7th International Conference*, vol. 2448 of LNAI, pages 1–9. Brno, Czech Republic: Springer-Verlag, Berlin.
- Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the special issue on web as corpus. *Computational Linguistics* 29(3):1–15.
- Kilgarriff, Adam and Martha Palmer, eds. 2000. *Special Issue on SENSEVAL*, vol. 34(1–2) of *Journal Computers and the Humanities*. Kluwer Academic Publishers.
- Kilgarriff, Adam and Joseph Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities* 34(1–2):15–48. Special Issue on SENSEVAL edited by A. Kilgarriff and M. Palmer.
- Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of Euralex*, pages 105–116. Lorient, France.
- Kilgarriff, Adam and David Tugwell. 2001. Word sketch: Extraction and display of significant collocations for lexicography. In *COLLOCATION: Computational Extraction, Analysis and Exploitation*. ACL Workshop, Toulouse.

- Kilgarriff, Adam and Colin Yallop. 2000. What's in a thesaurus? In *Proceedings of LREC 2000, the 2nd International Conference on Language Resources and Evaluation*, pages 1371–1379. Athens.
- Klein, Dan. 2005. *The Unsupervised Learning of Natural Language Structure*. Ph.D. thesis, Stanford University.
- Klein, Dan and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the ACL*.
- Knight, Kevin and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics* 24(4):599–612.
- Kohonen, Teuvo. 1997. *Self-Organizing Maps (Second Edition)*, vol. 30 of *Springer Series in Information Sciences*. Berlin: Springer.
- Kohonen, Teuvo, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Vesa Paatero, and Antti Saarela. 2000. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery* 11(3):574–585.
- Koskenniemi, Kimmo. 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Publications, University of Helsinki, Department of General Linguistics, Helsinki, Finland.
- Kurimo, Mikko and Krista Lagus. 2002. An efficiently focusing large vocabulary language model. In *International Conference on Artificial Neural Networks (ICANN'02)*, pages 1068–1073. Madrid, Spain.
- Lagus, Krista and Samuel Kaski. 1999. Keyword selection method for characterizing text document maps. In *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*, vol. 1, pages 371–376. IEE, London.
- Lagus, Krista and Mikko Kurimo. 2002. Language model adaptation in speech recognition using document maps. In *IEEE Workshop on Neural Networks for Signal Processing (NNSP'02)*, pages 627–636. Martigny, Switzerland.
- Lapata, Mirella and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics* 30(1):45–75.
- Leacock, Claudia, Martin Chodorow, and George A. Miller. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics* 24(1):147–165. Special Issue on Word Sense Disambiguation.

- Lee, Lillian. 1999. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.
- Lee, Yoong Keok and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of EMNLP-2002*, pages 41–47.
- Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: University of Chicago Press.
- Levow, Gina-Anne. 2003. Issues in pre- and post-translation document expansion: Untranslatable cognates and missegmented words. In *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages*, pages 77–83.
- Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL98*. Montreal, Canada.
- Lindén, Krister. 1993. Finnish language technology. In *Life and Education in Finland*. Helsinki, Finland.
- Lindén, Krister. 1997. Language applications with finite-state technology. *International Journal of Corpus Linguistics* 2(2):281–288.
- Lindén, Krister. 2004. Finding cross-lingual spelling variants. In *Proceedings of SPIRE 2004, the 11th Symposium on String Processing and Information Retrieval*. Padua, Italy.
- Lowe, Will. 1997. Semantic representation and priming in a self-organizing lexicon. In J. A. Bullinaria, D. W. Glasspool, and G. Houghton, eds., *Proceedings of the Fourth Neural Computation and Psychology Workshop: Connectionist Representations*, pages 227–239. London: Springer-Verlag.
- Lowe, Will. 2001. Towards a theory of semantic space. In J. D. Moore and K. Stenning, eds., *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 576–581. Mahwah NJ: Lawrence Erlbaum Associates.
- Magnini, Bernardo, Carlo Strappavara, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering* 8(4):359–373.
- Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

- Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics* 19(2):313–330.
- Martinetz, Thomas and Klaus Schulten. 1994. Topology representing networks. *Neural Networks* 7(3):507–522.
- Martinez, David and Eneko Agirre. 2000. One sense per collocation and genre/topic variations. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong.
- Masterman, Margaret. 1956. The potentialities of a mechanical thesaurus. *Mechanical Translation* .
- Masterman, Margaret. 1962. Semantic message detection for machine translation, using an interlingua. In *1961 International Conference on Machine Translation of Languages and Applied Language Analysis*, pages 437–475. London: Her Majesty's Stationery Office.
- McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287. Barcelona, Spain.
- McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. August 2004. Automatic identification of infrequent word senses. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, pages 1220–1226. Geneva, Switzerland.
- Mihalcea, Rada. 2004a. Senseval – evaluation exercises for word sense disambiguation, organized by acl-siglex. [<http://www.senseval.org/>]. Hosted by University of North Texas.
- Mihalcea, Rada. 2004b. Software and data sets – semcor. [<http://www.cs.unt.edu/rada/downloads.html#semcor>].
- Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff. 2004. The SENSEVAL-3 English lexical sample task. In *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28. Barcelona, Spain.
- Mihalcea, Rada and Philip Edmonds, eds. 2004. *Evaluation exercises for Word Sense Disambiguation*, Barcelona, Spain. ACL-SIGLEX, University of North Texas. [<http://www.senseval.org/>].

- Miller, George A., Christiane Fellbaum, Randee Teng, Susanne Wolff, Pamela Wakefield, Helen Langone, and Benjamin Haskell. 2003. Wordnet – a lexical database for the English language. [<http://www.cogsci.princeton.edu/~wn/index.shtml>].
- Mohammad, Saif and Graeme Hirst. 2005. Distributional measures as proxies for semantic relatedness. [<http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf>] (submitted).
- Mohri, Mehryar. 2003. Learning from uncertain data. In *Proceedings of The 16th Annual Conference on Computational Learning Theory (COLT 2003)*, vol. 2777 of *Lecture Notes in Computer Science*, pages 656–670. Washington D.C.: Springer, Heidelberg, Germany.
- Mohri, Mehryar, Fernando C. N. Pereira, and Michael D. Riley. 2003. At&t fsm library – finite-state machine library. [<http://www.research.att.com/sw/tools/fsm/>].
- Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In A. Joshi and M. Palmer, eds., *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 40–47. San Francisco: Morgan Kaufmann Publishers.
- Nienstedt, Walter. 2003. Tohtori.fi – lääkärikirja. [<http://www.tohtori.fi/laakarikirja/>].
- Nigam, Kamal, Andrew K. McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine Learning* 39(2–3):103–134.
- Oard, Doug and A. Diekema. 1998. Cross language information retrieval. In *Annual Review of Information Science and Technology*, vol. 33, pages 223–256.
- Ohtake, Kiyonori, Youichi Sekiguchi, and Kazuhide Yamamoto. 2004. Detecting transliterated orthographic variants via two similarity metrics. In *Proceedings of the 20th International Conference on Computational Linguistics, Coling 2004*, pages 709–715. Geneva, Switzerland.
- Pantel, Patrick André. 2003. *Clustering by Committee*. Ph.D. thesis, University of Alberta, Edmonton, Alberta, Canada.
- Pedersen, Ted. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*. Pittsburgh, PA.

- Pereira, Fernando, Naftali Z. Tishby, and Lillian Lee. 1993. Distributional clustering of english words. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 183–190. Ohio State University, Columbus, Ohio: Association for Computational Linguistics.
- Peters, Carol. 2000. Cross language evaluation forum. [<http://clef.iei.pi.cnr.it/>].
- Pimsleur, P. 1957. Semantic frequency counts. *Mechanical Translation* 4(1–2):11–13.
- Pirkola, Ari. 1999. *Studies on Linguistic Problems and Methods in Text Retrieval*. Ph.D. thesis, Tampere University, Tampere.
- Pirkola, Ari, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval* 4(3/4):209–230.
- Pirkola, Ari and Kalervo Järvelin. 2001. Employing the resolution power of search keys. *Journal of the American Society of Information Science* 52(7):575–583.
- Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.
- Pulkki, Ville. 1995. Data averaging inside categories with the self-organizing map. Helsinki University of Technology, Faculty of Information Technology, Laboratory of Computer and Information Science, Report A27, Otaniemi, Finland.
- Pustejovsky, James. 1998. *The Generative Lexicon*. The MIT Press.
- Qu, Yan, Gregory Grefenstette, and David A. Evans. 2003. Automatic transliteration for Japanese-to-English text retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–360. ACM Press. ISBN 1-58113-646-3.
- Qu, Yan, James Shanahan, and Janyce Wiebe, eds. 2004. *Exploring Attitude and Affect in Text: Theories and Applications*, vol. SS-04-07 of *AAAI Spring Symposium*, Palo Alto, CA. Stanford University, AAAI Press.
- Ramakrishnan, Ganesh, B. Prithviraj, and Pushpak Bhattacharyya. 2004. A gloss-centered algorithm for disambiguation. In R. Mihalcea and P. Edmonds, eds., *Proceedings of SENSEVAL-3, Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Evaluation exercises for Word Sense Disambiguation. ACL-SIGLEX, Barcelona, Spain: University of North Texas.

- Rapaport, William J. 2005. A (partial) bibliography (in chronological order) of (computational) theories of contextual vocabulary acquisition. [<http://www.cse.buffalo.edu/rapaport/refs-vocab.html>].
- Reifler, Erwin. 1955. The mechanical determination of meaning. In W. N. Locke and A. D. Booth, eds., *Machine Translation of Languages*, pages 136–164. New York: John Wiley & Sons.
- Resnik, Philip and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering* 5(3):113–133.
- Richens, Richard H. 1958. Interlingual machine translation. *Computer Journal* 1(3):144–147.
- Ritter, Helge and Teuvo Kohonen. 1989. Self-organizing semantic maps. *Biological Cybernetics* 61:241–254.
- Roget, Peter Mark. 1987. *Roget's Thesaurus*. Longman, Longman Edition edit by Betty Kirkpatrick edn. Original edition 1852.
- Salton, Gerard. 1968. *Automatic Information Organization and Retrieval*. New York: McGraw-Hill.
- Samuelsson, Christer. 2000. A statistical theory of dependency syntax. In *Proceedings of COLING-2000. ICCL.*
- Samuelsson, Christer and Atro Voutilainen. 1997. Comparing a linguistic and a stochastic tagger. In *Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the EACL*, pages 246–253. ACL, Madrid, Spain.
- Sanderson, Mark. 2000. Retrieving with good sense. *Information Retrieval* 2(1):49–69.
- Schütze, Hinrich. 1992. Context space. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 113–120. Cambridge, MA: AAAI Press.
- Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1):97–123. Special Issue on Word Sense Disambiguation.
- SENSEVAL-2. 2001. Training and testing corpora. [<http://www.cis.upenn.edu/~cotton/senseval/corpora.tgz>].

- Sinclair, John, ed. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London and Glasgow: Collins ELT.
- Somes, G. 1983. McNemar test. In S. Kotz and N. Johnson, eds., *Encyclopedia of statistical sciences*, vol. 5, pages 361–363. New York: Wiley.
- Spärck Jones, Karen. 1986. *Synonymy and Semantic Classification*. Edinburgh University Press. ISBN 0-85224-517-3. Originally published in 1964.
- Steyvers, Mark and Josh B. Tenenbaum. 2005. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science* 29(1).
- Stichele, R. Vander. 1995. Multilingual glossary of technical and popular medical terms in nine European languages. Tech. rep., Heymans Institute of Pharmacology, University of Gent and Mercator College, Department of Applied Linguistics. [<http://allserv.rug.ac.be/~rvdstich/eugloss/welcome.html>].
- Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, Danil M. Ogilvie, and associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.
- Tapanainen, Pasi and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of 5th Conference on Applied Natural Language Processing*, pages 64–71.
- Tishby, Naftali Z., Fernando Pereira, and William Bialek. 1999. The information bottleneck method. In B. Hajek and R. S. Sreenivas, eds., *Proceedings of the 37th Allerton Conference on Communication, Control and Computing*. University of Illinois, Urbana, Illinois.
- Towell, Geoffrey and Ellen M. Voorhees. 1998. Disambiguating highly ambiguous words. *Computational Linguistics* 24(1):125–146. Special Issue on Word Sense Disambiguation.
- Voorhees, Ellen. 2000. Text REtrieval Conference (TREC). [<http://trec.nist.gov/>].
- Voorhees, Ellen M., Claudia Leacock, and Geoffrey Towell. 1995. Learning context to disambiguate word senses. In T. Petsche, S.J.Hanson, and J. Shavlik, eds., *Computational Learning Theory and Natural Language Learning Systems 3: Selecting Good Models*, pages 279–305. Cambridge: MIT Press.
- Vossen, Piek. 2001. Eurowordnet. Tech. rep., University of Amsterdam, Department of Computational Linguistics. [<http://www.hum.uva.nl/~ewn/>].

- Vossen, Piek and Christiane Fellbaum. 2004. *WordNets in the World*. Global WordNet Association, [<http://www.globalwordnet.org/>].
- Voutilainen, Aro. 1995. A syntax-based part of speech analyser. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 157–164. Dublin.
- Voutilainen, Aro. 1997. Engcg tagger, version 2. In T. Brondsted and I. Lytje, eds., *Sprog og Multimedier*. Aalborg: Aalborg Universitetsforlag.
- Voutilainen, Aro, Juha Heikkilä, and Timo Järvinen. 1995. Engtwol: English morphological analyzer. [<http://www.lingsoft.fi/cgi-bin/engtwol>].
- Weaver, Warren. 1955. Translation. In W. N. Locke and A. D. Booth, eds., *Machine Translation of Languages*, pages 15–23. New York: Wiley & Sons. (Reprint of mimeographed version, 1949.).
- Weeds, Julie Elisabeth. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, University of Sussex.
- Wilks, Yorik. 1998. Language processing and the thesaurus. In *Proceedings of the National Language Research Institute*. Tokyo, Japan.
- Wilks, Yorick and Roberta Catizone. 2000. Can we make information extraction more adaptive? Tech. rep., The University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield, UK. [<http://www.dcs.shef.ac.uk/yorick/papers/doc/doc.html>].
- Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell. Translated by G. E. M. Anscombe.
- Yarowsky, David. 1995. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 189–196. Cambridge, MA.
- Yarowsky, David and Radu Florian. 2002. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering* 8(4):293–310.
- Yarowsky, David and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, pages 207–216.
- Zhang, Min, Haizhou Li, and Jian Su. 2004. Direct orthographical mapping for machine transliteration. In *Proceedings of the 20th International Conference on Computational Linguistics, Coling 2004*, pages 716–722. Geneva, Switzerland.

Zhang, Y. and P. Vines. 2004. Using the web for automated translation extraction in cross-language information retrieval. In *SIGIR 2004*, pages 162–169. Sheffield, United Kingdom: ACM.

Zipf, George Kingsley. 1935. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Boston, USA: Houghton Mifflin.

*“Differences of habit and language are nothing at all
if our aims are identical and our hearts are open,”
professor Albus Dumbledore said thoughtfully.*

— J.K. Rowling (2000)
Harry Potter and the Goblet of Fire
“The Beginning”