Chapter 4:Word sense disambiguation

30-47

4.1 APPROACHES OF WORD SENSE DISAMBIGUATION

Various approaches to WSD are often classified according to source of knowledge used in sense differentiation following are the techniques for word sense disambiguation

A. **Knowledge Based Method**

It represents the distinct category in word sense disambiguation. These methods use lexical knowledge bases such as dictionary[23] and thesauri and extract knowledge from word definitions and relation among word
and senses . Four main types of knowledge based method are as follows

1) **The Lesk Algorithm :** in this it computes the overlaps between words that are the number of words in common between the definitions of senses. For example consider the task of disambiguating the words pine and cone. The oxford Advanced Learners Dictionary defines four senses for pine and three senses for cone which are as follows

Pine:-
a) Seven kinds of evergreen tree with needle shaped leaves
b) Pine
c) Waste away through sorrow or illness
d) Pine for something, pine to do something
Cone:-
a) solid body which narrows to a point
b) something of this shape, whether solid or hollows
c) fruit of certain evergreen trees(fir, pine)

Here the third definition of cone and first definition of pine have the largest overlap among all sense combination with three words evergreen, tree and pine in common.

2) Measure of semantic similarity computed over semantic network : this include the methods for finding the semantic distance between concepts. Depending on size of context they are divided into two categories one is method applicable to a local context[24] and second method applicable to global context. Semantic similarity is the most

powerful constraint used in automatic word sense disambiguation. So by finding the smallest semantic distance we can find the appropriate senses of the word in given context.

3) Heuristic method: it consist of simple rules that can reliably assign sense to certain word categories which includes Most frequent sense, one sense per collocation and One sense per discourse

a) Most Frequent Sense: in general it is found that one meaning of word occurs more often than other meaning. So by using the very simple disambiguation method with the help of word frequency data we can assign to each word its most frequent meaning.

b) One Sense Per Discourse: according to Gale word tends to preserve its meaning across all its occurrences in a given discourse. If the meaning is identified in at least one such occurrence then it allows for the automatic disambiguation of all instances

c) One Sense Per Collocation: it states that a word tends to preserve its meaning when used in same collocation[25]. Nearby words provide strong and consistent clues to the sense of a target word.

4) Automatically or semi automatically acquired: -

A. Unsupervised corpus based method One common sense of “unsupervised” literary means “not supervised” and which includes any method that does not use supervised learning. In this approach it bootstrap from small number of sense tagged training example and

use that to build a simple model which tags a few more context. This method does not rely on external knowledge source such as machine readable dictionaries, concept hierarchies, or sense tagged text. They discriminate among word meaning based on information found in unannotated corpora and do not assign sense tags to words.

B. Supervised corpus based method:

It consist of training phase and testing phase .the training phase requires a sense annotated training corpus from which syntactic and semantic features are extracted to build a classifier .and in testing phase classifier picks the best sense of a word on the basis of surrounding words.

Regarding automatic word sense disambiguation one of the most successful approaches in recent years is supervised learning from example in which classification models are induced from semantically annotated corpora. The supervised approach to word sense disambiguation uses semantically annotated corpora to train machine learning algorithm to decide which sense to choose in which context.

C. Semi supervised method:

An important issue of supervised method is the knowledge acquisition bottleneck. It is difficult to find the required minimum number of occurrences per each sense of word. To overcome the knowledge acquisition bottleneck[26] faced by supervised method semi supervised method uses small annotated corpus.

4.2 APPLICATION OF WORD SENSE DISAMBIGUATION

There are various application of word sense disambiguation some of which are as follows

1) Machine translation: Word Sense Disambiguation is required in machine translation for words that have different translations for different senses. For example whenever we translate any sentence from

English to Hindi language it should give correct meaning. For example consider a sentence curiosity kills the cat. Here the meaning of this sentence if translated in Hindi gives different meaning because here the words have different senses in different context. If we take literal meaning of the word from one language then it translates it to another language then sometimes the translated sentence does not give the same meaning as of original language[27].

2) Information Extraction's is required for the accurate analysis of text in many applications. Word Sense Disambiguation is also used in text mining.

3) Information Retrieval: ambiguity must be resolved in some queries for example the noun "cricket" then what information should system retrieve because cricket is a insect as well as it is a sport so WSD plays very important role in information retrieval. Most of search engines do not use explicit semantics to prune out documents which are not related to user query. An accurate disambiguation of document base with possible disambiguation of the query words would allow it to eliminate document containing the same words used in different meaning and to retrieve document expressing the same meaning with different wordings

a) Cross language information retrieval: in cross language information retrieval the user presents a query of the usual form but some of the document s may be written in a different language. For example user types query in English language and wants document in German language, here it needs to convert that English query in German language and then retrieve the document. Here the problem of ambiguity comes which is why word sense disambiguation is use.

b) Question Answering: Question Answering[28] is the oldest natural language processing application. In this system specific questions are asked for example "when the computers are invented?" and it receives the concise answer rather than a set of relevant document. To retrieve concise pages we need to have correct sense of a particular word so word sense disambiguation plays a very important role.

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[29] 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)[30]. 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.

4.3 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[31] 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.

4.4 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[31], 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.

4.5 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[32] 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)[33], 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

4.6 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.

4.7 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[34] 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[35] 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 under specification 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.

4.8 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)[36], 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.

4.9The 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 Wittgenstein[37] an 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 linked 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[38], 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 mechanism.

A well-designed linguistically motivated GL with a trainable preference learning 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.

4.10 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)[39] 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 Lind´en (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 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. ff (2004) are that a large-scale

4.11 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[40], with unassailable 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 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 average, the proportion of the dominant sense therefore increases with n simply because the proportion of the first sense compared with that of the additional sense, 1/n , increases with . 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 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[41] 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.

4.12 Acceptability in Context

As soon as we start limiting the 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[42] is that the former deals 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 idiom hood 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)[43]. 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 epiphenomenon[44]. The valuable part is the context 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.

4.13 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) and the ongoing effort to provide samples for the sense inventory of WordNet by tagging corpora for the SENSEVAL (Mihalcea, 2004a)[45] 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.

a) 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 Language Target word Back translation

English deficit vaje, vajaus, alij ¨a¨am¨a; tilivajaus

shortfall vaje, alij ¨a¨am¨a

German Defizit vajaus, vaje, alij ¨a¨am¨a; kassavajaus, tappio;

tilivajaus; puutos, puute

Unterbilanz alij ¨a¨am¨a, vajaus, vaje, kauppavaje

Fehlbetrag vajaus, alij ¨a¨am¨a, tappio, virhemaksu

French d´eficit alij ¨a¨am¨a, miinus, tilivajaus; vajaus, vaje; tappio

Table 4.1: Translations of the Finnish source word alij ¨a¨am¨a into English, German and French with the back translations into Finnish. The shared back translations vaje, vajaus, alij¨a¨am¨a, tilivajaus are highlighted. morphology, topical clustering, and disambiguation of terms in glosses (Miller et al., 2003).

b) 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[46]. 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.

c)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)[47]. 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 – L¨a¨ak¨arikirja (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)