Exploiting translations for semantic annotation

Mona Diab

mdiab@umiacs.umd.edu

Ling895: PhD Candidacy Paper

Fall 2000

Advisor

Dr. Philip Resnik

Committee members
Dr. Amy Weinberg
Dr. Paul Pietrosky

Linguistics Department
Marie Mount Hall
University of Maryland, College park, MD 20742

Abstract

Large amounts of semantically annotated data, in more than one language, would provide a valuable resource for investigating issues of meaning representation cross linguistically. Unfortunately, such resources are currently unavailable due to the tremendous overhead of creating such data. Researchers in the computational linguistics community have proposed automated methods to approximate the human effort in sense annotating language. Yet, to date, all the proposed techniques deal with the one language at a time. This paper investigates a novel automated approach toward sense annotating two languages simultaneously. The proposed method is evaluated and compared against state of the art automated approaches. The yielded results are very promising. This research serves as an important stepping-stone for exploring lexicalization patterns and meaning representation issues cross linguistically.

Table of contents

I. Introduction

> "Many words have more than one meaning. When a person understands a sentence with an ambiguous word in it, that understanding is built on the basis of just one of the meanings…" [Kilgarriff, 1997].

Lexical ambiguity permeates language. The majority of words in natural language are polysemous. A word is polysemous if it has more than one meaning. Statements such as *'I walked by the bank'* are considered truly ambiguous for the listener/reader because of the ambiguous word *bank*, which may refer to either a *river bank* or *financial institution*. The listener usually resorts to wider contexts, such as conversational background or extralinguistic clues to determine the intended sense of *bank*. Similarly, the reader refers to the entire context in which the sentence occurs.

There are different types of polysemy. There are two main distinctions often referred to in the literature: **ambiguity** and **vagueness**[1] [Cruse, 1995; Dyvik, 1998; Kilgarriff, 1997; etc.]. Ambiguity is defined as word meanings that happen to share the same orthographic form, albeit arbitrarily, or etymologically. For example, *bank* is an ambiguous word since a *river bank* has little to do with money-saving unless historically people saved their money in river banks[2]. On the other hand, vagueness refers to senses of a word that are closely related to one another, for instance, a *newspaper* may have a *publication* sense, or a *building* sense or yet an *organization* sense. Researchers have devised linguistic as well as psycholinguistic tests to show that this fundamental difference exists in polysemy. Consider the sentence *'The newspaper costs 25 cents and fired its editor in chief'*. Native speakers of English consider it anomalous. It is a marked sentence because it appears to be conjoining two different senses of *newspaper*, the *publication* sense and the

---

[1] *Also, referred to as homonymy and polysemy in correspondence to ambiguity and vagueness.*

[2] *It is worth noting that the word for bank as a financial institution in modern standard Arabic is miSraf which also refers to a stream. It is possible that the two concepts related since the first is a place where money flows while the second is a place where water flows. Water and money are highly related sociologically*

organization sense. Moreover, in psycholinguistic studies, subjects, when given words in isolation, took longer times in lexical decision experiments with vague words than with ambiguous words using two different input modalities: visual and auditory [Rodd et al., 2000]. In addition to highlighting the different types of polysemy, these tests prove that words are represented with/as their senses in our mental lexicon.

1.1. General interest

The question of how senses are represented in the human mind has been the subject of extensive debate in psychology, computational linguistics and lexical semantics. The debate is, primarily, between two camps: an **enumerative** lexicon view of our mental model and a **generative** lexicon view.

The enumerative view advocates listing words with all their associated senses. The representation is static, as it requires updates whenever new meanings are introduced in language. Enumerative lexicon representations are usually silent when it comes to extensional uses of words in idiomatic expressions or metaphoric constructions. Traditional dictionaries and Machine Readable Dictionaries (MRD) are good representatives of an enumerative lexicon view.

On the other hand, the generative lexicon camp calls for an underspecified representation of words and provides the words with a generative capacity. The main idea is to represent words with minimal core **characteristics**. The generative lexicon model defines mechanisms for meaning extension based on the specified core characteristics depending on the appropriate contexts. Such models are argued for from a cognitive plausibility perspective. The words are represented underspecified, therefore, the model is flexible enough to deal with new meanings acquired on a frequent basis. Given an optimal set of characteristics for a word and the appropriate mechanisms for generation, the lexicon is rendered dynamic to process literal as well as extensional (idiomatic and metaphoric) senses of words [Hanks 2000; Pustejovsky, 1995; etc.]. Unfortunately, as appealing as it may sound, it has proven to be very challenging deciding which characteristics constitute

the core ones for the majority of words in language. To date, only a handful of words have been represented in a generative framework.

Researching ways in which a hybrid of both views can be attained is a valuable area of investigation. An integration of both camps would envelop the scalability of the enumerative lexicon and the generative capacity specified in the generative lexicon model, where entries in such a lexicon are organized based on the shared core characteristics. As mentioned before, deciding which characteristics are salient for a word poses a serious challenge. Questions such as *"What is the scope of a characteristic?"* or *"How does a word characteristic relate to the notion of a word sense/usage*[3]*?"* become relevant questions that require addressing. In addition to everyday words' usages in language, lexicographers decide the appropriate distinctive senses for dictionary entries based on an introspective consideration of these characteristics.

1.2. Current Goal

In this study, a sense is assumed the bearer of specific word characteristics. A word characteristic is defined as a generic concept that comprises semantic attributes, such as animate, inanimate, edible, etc. In case of sense ambiguity, senses have unique characteristics that distinguish them from other senses of the same word. For example, the *money-dealing* sense of *bank* bears the characteristic *financial institution*, while *river bank* sense bears the characteristic *edge of water* or *geological formation*. Accordingly, the two senses are ambiguous because the salient characteristic that sets them apart is not common. On the other hand, vague senses tend to be harder to deal with especially since they cross over into the realm of pragmatics, with metonymic usages playing a significant role in extending meanings. In vague cases, senses may share the same characteristics. For instance, *newspaper* has two vague senses: *publication* sense and *organization* sense, the core characteristic could be expressed as *reading product*.

---

[3] *A word 'usage' is a functional description of a word's usage in text. The term has less of an ontological commitment than a word 'sense'.*

From a linguistic perspective, exploring word senses and the relation between them is an interesting area of research. Cross-linguistic evidence for word senses and word meaning representations constitutes a very rich source of lexical information. The assumption is that core meaning characteristics are shared[4] among languages. These characteristics comprise semantic attributes that are reflected on a thematic level in the human mind. These semantic attributes can be thought of as primitives, which play a defining role in deciding thematic role assignment in argument structure. For example, the agentive thematic role is canonically associated with an animate semantic attribute in all languages. A characteristic may have more than one semantic attribute simultaneously. Translation of polysemous words brings the salient characteristics borne by the different senses to the foreground. For illustration, if the word *bank* is translated into *rive* in French, then implicitly, the *edge of the water* characteristic for *bank* becomes more salient than the *financial institution* characteristic, hence the choice of the appropriate sense *river bank* should be made, since it is the bearer of the appropriate characteristic. Moreover, given translations, it is interesting to explore the similarities and divergences between the patterns of meaning expressions in different languages. Large amounts of data that is sense annotated in several languages, would avail the linguistic community of a testbed to investigate these variations cross linguistically and to explore its consequences on our understanding of the structure of our mental lexicon.

Attaining large amounts of sense annotated or characteristic annotated data[5] is extremely time consuming and laborious, and requires trained specialized linguists and lexicographers. On the other hand, translators implicitly invoke these characteristics when they make lexical choices for a polysemous word in a language. Observing large amounts of data in translation, one can automatically derive the appropriate salient semantic characteristics by taking advantage of the translator's choice of lexical items in his/her repertoire. Unfortunately, sense annotated corpora are not available as of yet for any language, let alone texts and their translations.

---

[4] *It is not clear what the impact of cultural aspects is since they will affect to a certain degree the diversity in senses*

[5] *sense annotation is more coarse grained than characteristic annotation since senses may comprise more than one characteristic.*

Consequently, an automated data-driven approach is proposed to sense annotate words in two languages, simultaneously. The approach relies on exploiting translations as a source of sense distinction, thereby, obtaining large amounts of sense annotated data for two languages. Therefore, the main hypothesis is that instances of word translations bring to the foreground salient and defining characteristics for polysemous words. For example, the translation of *log* in the sentence *'please carry the log'* is *rondin*, in French. The translator's choice of *rondin* brings the *piece of wood* sense of the polysemous word *log* to the foreground. The choice of *rondin* as the translation suppresses other possible senses of *log* such as the *journal* sense or the *logarithm* sense. Given the verb *carry*, and the surrounding context in which the sentence was used, the translator picks the translation that highlights the appropriate semantic characteristic. It is worth noting that annotating polysemous words with their appropriate senses is an approximation toward the goal of identifying the salient characteristics that distinguish the different meaning dimensions associated with a word.

The paper is laid out as follows: Section 2 renders a problem description; Section 3 has a description of the approach; Section 4 lays out the evaluation of the method; Section 5 has a general discussion; followed by sections 6 and 7, which discuss future work and concluding remarks, respectively.

2. Problem Description

2.1. Problem Definition

The most accurate method of obtaining sense annotations for texts in a corpus is manually. Yet, it has proven to be a very expensive and labor-intensive endeavor because people performing the task need to be trained in the specific ontology utilized and have to be constantly checked for consistency, resulting in an enormous overhead. Over the course of the years, researchers in computational linguistics have proposed a variety of data driven (corpus based) methods of resolving lexical ambiguity in an automated fashion. Corpus-based methods rely on observations of the surface representations of words in running text. Corpus-based methods attempt to extract patterns of linguistic behavior that emerge from large amounts of data. Many of the corpus-based methods proposed utilize hand-crafted as well as automated knowledge resources, such as MRDs and computational lexicons, in addition to corpora, in order to arrive at better lexical ambiguity resolution [Ide & Veronis, 1998].

This current research is a statistical corpus-based method, which assumes the availability of a knowledge resource in addition to a parallel corpus. There are two main approaches within the data-driven framework that address the problem of sense annotation of data: **unsupervised** methods [Agirre et al. 2000; Litkowsky 2000; Lin 2000; Resnik 1997; Yarowsky, 1992&1995; etc.] and **supervised** methods [Bruce & Weibe 1994; Lin 1999; Yarowsky, 1993 etc.]. Unsupervised methods make no assumptions about the data, i.e. they do not require sense-annotated data as a prerequisite for the algorithm. Supervised methods, on the other hand, assume the availability of annotated data for training. On average, supervised methods yield better performance results than unsupervised methods [Kilgarriff & Rosensweig, 2000]. Supervised methods usually have a training phase where they tune a system to data that is already sense-annotated. Supervised methods need large amounts of such data in order to produce reliable results. Unfortunately, large amounts of sense-annotated data do not exist for nearly all languages. Moreover,

supervised methods are inherently highly tuned to the training data. Hence, the same supervised algorithm that was trained on an economic genre corpus such as the *Wall Street Journal*, will not be able to perform as well if applied to a literature genre corpus, such as George Orwell's *Nineteen Eighty Four*. Accordingly, Supervised methods are not portable to different kinds of corpora. On the other hand, unsupervised methods lack the sensitive tuning to the kind of corpus under investigation. Unfortunately, the unsupervised methods' lack of sensitivity to a specific corpus genre negatively affects the sense annotation quality. Yet, they have the advantage of not depending on the availability of sense-annotated data. Accordingly, they can be utilized as a means of obtaining and supplying large amounts of word sense annotated data – albeit noisy – that would be used for bootstrapping supervised systems.

To date, all automated methods aim at solving the problem for one language, usually for the language with the most available knowledge resources. This constitutes a problem for languages with scarce resources (low density languages), which are starting to take their place on the global research agenda due to the widespread use of the World Wide Web (WWW). Unfortunately, the majority of the language processing tools available has been tailored to address the former type of languages, ignoring low density languages in the process. Moreover, only a handful of the available approaches address the issue of automatically sense annotating text in a corpus (running text) on a large scale. The majority of the systems evaluate their performance against a handful of the available data, hence creating a scalability problem. Furthermore, very few systems are evaluated against the same material. It was only recently that the community decided to create a standard for word sense disambiguation, which resulted in the first SENSEVAL [Kilgarriff & Palmer, 2000].

This research investigates the feasibility of automatically sense annotating (tagging) large amounts of data in corpora using an unsupervised algorithm, targeting two languages simultaneously, only one of which has an available sense inventory. This will result in large amounts of annotated data for a language that already has knowledge resources, which would be utilized by supervised algorithms. Furthermore, it will have bootstrapped

sense tagging for a low density language. Moreover, links for the low density language in the established high density language sense inventory will be created, consequently, bootstrapping the creation of sense inventories for low density languages. Finally, it will provide a rich repository of cross lingual sense annotated data that can be researched further.

2.2. Significance

Large amounts of sense-annotated data in different languages provide a valuable resource for researching issues of lexicalization patterns cross linguistically. As mentioned before, it will help explore how polysemous words express their different senses across languages. If consistent patterns are detected, it may help identify and make explicit core characteristics for such words, which in turn, may aid lexicographers in building knowledge resources.

From a linguistics perspective, knowing the sense of the words involved in a sentence allows for making better predictions regarding the degrees of its acceptability, therefore, rendering a better model of human's understanding of language. For instance, in the following sentences:

1.a. I ran a mile in 10 minutes.
1.b. ?I ran a store in 10 minutes.

Sentence (1.b.) is a less accepted usage of the word *run*. Sentence (1.b.) assumes the *managing* sense of run but, in general, people do not manage places in 10 minutes. Informants get an acceptable reading of the sentence if the tense of *run* is modified to the future tense as *'I will run a store in 10 minutes'*, allowing for the reading of the verb *run* as *start managing* as in *'I will start managing the store in 10 minutes'*. Furthermore, knowing the semantic characteristics of the object words provides insight into the aspectual information borne by the verb. For example, in the following examples:

2.a. I baked a cake.
2.b. I baked a potato.

In (2.a.) and (2.b.), one distinguishes different senses for the verb *bake*. In (2.a.) *bake* has a *create* sense since a cake bears the semantic characteristic *artifact*, therefore allowing for an *achievement* reading of the verb. On the other hand, *bake* in (2.b.) has a *cooking* sense, therefore a *process* reading, because it is baking a potato, hence, a *natural object* [Bergler, 1995].

Almost all natural language processing (NLP) applications would benefit tremendously from the availability of sense-annotated data. Words with multiple senses constitute a serious bottleneck for many NLP applications such as Machine Translation (MT) systems, Cross Language Information Retrieval (CLIR) systems, Natural Language Understanding (NLU) applications, as well as Parsing systems.

In MT, given a sentence *'I walked by the bank'*, the system will fail to know which *bank* is being referred to, the *money bank* or the *river bank*. To date, most MT systems work on the scope of a sentence. Hence, if an MT system has the sense information for the word *bank* available, it will be able to translate it into the appropriate corresponding lexical item in the foreign language. Similarly, for CLIR, if the user enters a query *'log information'*, the system could translate the word *log* as *rondin* and find all the documents in French containing the word *rondin*. Yet, the user could have intended the *journal* or the *logarithm* sense of *log*. Therefore, as in the MT case, the quality of the output is seriously affected by the choice of the translation word sense. In NLU, polysemy plays a significant role in automatic message understanding, especially, in situations where it is critical to understand the intent of the user and perform a task. Parsing systems could benefit from sense information. Sentences such as (1.b.) can be ruled out given the sense information for the verb *run* (*manage* sense) due to the semantic characteristics of its argument *store*, which is interpreted in its *locational* sense.

2.3. Relevant Background

Using lexical translations as a source of sense distinction is an idea that has been in existence since the early 1990's [Brown et al. 1991; Dagan et al. 1991; Gale et al. 1992]. The key observation is that when a polysemous word in one language (**L1**) is translated into another language (**L2**), the polysemous word in L1 is translated into several distinct L2 words in different contexts corresponding to the L1 word's various senses.

Resnik & Yarowsky [Resnik & Yarowsky, 1999] investigated various distance measures and translingual sense inventories by analyzing native speakers annotations of 222 polysemous contexts across twelve different languages. They showed that monolingual sense distinctions could be discriminated in some set of second languages. Moreover, their findings suggested a correlation between language family distance and the extent to which polysemous words would express their various senses as distinct words, therefore the farther the family distance of L1 from L2, the better the sense distinction.

In an independent study, Dyvik [Dyvik, 1998] conducted a study on a set of Norwegian polysemous words and their translations into English. He proposed an unsupervised method that did not rely on any external resources for sense distinction. He discovered word senses in a corpus by using translations and their reverse translations, i.e. finding the translations of a Norweigian polysemous word in the English text then looking for the translation of the English words that corresponded to the original Norweigian word in the Norweigian text and so on, back and forth. He concluded that translation could indeed be used reliably for sense distinction. Exploiting translations enabled him to discover appropriate senses for the majority of the Norweigian polysemous words investigated.

Finally, a recent study by Ide [Ide, 2000] aimed at exploring the extent to which polysemous words and their equivalents lexicalize differently across five different languages belonging to four different language families: Germanic, Slavic, Finno-Ugrec, and Romance. The languages were English, Slovene, Estonian, Romanian, and Czech. The data was extracted manually from an on-line parallel corpus, comprising translations

of George Orwell's *Nineteen Eighty Four*. She concluded, contrary to Resnik & Yarowsky, that the evidence was weak for a correlation between language family distance and lexicalization pattern cross linguistically, yet she noted that translation could successfully be used as a filter for sense distinction.

2.4. Underlying Hypothesis

Given the promising indications from previous research, this study explores the relation between the translations of multiple instances of a polysemous word in a corpus. The basic assumption is that if several words $(w_1, w_2, \ldots, w_x)$ in L1 are translated into the same orthographic form in L2, then $w_1, w_2, \ldots, w_x$ share some meaning characteristic that brings the corresponding sense for each of these words to the foreground. *This makes the crucial assumption that the foreign word is not ambiguous.* Figure 1 below illustrates the assumption.

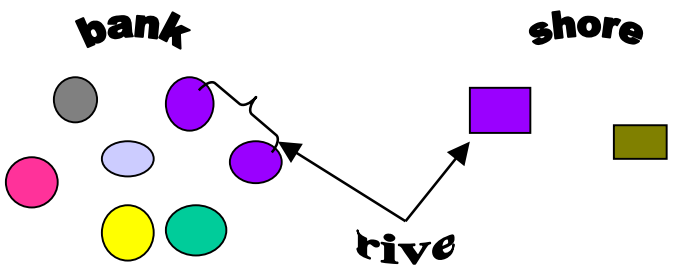


Figure 1: common attributes between words in text

In figure 1, the purple color is an indication of shared meaning characteristic(s) between the words *bank* and *shore*. Choice of different geometrical shapes indicates that the senses are not necessarily identical. As illustrated in the diagram, *rive* is the chosen French translation for both words. Then, *rive* has highlighted the shared meaning component between the two words in English. In this case, the shared characteristic is a concept such as *edge of the water* or *geological formation*.

2.5. High level method description

2.5.1. General Outline

Texts in translation (**parallel corpora**) are required in order to investigate the feasibility of this general hypothesis. If the task of sense annotating the parallel corpora is manually attempted, the study will require the knowledge of speakers of both languages. For instance, if the two languages are English and French, the task would be twofold: *1. to locate polysemous words in English and their corresponding translations in the French text. 2. Tag the English words with the French words that they were translated into.* This task will point out the English words, which were translated into the same orthographic forms in French. Moreover, based on the bilingual speakers knowledge of both languages, they can explicitly indicate the meaning dimensions (characteristics) that the different English words that mapped to the same orthographic form in French have in common. Unfortunately, this is a labor-intensive exercise. Moreover, since the English words are being tagged by their corresponding translations in French (2), the resulting tags will be corpus specific. Therefore, the need arises for a method of automatically discovering word mappings (alignments) in corpora and a standard independent sense inventory that is computational in nature with a well-defined distance measure between the words' senses. Once the words in one language are annotated with their appropriate sense tags, the tags are mapped over to the other language (the translation language). Accordingly, the sense chosen for a word in one language is the same sense projected in the foreign language if lexicalized[6]. This assumes that the translator, while translating the text, chose a similar sense of the word being translated in the foreign language. Due to divergences in which languages represent meaning, the mapping of sense tags is not necessarily straightforward. There is on-going research in the area of mapping senses cross linguistically on the taxonomic level by developers of the inventory EuroWordNet [Alonge et al., 1998; Rodriguez et al., 1998] attempting to create an interlingual index

[6] *languages do not necessarily express concepts in the same lexical surface representation, for instance a word in English can be translated into a phrase in German and vice versa or even a morpheme in some other language, languages tend to mix different levels of granularity, perhaps depending on socio-linguistic factors. Yet, in translation the meaning is preserved.*

(ILI) indicating cases of sense variations –divergences and/or conflations - across the different languages participating in the inventory [Vossen et al., 1999]. Yet, for purposes of the current research, the polysemous word instance in a language and its translation will bear the same sense tag. For example, if an instance of the word *bank* in an English-French parallel corpus translates to an instance of the word *rive* in French, and *bank* is assigned sense tag #2 from a particular sense inventory, then the instance of *rive* on the French side will be assigned the same sense tag.

2.5.2. Required Resources

Two knowledge resources are required for this study: *1. a parallel corpus; 2. a sense inventory for one of the languages.* The need for a parallel corpus stems from the need for large amounts of data in translation, in order to make solid conclusions about the observed data, especially that data representation in a sample of natural language is usually sparse. A number of years ago, the availability of large corpora in translation would have constituted a major impediment to the realization of this investigation. Recently, techniques have been proposed to acquire parallel corpora automatically from the WWW [Nie, 1999; Resnik, 1999], providing the linguistics community with large amounts of data in translation, with high accuracy and relatively minimal manual labor. The second resource is a sense inventory for words in one of the languages, where each word is represented with its/as its corresponding senses. As a first approximation, such a resource is required for only one of the languages under investigation. The current assumption is that since the data are translations of one another, the translator is trusted to have chosen the most faithful lexical translation that conveys the sense of the original word by preserving the salient meaning characteristic. It should be acknowledged that this approximation does not take into consideration possible sense variations cross linguistically, which is an issue farther discussed in section 6 of this paper. The chosen sense inventory should be rich enough to provide maximum coverage for the corpus utilized.

2.6. Research Hypotheses

- Translations of multiple instances of the same word in text expose salient/relevant characteristics of the targeted corresponding translated words.
- In a token aligned parallel corpus, it is possible to accurately automatically sense tag large amounts of data for more than one language simultaneously, given a sense inventory for only one of the languages and an appropriate distance measure between word senses.

2.7. Performance Criterion

In this study, the method has to achieve accuracy rates in the sense tagging task that exceed a chosen baseline. ***Accuracy is a measure of how many times was the method able to correctly annotate a polysemous word with its appropriate sense tag.***

3. Proposed method

The method proposed in this study is a data driven unsupervised algorithm for automatic large-scale word sense tagging for the two languages of a parallel corpus, simultaneously. The approach is unsupervised in as much as it assumes the absence of annotated data as a given, yet it is relies on the availability of a computational word sense inventory for one of the languages. Figure 2 below illustrates a high level view of the method.

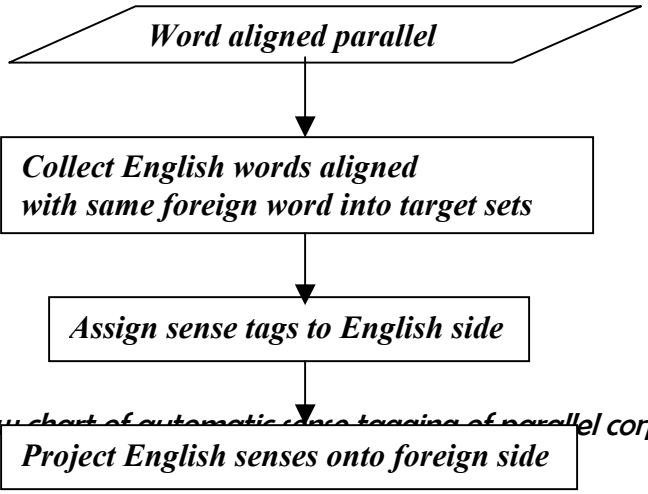


Figure 2: flow chart of automatic sense tagging of parallel corpora

3.1. Word Aligned Parallel corpus

The approach assumes the availability of token aligned parallel corpora. A token refers to a space delimited unit in a tokenized text in language. It refers both to word instances and punctuation. Figure 3 shows the kind of alignment expected by the algorithm.

La maison grise est grande .

The gray house is big .

Figure 3: A sample token alignment in a parallel corpus

Every token in one corpus is aligned to a token on the translation side. Figure 3 illustrates a very simple aligned sentence. In some cases, tokens in one language align with more than one token in the translation corpus. Tokens can align to NULL on the translation side if there is no correspondent. There exist automated methods to token-align parallel corpora that are sentence aligned, an example of which is the GIZA system, which is part of the Egypt package [Al Onaizan et al. 1999] for statistical machine translation based on the IBM models 1-4 [Brown et al. 1991]. The current algorithm assumes that the parallel corpora are obtained token aligned.

3.2. Create Target Sets

3.2.1. Identifying the aligned tokens

Once the token aligned corpora are obtained, all instances of words (tokens) that are aligned with the same orthographic form in the translation corpus are collected paired. Throughout this research paper, the language of the corpus that has the sense inventory available to it is referred to as the **target** language while the translation is referred to as the **source** language[7]. In all the illustration figures below, English is the target language and French is the source language.

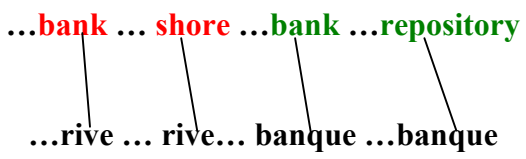


Figure 4: tokens aligned in the parallel corpus

In figure 4, the different instances of *rive* align with an instance of *bank* and an instance of *shore* on the English side of the corpus. Similarily, somewhere else in the corpus, *bank* and *repository* align with different instances of *banque* in French. The dots indicate running text.

Figure 5 shows different tokens *bank* and the different tokens of *shore* that aligned with tokens of the French word *rive* with their location information in the corpora.

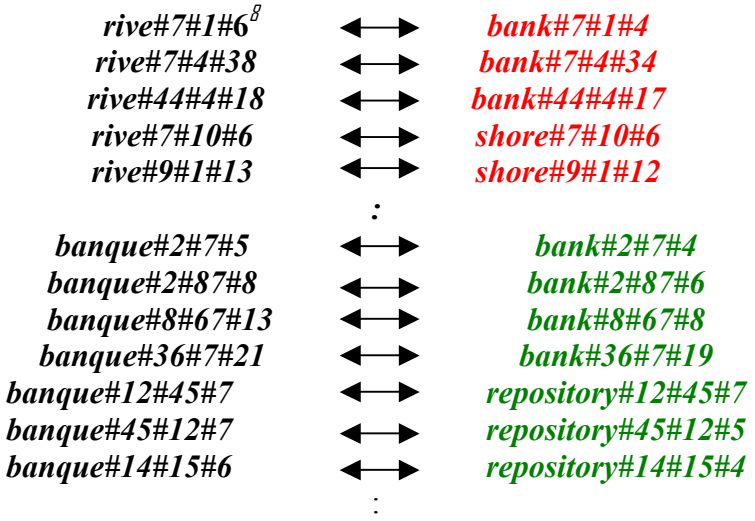


Figure 5: aligned tokens from the source to the target side

3.2.2. Conflating the alignments

All the tokens on the source side that share the same orthographic form are conflated into a word type. In the example illustrated in figure 5, all the tokens *rive* are conflated in the

---

[7] *Adopting the terminology of noisy channel in information theory*

[8] *rive#7#1#6 refers to an instance of the word rive in document #7, sentence #1 at position #6 in the sentence*

word type **RIVE** by removing the location information, and likewise for the tokens *banque* which are, in turn, conflated as **BANQUE**. The corresponding target language tokens are grouped in a token set as illustrated in figure 6.

RIVE: ***{bank#7#1#4, bank#7#4#34, bank#44#4#17, shore#7#10#6, shore#9#1#12, …}***
BANQUE: ***{bank#2#7#4, bank#2#87#6, bank#8#67#8, bank#36#7#19, repository#12#45#7, repository#45#12#5, repository#14#15#34, …}***

Figure 6: French source word and the corresponding English instances

The next step is to reduce the set of target tokens to target word types. The resulting set of word types is referred to as a ***target*** set. Figure 7 illustrates the target set for **RIVE** and **BANQUE**.

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

Figure 7: the final target sets for the source words RIVE & BANQUE

3.3. Sense Assignment to target language words

Once the target sets are obtained, they may be assigned the appropriate sense tags from a word sense inventory. At this point, the first research hypothesis is addressed: translations of multiple instances of the same word (in the current example, **RIVE**) in text expose salient/relevant dimensions of meaning – characteristics - for the two target word types (**BANK** and **SHORE**). It is crucial that different target words align with the same source word for this hypothesis to be tested, i.e. if a source word has a target set with only one word type, then the research hypothesis is not applicable. A distance measure needs to be defined to measure the similarity between the word senses of the target set. Therefore, a similarity function ***sim*(w_x, w_y)**, where ***sim*** calculates the distance between all the senses of word w_x and word w_y, for all the words in the target set is defined. The goal is to

maximize the similarity among the word senses across the target word types, hence, the value of ***sim***. Accordingly, the resulting distance metric is ***max(sim(w_x, w_y))***, which is an optimization function because it aims at choosing the senses that are most similar among all the senses of all the words in the target set. For instance, given the target set for **RIVE** as {**BANK, SHORE**}, all the senses corresponding to the two words in a sense inventory are compared and the ones that are the most similar according to the defined similarity metric are chosen as the appropriate tags for the respective word types. Looking up the word **BANK** in the *Collins Cobuild* dictionary [Sinclair, 1993], it comprises five nominal senses, while **SHORE** has two senses listed. Consequently, the chosen ***sim*** function will compute 2X5 comparisons, each comparison resulting in a similarity value. The sense tags (numbers) that yield the highest similarity value are assigned to the word types respectively. For illustration, the five listed senses for **BANK** are:

1. *A bank is an institution where people or businesses can keep their money…*
2. *the bank in a gambling game is the money that belongs to the dealer or to the casino management*
3. *a bank is the raised ground along the edge of a river or a lake*
4. *a bank of something such as computer data or blood is a store of it that is kept ready for use when needed*
5. *a bank of switches, keys, etc on a machine*

The two senses listed for **SHORE** are:

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

The different senses for the word **REPOSITORY** are:

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

Given the above mentioned definitions for the different sense entries for the target words **BANK** and **SHORE** and an appropriate similarity function, one would expect that sense

#3 for **BANK** and sense #1 for **SHORE** would yield the highest similarity value. Consequently, the target word types **BANK** and **SHORE** are assigned those senses. In this case, the similarity function would be a computation of the amount of overlap between the words in the definitions of the different senses of the words compared. This similarity computation was proposed and tested earlier on by Lesk [Lesk, 1986]. It is worth noting that, the salient characteristic is the concept of *edge of the river* as it is repeated in the definitions of the respective chosen senses. Likewise, for the words **BANK** and **REPOSITORY** in correspondence with the source word **BANQUE**: **BANK** is assigned sense #4 and **REPOSITORY** is assigned sense #2. In this case, the salient characteristic is *a place to keep objects of a kind.* The resulting target tag set is illustrated in figure 8, as well as, the senses propagated to the tokens corresponding to the word types.

RIVE: {BANK$_3$, SHORE$_1$, …}
BANQUE: {BANK$_4$, REPOSITORY$_2$, …}

Type tag set

RIVE: {*bank#7#1#4$_3$*, *bank#7#4#34$_3$*, *bank#44#4#17$_3$*, *shore#7#10#6$_1$*, *shore#9#1#12$_1$*, *…}*
BANQUE: {*bank#2#7#4$_4$*, *bank#2#87#6$_4$*, *bank#8#67#8$_4$*, *bank#36#7#19$_4$*, *repository#12#45#7$_2$*, *repository#45#12#5$_2$*, *repository#14#15#34$_2$*, *…}*

Token tag set

Figure 8: The target sets with their sense tags assigned

3.4. Project target sense tags to source tokens

Finally, the target sense tags are propagated to the source side of the corpus. This is a direct mapping step.

rive#7#1#6[BANK$_3$] ⟷ *bank#7#1#4$_3$*
rive#7#4#38[BANK$_3$] ⟷ *bank#7#4#34$_3$*
rive#44#4#18[BANK$_3$] ⟷ *bank#44#4#17$_3$*

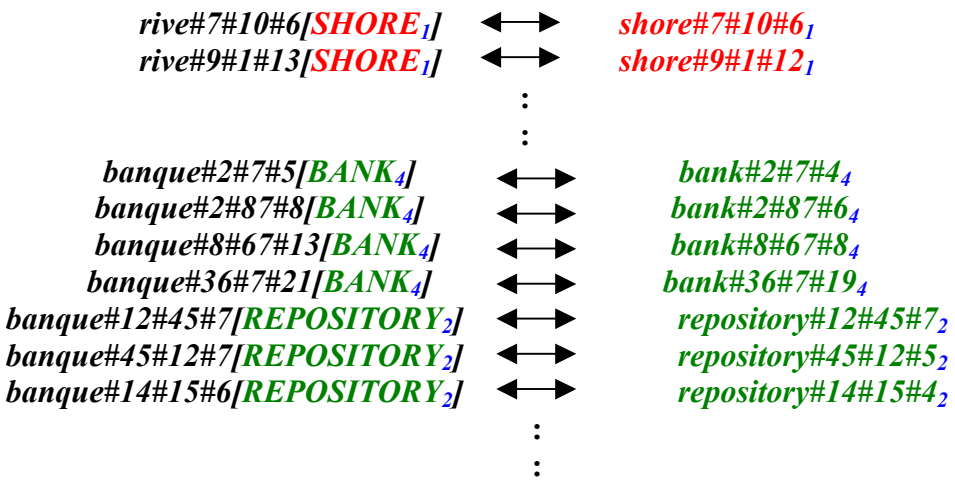


Figure 9: tagging the source language

In figure 9, *rive* and *banque* are assigned the senses corresponding to the target language sense inventory entries, thereby creating links for the French words in the *Collins* dictionary. Furthermore, even though the tokens are annotated with the sense tags from the dictionary, they are in effect also annotated with the salient characteristic explicitly. Therefore, in the case of instances of *rive* and its corresponding translation *bank* (expressed here as *rive-bank*) and *rive-shore* pairs, each of the instances is annotated with the characteristic *edge of the water*. Similarly, the pairs *banque-bank* and *banque-repository* are annotated with the salient characteristic: *a place to keep things of a kind*. Accordingly, the translations bring the salient characteristic of these polysemous words to the foreground. (first research hypothesis)

4. Evaluation

4.1. Materials

4.1.1. Corpora

Typically, a manual annotation has the best quality of sense annotations. A manual annotation constitutes a gold standard to be evaluated against. In order to evaluate the proposed approach, the need arose for a corpus that has been manually annotated with sense information. Unfortunately, the only corpus that has a sizeable percentage of it manually annotated is the Brown corpus of American English (BAE) [Francis & Ku□era, 1982]. BAE has the advantage of having had nearly one fifth of it manually sense annotated covering different parts of speech. BAE is a balanced corpus of approximately one million words. The fact that the BAE is balanced indicates that it offers the variability in contexts as it covers a variety of genres and topics. Yet, the algorithm requires a parallel corpus and the BAE, as the name indicates, is an English corpus only, which does not exist in translation. Consequently, a decision was made to approximate a human translation of the entire BAE using commercially available machine translation (MT) systems. MT systems have the benefit of performing the job relatively faster than a human translator and they are economically more feasible, however, the quality of the MT produced translations is much lower than that produced by a human. Given an MT system, it is relatively easy to translate the corpus into more than one language, therefore, allowing room for testing more hypotheses regarding meaning representation in different languages.

Two different commercially available MT systems are used to translate BAE: Globalink Pro 6.4 (GL) and Systran 2.0 (SYS). The decision to use two MT systems assumes that the systems translate contexts differently in addition to the fact that they employ different dictionaries in the process, thereby, allowing for variability in the translation words. BAE is translated into 3 different languages: French, German, and Spanish. The underlying

assumption here is that different languages deal with ambiguity in different ways, accordingly, a polysemous word in English may preserve the same kind of polysemy in French but have different words for the senses of the word in German. EuroWordNet exists for all three languages, therefore availing this research of a knowledge resource in which to investigate the quality of the sense annotation on the source language side of the parallel corpus. Furthermore, both MT systems claim to produce good quality translations in three chosen languages. Throughout this is evaluation, the English corpus is the target corpus and the foreign side is the source corpus. Thereby, the translation resulted in 6 parallel corpora, (two MT systems X three languages), with BAE as the target language.

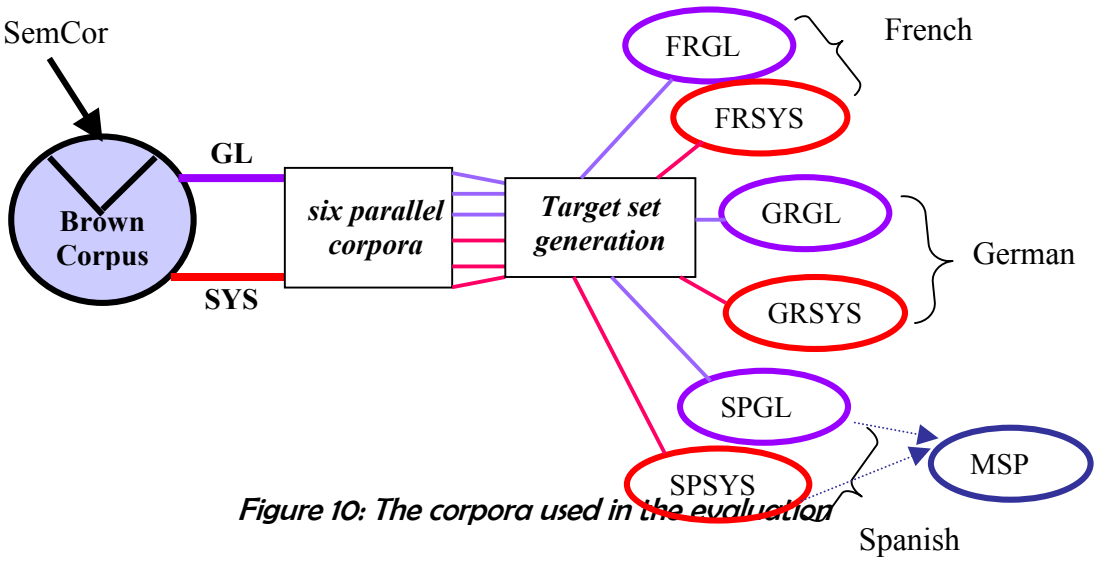


Figure 10: The corpora used in the evaluation

The corpora were as follows: French, German, and Spanish produced by GL, and French, German, and Spanish produced by SYS. Since the parallel corpora were artificially created, an automated token alignment system was used, GIZA. The GIZA system is part of the EGYPT statistical machine translation package [Al Onaizan et al., 1999]. EGYPT is an implementation of IBM models 1-3. [Brown et al., 1993]. Each of these models produces a Viterbi alignment. The models are trained in succession where the final parameter values from one model are used as the starting parameters for the next model. Given a source and target pair of aligned sentences, GIZA produces the most probable token-level alignments. Multiple token alignments are allowed on the target language side, i.e. a token in English could align with multiple tokens in a foreign language.

Tokens on either side could align with nothing, designated as a NULL token. GIZA requires a large corpus in order to produce reliable alignments, hence, the use of the entire BAE, not just the manually sense tagged portion. In figure 10, FRGL, GRGL, SPGL, FRSYS, GRSYS, SPSYS, refer to the target sets after the corpora are token aligned by GIZA and the target sets are generated as described in section 3.2. For instance, for the parallel corpus pair French Globalink translation and BAE, the French word types and their corresponding type target sets in English are referred to as FRGL; FRSYS is in reference to the same idea but using Systran as the MT system. Likewise for the rest of the translations, GRGL refers to the source German words translated by Globalink and aligned with English target sets, and the rest follows for the German translation using Systran and the Spanish translations. Finally, MSP refers to merging the target sets for both of the SPGL and SPSYS. The driving hypothesis is that merging the results of two translations and their alignments will increase the variability in the target sets, hence come closer to a human translation. For example, the words **SHORE**, **BANK** are in the target set of **ORILLA** in SPGL, and **COAST**, **BANK**, & **SHORE** are in the target set for **ORILLA** in SPSYS, the union of the target sets is taken and the result is a merged target set for **ORILLA** as follows: {**BANK, COAST, SHORE**}.

4.1.2. Sense Inventory

Since a portion – two hundred thousand words, approximately one fifth - of BAE was manually sense tagged using WordNet 1.6. [Fellbaum, 1998], it was decided to use WordNet 1.6. as the sense inventory. WordNet is a computational semantic lexicon for English. It was constructed by hand. It is a large-scale enumerative knowledge base. WordNet is freely available. It combines the knowledge found in traditional dictionaries, as senses of words and it defines synsets of synonymous words, which represent single lexical concepts. A word that is represented with several synsets is ambiguous. Furthermore, it organizes these concepts in a taxonomic manner, into a hierarchy[9]. The hierarchy is organized such that the more specific concepts are lower than the more abstract concepts. WordNet defines several types of semantic relations:

hyponymy/hypernymy, antonymy, meronymy, etc. For instance, a synset is a hypernym of another synset if it is a broader concept, accordingly, FOOD is a hypernym of the concept FRUIT. WordNet comprises 4 part of speech databases: a noun, verb, adjective, and adverb database. The noun database is the richest of the four databases as it comprises approximately 66,000 concepts and has the depth of 15 nodes, thereby, nearly four times the size of the verbs database and three times the size of the adjectives database. In the current study, the focus is only on the nouns in BAE, consequently, only the noun database in WordNet 1.6. is of direct interest at this stage. The majority of the concepts in the noun database are linked via an identity relation referred to as the IS-A relation. A fragment of WordNet is illustrated in figure 11.

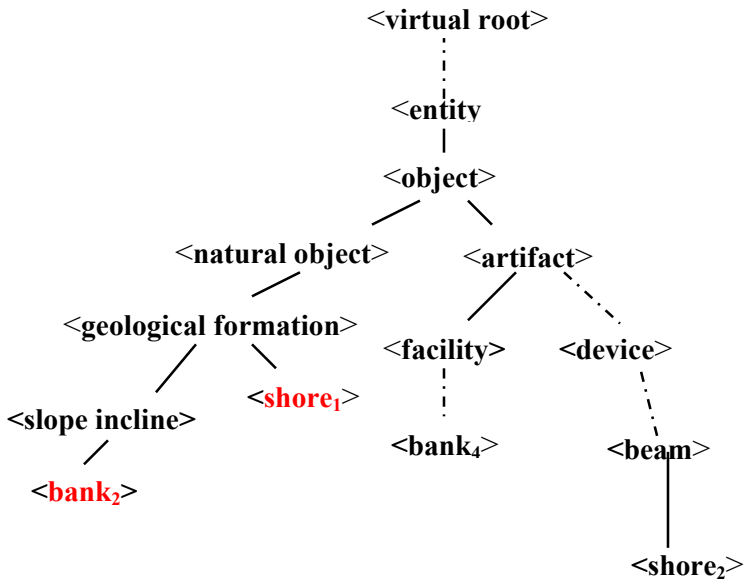


Figure 11: an excerpt from Wordnet 1.6

Figure 11 illustrates part of wordnet, in the sample, we have two senses of the word *bank* and two of the senses for *shore* shown. The dotted lines indicate omitted nodes.

4.1.3. Test set

[9] *WordNet allows for multiple inheritance, therefore a synset may have more than one parent*

As mentioned earlier, roughly one fifth of BAE is manually sense-tagged by the WordNet group, which resulted in a semantic concordance (Semcor) [Miller et al. 1994]. In the majority of the cases (nearly 85%), the annotators decided on a unique WordNet 1.6 sense for polysemous word instances in context. They annotated the four parts of speech, which correspond to WordNet's four part of speech sense inventories. It is worth noting that the inter-annotator agreement was at a low 78.6% overall, and as low as 70% for words with eight or more senses [Fellbaum et al., 1998].

For the current study, the human annotated data constitutes a gold standard against which to evaluate the proposed approach. Only polysemous nouns are considered for testing[10]. Hence, part of speech tags that are available in the Penn Tree Bank are used to identify the nouns in BAE. The **test set** includes only the polysemous nouns in WordNet 1.6 that occur in Semcor. The test set has 58372 noun instances of 6824 noun types. In cases where the manual annotator chose more than one sense, only the first manually chosen sense is considered.

As for the baseline, there are two commonly used baselines: 1. a **random** baseline (RBL), where a sense tag is assigned to the word in the corpus at random, from the list of available senses for a noun in WordNet; 2. a **most frequent sense** baseline (FBL), where a word is assigned the most frequent sense listed for it in WordNet 1.6, where senses are listed in order of their frequency in language in WordNet 1.6. It is important to note that FBL is more appropriate as baseline for supervised methods since the frequencies of the senses are obtained from sense annotated corpora [Resnik, personal communication].

4.2. Distance measure

[10] *There are no inherent restrictions in the method for applying it to other parts of speech.*

In order to assign the appropriate senses to the English words in the target sets, i.e. assign a **tag set** – a tag set can have one or more sense tags - to each word, an algorithm proposed and implemented by Resnik [Resnik, 1999] is used. The algorithm is an optimization function called Disambiguate_class. Given a target set of words in English, Disambiguate_class calculates the pairwise similarity across all the senses of the words in the target set and assigns the highest confidence scores to those senses that maximize an overall similarity value across the whole set of words' senses in the target set. Disambiguate_class is based on an information theoretic similarity measure, where the distance between the senses is measured in terms of ***information content*** the senses share. The algorithm assumes the presence of a taxonomy that has nodes with associated probabilities from a corpus. Information content is measured as in the following equation:

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)] \qquad 1$$

Where $S(c_1,c_2)$ is the set of concepts that subsume both c_1 and c_2 . A concept that obtains maximum similarity value is the most informative subsumer. Due to the structure of WordNet, no concept is less informative that its superordinates in the hierarchy. Accordingly, in WordNet, the higher up a node is in the hierarchy the less informative it is. The probabilities are calculated as a function of a concept's frequency in a corpus and a node's frequency includes an aggregate of the frequencies of its children in the hierarchy. In this study, similarity is computed among words in the corpus rather than concepts, therefore, the equation is defined as follows:

$$wsim(w_1, w_2) = \max_{c_1, c_2} [sim(c_1, c_2)] \qquad 2$$

Where c_1 ranges over the senses of w_1 and c_2 ranges over the senses of w_2. and $sim(c_1,c_2)$ is calculated as in equation 1.

Given the above similarity measure, Resnik, defines the algorithm Disambiguate_class. Disambiguate_class calculates the distances between all the senses of the words in a

given set of words using the above taxonomically-defined distance measure. The underlying assumption is that at least one sense from each one of the words in the set of words is relevant. The senses that contribute the most to the overall maximization of the similarity value for the set are assigned the highest confidence score, which is a score between 0-1. In order to illustrate the algorithm, consider the words *shore* and *bank* in figure 11. The distance between all the senses of *shore* and all the senses of *bank* is computed, rendering 2X10 comparisons, as *bank* has 10 senses in WordNet 1.6 and *shore* has two senses. The distance between any two senses of the two words is measured by the amount of information content in the lowest subsuming node for both words' senses. Calculated over all the senses of both words, the most informative subsuming node has the maximum similarity value between the two words in question. In figure 11, $bank_2$ and $shore_1$ are subsumed by several parents: *geological formation, natural object, object, entity*. Yet, *geological formation* is the narrowest (lowest) most informative subsumer. $bank_2$ and $shore_1$ are closer to one another than $shore_1$'s distance from $bank_4$, since $bank_2$ and $shore_1$ share a more informative subsumer than $bank_4$ and $shore_1$, i.e. $bank_2$ and $shore_1$ share the subsumer *geological formation* which is more specific than the subsumer *object* shared by *bank4* and *shore1*. Likewise, $bank_4$ and $shore_2$ have a more generic –less informative - subsumer *artifact* if compared against *geological formation* for $bank_2$ and $shore_1$. *geological formation* is considered the semantic characteristic for the words bank and shore in their respective contexts. Therefore, in the excerpt in figure 11, $bank_2$ and $shore_1$ are the most similar of the senses for the two words and they are assigned the highest confidence scores. Figure 12 illustrates the confidence assignment by Disambiguate_class on a given target set {BANK, SHORE, COAST}, where all three words are polysemous.

Word 'bank' (10 senses)
1. 0.0000 depository_financial, bank, banking_concern, banking_company: a financial institution that accepts deposits and channels the money into lending activities
2. 1.0000 bank: sloping land (especially the slope beside a body of water);
3. 0.0000 bank: a supply or stock held in reserve especially for future use
4. 0.0000 bank, bank_building: a building in which commercial banking is transacted

5. 0.0000 bank: an arrangement of similar objects in a row or in tiers
6. 0.0000 savings_bank, coin_bank, money_box, bank: a container (usually with a slot in the top) for keeping money at home
7. 1.0000 bank: a long ridge or pile
8. 0.0000 bank: the funds held by a gambling house or the dealer in some gambling games
9. 1.0000 bank, cant, camber: a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force
10. 0.0000 bank: a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)

Word 'shore' (2 senses)
1. 1.0000 shore: the land along the edge of a body of water (a lake or ocean or river)
2. 0.0000 shore: a beam that is propped against a structure to provide support

Word 'coast' (2 senses)
1. 1.0000 seashore, coast, seacoast: the shore of a sea or ocean
2. 0.0000 coast: the area within view; "the coast is clear"

Figure 12: Confidence score assignment by disambiguate_class

In figure 12 illustrates actual output from the algorithm disambiguate_class applied to the set {BANK, COAST, SHORE}. disambiguate_class assigns senses 2, 7 and 9 a 100% confidence score for the word **BANK**, therefore, all three senses are equal contributors to the overall maximization function. It is worth comparing the sense assignment from WordNet with that illustrated in figure 8, where a unique sense is assigned to the word **BANK**. If the glosses for senses 2,7, and 9 are compared, one will notice that they are very similar in meaning. This illustrates the fine granularity of WordNet in representing word senses. Both **SHORE** and **COAST**'s first sense are assigned the highest confidence score. Consequently, the final tag set for this target set is {BANK$_{2,9,7}$, COAST$_{1}$, SHORE$_{1}$}.

4.3. Experimentation

4.3.1. Experiment environment

The algorithm is developed using a combination of C and Perl code on a Sun Solaris 2.6. platform.

4.3.2. Preprocessing

BAE is detokenised for translation purposes. Once the 6 translations are obtained (cf. sec 4.1.1), they are tokenized since the GIZA system expects the corpora to be sentence aligned and tokenized. The tokenization involves replacing reduced forms such as *j'ai* in French by its equivalent *je ai*, also separating out punctuation marks from the text. Appendix A has a full listing of the scripts used for tokenization. Since the study is dealing with Latin based languages, tokenization is not a serious impediment, which would have not been the case if this study involves an oriental language such as Chinese where segmentation constitutes a serious bottleneck.

The manually assigned part of speech tags (POS) provided in the Penn TreeBank [Marcus et al., 1993] are used to identify the nouns in the BAE corpus. A problem was encountered because the lexicalization that is in the Semcor data is different from that of the Penn TreeBank, therefore, in cases where the Penn TreeBank assigned a POS tag to a compound and Semcor assigned sense tags to the components of the compound, the Brill POS tagger [Brill, 1994] is used.

Sentence alignment is straightforward since the MT systems translated the corpora sentence by sentence respecting the sentence boundaries. Therefore, the POS tagged BAE and its translation into one of the three languages are tokenized and sentence aligned and finally passed on to the GIZA system for token alignment.

4.3.3. Experiment conditions

Since the alignments and the translations are completely automated, the quality of the target sets is noisy. For example, the French word **CATASTROPHE** is aligned with the English word types **{CATASTROPHE, DISASTER, SHOCKER, TRAGEDY}**. The word **SHOCKER** is an outlier in this set, since all the other words are closer to one another in meaning. The fact that **SHOCKER** is in the target set, affects the overall confidence score assignment of disambiguate_class for this set of words. For example, given the four words in the set, disambiguate_class assigns the associated senses the following confidence scores:

Target Set: **CATASTROPHE DISASTER SHOCKER TRAGEDY**

Word 'catastrophe' (3 senses)
0.5000 *[1] calamity, catastrophe, disaster, tragedy, cataclysm: an event resulting in great loss and misfortune*
0.5000 *[2] catastrophe, disaster: a state of extreme (usually irremediable) ruin and misfortune*
0.0 *[3] catastrophe, cataclysm: a sudden violent change in the earth's surface*

Word 'disaster' (3 senses)
0.5000 *[1] catastrophe, disaster: a state of extreme (usually irremediable) ruin and misfortune*
0.5000 *[2] calamity, catastrophe, disaster, tragedy, cataclysm: an event resulting in great loss and misfortune*
0.0 *[3] disaster: an act that has disastrous consequences*

Word 'shocker' (2 senses)
0.0000 *[1] shocker: a shockingly bad person*
1.0000 *[2] shocker: a sensational message (in a film or play or novel)*

Word 'tragedy' (2 senses)
0.5000 *[1] calamity, catastrophe, disaster, tragedy, cataclysm: an event resulting in great loss and misfortune*
0.5000 *[2] tragedy: drama in which the protagonist is overcome by some superior force or circumstance; excites terror or pity*

Figure 13: Actual output from Disambiguate_class

Condition 1 of the experiment is the default condition referred to as **Classim.** In Classim, the words in the target set are assigned the sense tags or synset numbers that score the highest confidence value by disambiguate_class, highlighted in red in figure 13. Classim, is a maximization function over the full target set. Therefore, the resulting tag set for the given target set is as follows: {***CATASTROPHE*$_{[1]}$, *CATASTROPHE*$_{[2]}$, *DISASTER*$_{[1]}$, *DISASTER*$_{[2]}$, *SHOCKER*$_{[2]}$, *TRAGEDY*$_{[1]}$, *TRAGEDY*$_{[2]}$}** in the default condition. The default condition, Classim attempts to find a global optimum tag or set of tags for each of the words in a target set. Accordingly, it generates many tags or too few tags that do not always satisfy the different contexts in which the polysemous words occur.

Consequently, two alternative conditions are devised, where the optimization function is localized to pairwise comparisons of the senses of each pair of words in the set. The two additional conditions are: condition 2, referred to as **Pairsim_1**, and condition 3 referred to as **Pairsim_all**. For both conditions 2 and 3, the data was submitted as pairs of words to disambiguate_class. For Pairsim_1, only senses that scored a 1.000 confidence score are considered. A sense of each of the two words has to have been assigned a 1.0 confidence score, by disambiguate_class, in the pairwise comparison, in order for the sense to be used in the final sense tag set for a specific word. On the other hand, for condition 3, Pairsim_all, all the senses that achieve maximum score in a pairwise comparison are assigned to the final tag set for the target words. Figure 14 illustrates the confidence scores assigned to the words in the pairwise comparison of the target set illustrated in figure 13 above.

A. Target Set: TRAGEDY SHOCKER

Tragedy	*shocker*
[1] 1.0000	*[1] 0.5000*
[2] 0.0000	*[2] 0.5000*

B. Target Set: TRAGEDY CATASTROPHE

Tragedy	*Catastrophe*
[1] 1.0000	*[1] 1.0000*
[2] 0.0000	*[2] 0.0000*
	[3] 0.0000

C. Target Set: TRAGEDY DISASTER

Tragedy	*disaster*
[1] 1.0000	*[1] 0.0000*
[2] 0.0000	*[2] 1.0000*
	[3] 0.0000

D. Target Set: DISASTER SHOCKER

Disaster	*shocker*
[1] 1.0000	*[1] 0.5000*
[2] 0.0000	*[2] 0.5000*
	[3] 0.0000

E. Target Set: CATASTROPHE SHOCKER

Catastrophe	*shocker*
[1] 0.0000	*[1] 0.5000*
[2] 1.0000	*[2] 0.5000*
	[3] 0.0000

F. Target Set: CATASTROPHE DISASTER

Catastrophe	*disaster*
[1] 1.0000	*[1] 0.0000*
[2] 0.0000	*[2] 1.0000*
[3] 0.0000	*[3] 0.0000*

Figure 14: Pairwise comparison of the target set

In figure 14, sense numbers are between square brackets and the real numbers are the confidence scores assigned by disambiguate_class. The senses that make it to the final tag set for this target set in Pairsim_1 are highlighted in red. The senses highlighted in blue are the senses assigned according to condition 3, Pairsim_all, which assigns both the senses highlighted in red as well as those highlighted in blue to the final tag set.

Illustrating condition 2, in figure 14 section A, none of the senses is used to tag neither TRAGEDY nor SHOCKER, even though sense #2 for tragedy is assigned a 1.0 confidence score. Similarly, in sections D and E, no senses are assigned in this pairwise comparison. Accordingly, only where disambiguate_class was very confident in one of the senses per word, in the pairwise comparison, is a sense chosen for tagging. In this case, the final tag set for the target set is {$CATASTROPHE_{[1]}$, $DISASTER_{[2]}$, $TRAGEDY_{[1]}$}. By considering the associated glosses for the chosen senses, in figure 13,

the final tag set is a minimal. Furthermore, Pairsim_1 has weeded out the senses of **SHOCKER** from the list of final tag set.

On the other hand, condition 3, is more generous as it yields the following tag set for the current target set of words - the union of the senses highlighted in red and blue in figure 14: {***CATASTROPHE*****$_{[1]}$**, ***CATASTROPHE*****$_{[2]}$**, ***DISASTER*****$_{[1]}$**, ***DISASTER*****$_{[2]}$**, ***SHOCKER*****$_{[1]}$**, ***SHOCKER*****$_{[2]}$**, ***TRAGEDY*****$_{[1]}$**}. In this tag set, Pairsim_all assigns the same senses to the words in the target set as in Pairsim_1 in addition to assigning senses to the outlier word. But, it is different from Classim since it does not yield the second sense for **TRAGEDY**. The second sense for **TRAGEDY**, as defined in figure 13, it is not a correct sense to be assigned because it has little in common with **CATASTROPHE** and **DISASTER**. Pairsim_all has the disadvantage of not being able to rule out a complete outlier, such as **SHOCKER** in this example. Yet, it has the advantage of choosing more of the appropriate senses for **CATASTROPHE** and **DISASTER** than Pairsim_1, which eliminates a possible sense for each.

4.3.4. Evaluation metric

Only the polysemous noun instances in the target sets, that were also in the manually sense tagged Semcor data, are evaluated. Multiple sense assignment is allowed especially since WordNet is extremely fine grained. However, the majority of unsupervised methods, for the English language sense annotation task, assign only one sense per polysemous word. Therefore, in order to create a basis for comparison against other reported methods in the literature, only one sense tag is allowed per word. For evaluation, the multiple senses per word are assumed to have equal weight, i.e. the sense tags are uniformly distributed in these cases. Accordingly, the ties between multiple sense tags are broken by choosing the most frequent sense among the assigned senses in the tag set. The evaluation is a rigorous evaluation metric, partial credit for assigning a close sense tag in the tag set is not allowed. Only exact matches with the manually chosen tags in the test set are considered correct. Accuracy was measured as follows:

$$accuracy\% = \frac{\#\,correct}{total} \times 100 \qquad 3$$

Coverage is measured as the percentage of the data in the test set – Semcor polysemous nouns - that is covered by the tag set.

4.4. Results

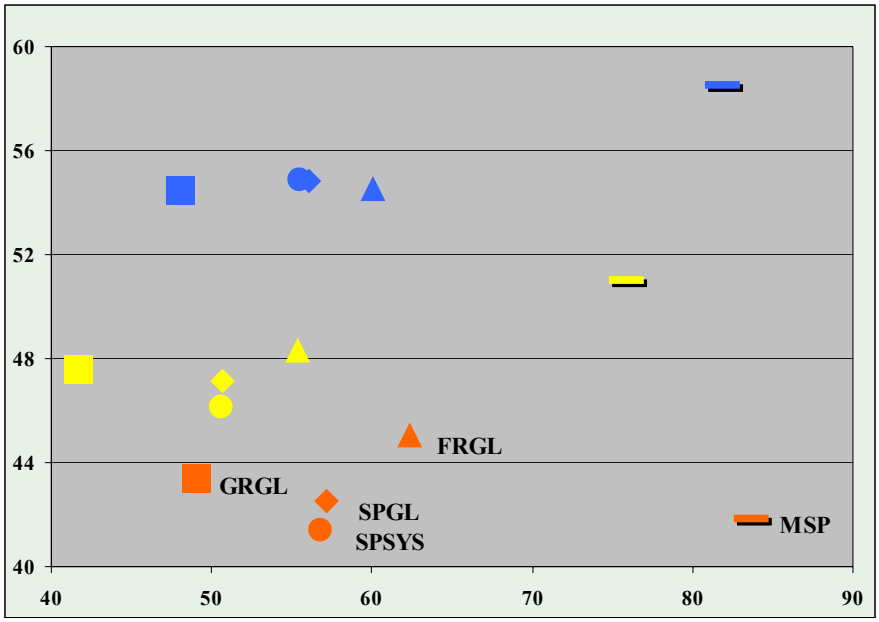
In table 1, the results for the three experimental conditions are presented with the two metrics accuracy and coverage.

Corpus	*Classim*		*Pairsim_1*		*Pairsim_all*	
	Coverage	*Accuracy*	*Coverage*	*Accuracy*	*Coverage*	*Accuracy*
FRGL	62.0	45.0	55.4	48.2	60.1	54.5
GRGL	49.0	43.5	41.6	47.6	48.0	54.5
SPGL	57.2	42.5	50.7	47.1	56.1	54.8
SPSYS	56.8	41.4	50.6	46.1	55.5	54.9
MSP	83.6	41.6	75.8	51.0	81.8	58.5
RBL	28.6%					
FBL	67.6%					

Table 1: results for experimental conditions 1, 2 & 3

Table 1 illustrates the results for experimental conditions 1, 2 and 3. **Coverage** indicates the percentage of polysemous nouns from the Semcor data for which the proposed method selected a sense. The method did not have a 100 % coverage because not all of the words that are tagged are in Semcor, since the whole BAE is used. Moreover, in many of the cases none of the words in the foreign data had a polysemous word from Semcor in the target sets. Or in some cases the target set will only have single words. Moreover, the method hinges upon the presence of words that are similar in the data, like BANK and SHORE, if they do not exist then the method will not be able to perform the task. **Accuracy** is measured based on the evaluation metric in equation 3. All the results are statistically significant at $p < 0.05$ according to the *Z score* test for the difference between population proportions. RBL and FBL are the baselines (see section 4.1.3).

Graph 1 illustrates the results where the x-axis is **coverage** and the y-axis is **accuracy**, both measures are in percentages. The baselines are not shown on the graph since FBL is above the highest accuracy rate achieved and the random baseline, RBL, is significantly below the lowest accuracy rate achieved.



Graph 1: results of conditions 1,2, & 3

Graph 1 shows a pictorial view of the results obtained in table 1, for the three conditions of the experiment. The red color indicates condition 1, the default condition Classim; yellow indicates Pairsim_1 results, condition 2; blue indicates the results from condition 3, Pairsim_all. The shapes of the markers are consistent across the three conditions for the alignments across the three languages.

4.5. Discussion of quantitative results

None of the reported accuracy rates obtained exceeded the FBL of 67.6% accuracy rate. As mentioned above, FBL is an appropriate baseline for comparing supervised methods

not unsupervised methods, therefore the appropriate baseline in this case is RBL, which is significantly exceeded in all three conditions of the experiment. Therefore, the second research hypothesis is accepted.

The results are considered reasonably good especially when compared with related work [cf. section 4.6]. There is a cascading of error affect in this experiment, due to the usage of both commercial MT system to translate BAE and the use of an automated alignment method. The quality of the translations is especially affected since the BAE covers many different genres. This in turn has a negative effect on the alignment quality, which even with a genuine parallel corpus in a limited domain is reported to achieve accuracy rates of only 89%, English to German alignments [Och & Ney, 2000a].

Results from condition 3, Pairsim_all outperform results from condition 2, Pairsim_1, which in turn outperform results from condition 1, the default condition, Classim, in terms of **accuracy**. Coverage deteriorates for both conditions 2 and 3. In fact, condition 2 depicts the worst coverage, which is expected due to the rigorous nature of sense assignment. Results from condition 3 significantly improve accuracy with a slight deterioration in coverage from the default condition, which reports the best coverage across all the conditions.

The merged Spanish alignments, MSP, yield the highest accuracy for both Pairsim_1 and Pairsim_all, an improvement of ~4% over the individual Spanish translation alignments SPGL and SPSYS, which was expected since the variability in the target sets increased due to the incorporation of data from two translation sources. Yet, MSP yielded worse accuracy rates than SPGL (~1%), and slightly better results than SPSYS (0.2%) in the default condition. It is suspected that the worse results are due to the same factor, which caused an increase in the performance for conditions 2 and 3: variability in the target sets. The global nature of the similarity measure in the default condition lead to a dampening in the signal by increasing the noise in the sets therefore, which eventually lead to the assignment of sub-optimal tags. In particular in the cases where the target sets had many word types, there appeared to be a spectrum of word meaning similarity. For example, in

the following target set {**AGITATION, BUSTLE, COMMOTION, TURMOIL, RESTLESSNESS, FLURRY, FUSS**}, it is easy to detect relatedness rather than similarity. The words **BUSTLE, FLURRY,** and **FUSS** seem to be most similar to one another in this set, in fact the sense entries for these words is exactly the same synset number in WordNet; **COMMOTION** and **TURMOIL** seem to be most similar to one another; and yet **RESTLESSNESS** and **TURMOIL** can also form a cluster. **AGITATION** can fit into any cluster since it is equidistant from all the subsets. Yet, in a setting where an algorithm is trying to achieve a global optimum sense tagging, **AGITATION** will be swayed more toward being assigned the sense in tune with the majority cluster. The global optimum for all the polysemous words in the set will suppress some senses that are appropriate for some of the contexts. The notion of an optimal target set size is an interesting idea to pursue. Upon qualitative inspection, the data suggests that the optimal target size is 3-4 word types. The approach is hypothesized to yield results in cases where there is similarity, rather than relatedness[11].

The three chosen languages are close to English, French and Spanish share the Latin roots with English, while German shares the Anglo Saxon roots. Yet, French and German are closer to English than Spanish is. One would expect them to yield similar results. Interestingly, this is not the case. The results of the French alignments, FRGL, significantly outperform the results from the German alignments, GRGL, for all three conditions in sense tagging accuracy. As a speculation, this observation may be related to the quality of the translation. It could be that Globalink is better at translating into French than it is into German. Moreover, it could be attributed to the highly agglutinative nature of the German language[12].

There was no statistically significant difference ($p<0.05$ confidence) for both measures – accuracy and coverage - between the Globalink and the Systran translations for all three conditions in the Spanish, SPSYS and SPGL, data.

---

[11] *A formal study of the optimal size target set needs to be performed before conclusions can be drawn.*

The results from SPGL and SPSYS were expected to outperform the results from FRGL and GRGL – according to the conclusion drawn in [Resnik & Yarowsky, 1999] correlating language distance with sense distinction – yet, this is not the case. The pattern for condition 1 and 2 is the same: FRGL accuracy rates are the highest, followed by GRGL rates then SPGL & SPSYS accuracy rates. In condition 3, the three language alignments yielded similar accuracy rates.

In terms of coverage, GRGL, yielded the worst results across all three conditions. The Spanish data, SPGL as well as SPSYS, followed with better coverage results across the different conditions. Spanish is a highly inflectional language, which could have contributed to the poverty in coverage as well. For the individual languages, French, FRGL, yielded the best coverage results across all conditions. As expected, the overall best coverage was obtained from the merged target sets from the Spanish, MSP, with a margin of ~25% across all conditions when compared with the individual Spanish target sets, SPGL & SPSYS. MSP clearly surpassed the other two data sets, GRGL and FRGL.

4.6. Related work

To date, all automated methods proposed in the literature to sense annotate large amounts of data have targeted one language only at a time. This is in clear contrast to the current approach, where the proposal is to sense annotate two languages simultaneously. Of the known unsupervised methods proposed in the literature, only three studies relate to our approach since they evaluated against the Semcor data as well.

Resnik [Resnik, 1997] proposed an unsupervised technique that annotates data with their sense information based on selectional preference association strength between a predicate and its argument. The basic intuition is that predicates that select strongly for their arguments, in fact, select for a specific sense of a word rather than a word. The

---

[17] *No morphological analysis was performed on the data in any of the languages that participated in the study*

algorithm calculates the selectional association as the difference between the prior probability of a concept (in this case a noun), as its probability of occurring as an argument ***prob(c)*** in a specific type of relation ***r*** and its posterior probability ***prob(c|pred)*** using relative entropy. Therefore, the selectional association of a specific predicate for a particular concept ***c*** is equivalent to the proportion of selectional strength it exercises on its argument. For instance, the probability of *human* appearing in a subject relation in a corpus is very high, yet given a predicate *buzz*, the probability of *human* occurring in the subject position is diminishes tremendously, therefore, *buzz* has a strong selectional association since the difference between the prior probability of *human* is very different from its posterior conditional probability. Since the data are not annotated with concept information, Resnik considered each occurrence of a word as evidence distributed uniformally among all the concepts to which it belongs to in WordNet. Accordingly, if the word *water* belongs to 12 classes in wordnet, each of the concepts of which *water* is a hyponym is assigned a prior probability of 1/12. The sense annotating algorithm assumes that the training data will bias the selection of a specific hypernym concept for disambiguation according to the evidence seen. Therefore, if *eat* is observed as the predicate of words like *fruit*, *vegetable*, and *cheeses* in the training data, and *eat* occurs with *meat* in the test data, then the system will assign the FOOD sense to the polysemous word *meat.* The algorithm will favor the FOOD sense rather than the COGNITION[13] sense, since there is more evidence for *eat* selecting for FOOD type words than COGNITION type words in the training data. The problem with this assumption lies in the fact that the predicate has to be a strong selector itself[14]. For instance, a verb such as take – which has weak selectional preferences – would be equally likely to choose all the concepts associated by all of its arguments if they occurred with the same frequency, with the assumption that there is enough balanced data representing different modes of expression for these predicates. Moreover, another drawback to this approach is the sensitivity to the training data. For instance, one can easily envision a scenario where the corpus has many metaphorical uses of the verb *eat* (biasing the system toward non edible concepts) while only one instance of the verb is *eat* used in its literal sense (with

[13] As in the meat of the topic or question

arguments that are edible). Accordingly, when deciding the appropriate sense for meat as argument for eat in the test set, the system will decide on the COGNITION sense rather than the FOOD sense based on its bias toward other than edible concepts. Resnik investigated five different relations for sense disambiguation: verb-object, verb-subject, modifier-head, head-modifier, and adjective-noun. His approach crucially relies on the availability of a parsed corpus to give structural information for the training phase. He reported results on the Semcor data of WordNet 1.4, which is an older version than the one used in the current study. His approach yielded a maximum of 44.3 % accuracy rate for the verb-object relation. Similar to the current evaluation, he only considered ambiguous nouns.

Abney and Light [Abney & Light, 1999] present a similar approach to word sense disambiguation also using selectional preferences. The main difference between their model and Resnik's model is the adoption of a stochastic generative model for the estimation of the parameters associated with the concept classes to which a word belongs. They associated a Hidden Markov Model with each predicate relation pair. They used the Estimization Maximization algorithm to estimate the different parameters of the system. Their approach suffers the same drawbacks as the previous ones. They evaluated their system's performance against the Semcor data. They reported a maximum accuracy rate of 42.3%.

The third unsupervised method that was evaluated against the Semcor WordNet data is an approach based on learning Selectional preferences using Bayesian Networks. The method as proposed by Ciaramita & Johnson [Ciaramita & Johnson, 2000], performs sense disambiguation as a side effect. The approach is also based on the Resnik model mentioned above. Given a Bayesian Network, Bayesian inference can be used to estimate both marginal and posterior parameters, which in turn help derive the prior probabilities for the concepts. They represented WordNet as a Bayesian Network where the synsets in

---

[14] *Or it should be disambiguated into the various possible senses itself can conflate*

WordNet were represented as nodes in the Bayesian Network. They report results of 51.4% accuracy rate.

All three unsupervised methods depend on the availability of parsed data as a source of structural information, in contrast to the current approach, which does not require additional structural information to perform the sense annotation task. Moreover, the evaluation in all three cases was against previous Semcor data. Even though their results are not directly comparable to the current evaluation, there is reasonable evidence that the current approach performed significantly better (58.5% accuracy rate for MSP data) with minimal resources including less than perfect translations. It would be interesting to investigate the effect of combining structural information such as selectional preference especially when exploring the approach applied to predicative parts of speech, like verbs.

Recently, a workshop was dedicated to the evaluation of automatic systems performing sense annotation of corpora, SENSEVAL [Kilgariff & Palmer, 2000]. It is not possible to directly compare the current results with systems that participated in the SENSEVAL effort because the workshop evaluated the systems against a test set from a different knowledge resource, Hector. The Hector database/dictionary resulted from a pilot study in corpus analysis for lexicographic purposes. Lexicographers from Oxford dictionaries did the corpus analysis on a sizeable amount of data (17.3 million words) of British English from the 1980's and 1990's. The Hector dictionary is an enumerative lexicon that has 220,000 tokens over 1400 dictionary entries. It is relatively smaller than WordNet 1.6, which has 66,000 concepts in the noun taxonomy alone. The Hector database arranges the word senses in as a shallow hierarchy. WordNet represents words as concepts as opposed to the representation adopted in Hector, where the dictionary entries are words. In the Hector database, polysemous words are divided into their respective senses. The senses are ordered by both their frequency of occurrence in the analyzed corpus and by their semantic flow: the first sense is not the most common one only but also the psychologically prior sense[15]. The lexicographers on the project had a notion of a

[15] *http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/HECTORcorp.asc.*

core sense - a sense from which other senses have developed - which was always put first regardless of relative frequency[16]. Inherent in the adopted encoding is a sense of distance among the respective senses of an entry. The senses are organized in a hierarchical manner in some cases: ambiguous senses are on the same level while vague senses are represented hierarchically. As an example, the polysemous word accident has the following entry:

Accident

1. **Crash**
1.1. ***crash****: an unfortunate or disastrous incident not caused deliberately; a mishap causing injury or damage; in particular, a crash involving road vehicles.*
1.2. ***waiting****: an accident waiting to happen, and variants. A potentially disastrous situation, usually one caused by negligent or faulty procedures; also, a person to cause trouble*
1.3. ***happen****: in proverbial expressions referring to the inevitability of mishaps*
1.4. ***pee****: an incident of incontinence, especially by a child*

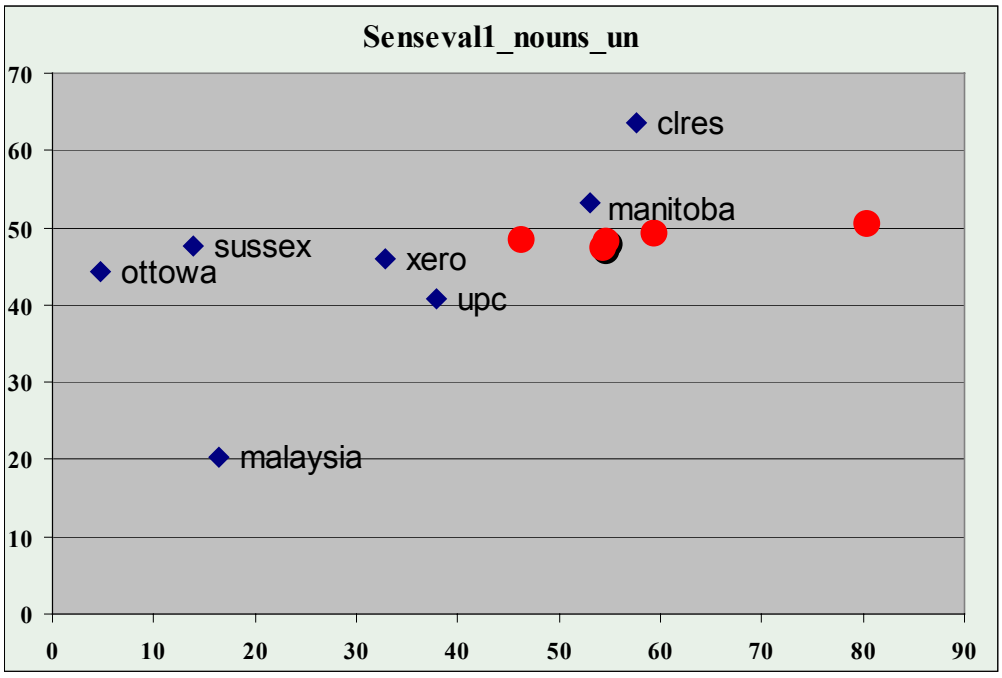
2. **Chance**
2.1. ***chance****: in general, something that happens without apparent or deliberate cause; a chance event or set of circumstances*
2.2. ***by accident****: by chance; unintentionally*

Figure 15: An example of a polysemous entry in the Hector database[17]

As illustrated in figure 15, the Hector database has a detailed level of granularity for its entries. For the 24 nouns used in SENSEVAL, the mean polysemy in the Hector database is 3, while the mean polysemy for the same items in WordNet 1.6. is 3.25. Comparing the means with a paired t-test ***(N=24, t(23)=0.663, at p<0.05 confidence level)***, it is clear that the two means are not significantly different from one another. Therefore, for those 24 nouns, the two databases are of comparable granularity.

[16] *It is not clear how they arrived at this notion of coreness, there were no reports of evidence supporting such notions in the document describing the acquisition procedure*

In SENSEVAL, the systems were evaluated against a handful of data from 4 categories: nouns, adverbs, adjectives and verbs, as well as, an 'unidentified' category. In the following graph, the average of the current results are plotted against results obtained from the unsupervised systems on the nouns subtask that participated in the workshop.



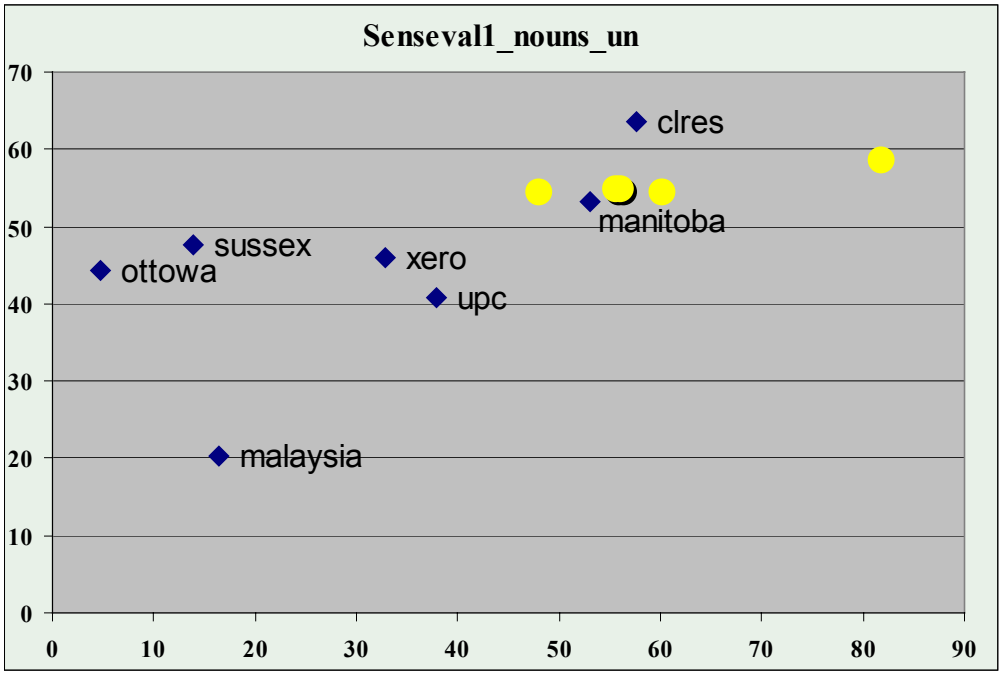
Graph 2 : SENSEVAL results for unsupervised systems on nouns subtask against average of current results

Graph 2 illustrates the results of the unsupervised systems on disambiguating nouns plotted against the current investigation's results shown in red. The x-axis represents coverage and the y-axis represents accuracy. The current results are the average of *accuracy* and *coverage* taken across the three conditions in the experiment. On average, the current approach yielded results that are high on the accuracy scale, roughly the third system as shown on the graph. The highest accuracy rate of 50.5% achieved by the merged alignments for the Spanish translations, MSP, comes second to the Manitoba system which yielded accuracy rates of 53.3% on the task. Considering only the best

---

[17] *Examples were omitted in interest of space. A full listing of the words that were chosen for the SENSEVAL exercise can be found on the SENSEVAL web page at http://www.itri.brighton.ac.uk/events/senseval/*

results obtained from the current method against the SENSEVAL workshop results, the highest accuracy rate obtained from experimental condition 3, Pairsim_all for MSP, is 58.5% accuracy rate, which would place the current system second only to CLRES. The current method's is placed halfway between the two best unsupervised systems in the workshop: exceeding Manitoba system achieved accuracy rate by 5% and below CLRES (63.5% accuracy rate) by 5%. Graph 3 below shows the results from Pairsim_all plotted against the SENSEVAL results.



Graph 3: Pairsim_all results vs. SENSEVAL unsupervised methods on noun subtask

5. General discussion

The proposed approach yields very promising quantitative results relative to other unsupervised methods even though the translations that are used are not genuine translations. The current method is the first of its kind to address sense tagging two corpora in two languages simultaneously and bootstrap the process for a low density language. The approach has the advantage of requiring minimal resources for only one

language in the parallel corpus language pair. The quality of the sense annotations on the foreign side (bootstrapped side) of the parallel corpus still requires evaluation. It will be interesting to investigate the correlation between the quality of the sense tagging on both sides of the parallel corpus. It seems that there is a close connection, the higher the accuracy of one, the higher the accuracy of the other. Moreover, it creates links for the words of the foreign language in an established ontology. In this study, in effect, links are created for the French, German and Spanish words in WordNet 1.6. Furthermore, large amounts of sense annotated data are created for English that can be used by supervised methods. The approach is fully automated. It is applied on a large scale. It is modular in design, for example, the knowledge source of word senses in the target language can change to an MRD and the distance measure will change accordingly to some function of calculating the amount of overlap between the words in the sense definitions of the words in question. Therefore, both the distance measure and the ontology can change without affecting the skeleton of the overall approach. Importantly, having such large amounts of sense annotated data facilitates discovering and studying different interesting aspects of lexical semantics cross linguistically.

By annotating the corpora on a large scale with the appropriate word senses, the proposed method is able to bring the salient characteristics shared by different senses of polysemous words to the foreground. As an example, given **BANK** and **SHORE** and their correspondence to different instances of **RIVE** in French, **BANK** and **SHORE** share the common characteristic *geological formation* which is ***explicitly*** their most informative subsumer, according to the utilized distance measure and WordNet 1.6. Therefore, it is conceivable to tag instances of the three words in both English and French with the salient characteristic, rendering a parallel corpus that is semantically annotated with characteristics rather than senses. Such a tagging is coarser in nature but very interesting from a linguistic perspective. Crucially, the sense information is required in order to identify the salient characteristics in the target language.

The method is limited by the cross linguistic lexicalization patterns. For example, if a word in L1 is not lexicalized as a word in L2, then the approach ignores it, hence, the

usage of more than one language source for the evaluation. Ide [Ide 2000] has reported a study where she calculates the percentage of the time an English word is translated as a lexical item in four different languages in the parallel text *Nineteen eighty four*. She found that only 86.6%, on average, of the words are translated as single lexical items in the foreign language. It would be interesting to calculate the same statistic for the corpus under investigation, since it sets an upperbound on the coverage performance of the approach with respect to the corpus examined.

The current method essentially depends on variability in the contexts allowing for the usage of words and other words that are close to them in meaning in the text. As noted throughout the paper, **BANK**, **SHORE** and **COAST** are very close in meaning and the closeness is highlighted by the fact that the three of them are translated as the same word **ORILLA** in Spanish. Yet, distant words may align with the same source word. For example, **AMORCE** in French may align with {**INITIATION, BAIT, CAP**}, which are all correct translations of the French word but they are distant from one another, therefore they will affect the accuracy results negatively by assigning inaccurate sense to the respective words. Moreover, in other cases, if one of the words in the target set is misaligned (an incorrect translation), it leads the results astray especially in the default condition. Experimental conditions 2 and 3, attempt to address this issue but the problem worsens if the outlier word is a monosemous word, since it gets a default confidence score of 1.0, therefore maintaining it in the set while potentially biasing the choice of senses for the other members of the target set.

The data is evaluated qualitatively. The intent is to devise methods for improving the sense annotation task on the target language side of the corpus. It is interesting to investigate the results if the approach had accurate translations and alignments. The intuition is that, given a perfect alignment and a perfect translation, the performance of the approach should significantly improve in accuracy. Accordingly, a set of 18 target sets were hand picked at random from the set of French words that started with the letter A in the FRGL data. The chosen target sets met both criteria of translation quality and alignment quality based on the author's bilingual knowledge of French and English. The

words in the target set had to be polysemous and included in the Semcor data. A sample from the 18 target sets is presented below.

French Target set
ABSURDITÉ: {ABSURDITY, NONSENSE}
ACCIDENT: {ACCIDENT, CRASH, WRECK}
ACCUSATION: {ACCUSATION, FRAMING, INDICTMENT}
ADVERSAIRES: {ANTAGONISTS, OPPONENTS, CONTESTANTS}
AGRICULTURE: {AGRICULTURE, FARMING}
APPARANCE: {APPEARANCE, LOOK}

figure 16: A sample of good alignments with good translations

It is worth noting that the cases are chosen blindly with absolutely no foreknowledge of their individual accuracy achievement. The majority of the 18 cases (14) have a cognate in the target set for the source French word. Furthermore, for all the cases that are deemed good alignments and good translations, the target sets never exceed 4 word types, which supports the notion that there seems to be an optimal size for target sets, i.e. if the set contains more than 4 elements then probably it is too wide a cluster. Since the data is already sense tagged, accuracy rate is calculated on those 18 target sets only. The accuracy rate yielded is 78% in the default condition (Classim), which is significantly above the FBL – the most frequent sense baseline – as well as the results yielded in FRGL for the same condition (45% accuracy). If the results are extrapolated to the whole data, it could potentially rival supervised methods on the nouns subtask in SENSEVAL, where the best supervised system, Durham, achieves an accuracy rate of 83% followed by the systems Hopkins and Tilburg and Ets-pu at 80% accuracy rates [Kilgariff& Rosenzweig, 2000].

In order to achieve such results, the algorithm requires very good translations and very good alignments. Attaining very good translations depends on the meticulousness of the person doing the translations. Given the current state of the art in machine translation exemplified by the commercially available machine translation systems (Globalink, Systran, Logos, etc), obviously human translations are far superior in quality. There are

many examples of very accurately done translations like the Bible [Resnik et al., 1999] which could serve as an interesting test bed for current approach, but might be too small – ~800,000 words - for the alignment tool to be able to produce reliable alignments[18]. On the other hand, acquiring large amounts of text in translation has been facilitated by techniques proposed in the literature [Nie et al., 1999; Resnik,1999]. The main foreseeable problem with using genuine parallel corpora is the evaluation. In the current experiment, the approach is evaluated against a manually annotated test set which are very laborious to obtain. The manual sense annotation usually requires trained lexicographers who spend a significant amount of time deciding on the appropriate tags from a predefined ontology. Due to the time required for such an effort, alternative means need to be devised to evaluate the sense tagged results.

On the other hand, improving the quality of the automated token alignments is underway. Och & Ney [Och & Ney, 2000b] have reported results of 94% on the alignment quality when measured on the Hansards parallel corpus using a more advanced version of the GIZA tool, which incorporates IBM model 4.

Throughout the investigation, the approach assumes simplistically that the source language words are either monosemous or vague at the most but not ambiguous. Observing the data closely indicates otherwise. Figure 17 illustrates some of these cases.

1. CANON: {CANNON, CANNONBALL, CANON, THEOLOGIAN}
2. BANDES: {BAND, GANG, MOB, STRIP, STREAK, TAPE}
3. BAIE: {BAY, BERRY, COVE}

Figure 17: examples of target sets comprising multiple clusters

In the given example, figure 17, in 1-3 the different clusters are highlighted in different colors. Upon inspecting 1, one can deduce that **CANON** in French, is polysemously ambiguous as it is used as the translation for both these English words: **CANNON** and

[18] *It is worthwhile to test the lower bound on the amount of data that can be aligned automatically by GIZA and produce reliable alignments.*

CANON. In 2, **BAND** is not highlighted since itself is a polysemous word that fits in both subclusters namely {**BAND, GANG, MOB**} and {**BAND, STRIP, STREAK, TAPE**}, likewise for {**BAY, BERRY**} and {**BERRY, COVE**}. Cases 1-3 show that the words ***CANON, BANDES, BAIE*** in French are polysemous words that are ambiguous. In case of ***BANDES*** and ***BAIE*** they are also vague words. The presence of such subclusters in the target set has a definite negative effect on the quality of the sense tagging. One way to resolve this problem is to gather distributional features of the target data [Diab&Finch, 2000; Pereira et al, 1993; Schütze, 1992; etc.] and apply automatic clustering techniques. Once clustering is applied, the appropriate target sets are discovered and, simultaneously, the clustering will discover in an automated unsupervised manner the number of senses for polysemous words in corpus of a language with scarce resources.

The evidence in the data suggests that if a source word is assigned multiple senses, then it is a vague word, which shares characteristics across its different senses. Therefore, the proposed method by allowing multiple sense assignment can be viewed as an automated method for discovering ambiguous vs. vague polysemy in a low density language.

In this investigation, around 15 percent of the test set data is manually tagged with more than one sense. The current evaluation only considers the first manually listed sense, which definitely has a negative impact on the results. The evaluation measure needs to deal with multiple sense assignment in a more sophisticated manner, rather than breaking the ties with the most frequent sense, for example assigning probabilities to the different senses that are associated with the same word. Furthermore, the evaluation metric needs revision by assigning partial credit to a sense tag if it is close enough to the correct tag. Therefore, incorporate Melamed & Resnik's suggestions for a more sensitive evaluation measure [Melamed & Resnik, 2000].

The overlap between the three language alignments is investigated. In terms of coverage, Pairsim_all has the best coverage if looking at two languages at a time. FRGL and GRGL overlap 34.4% of the time with the Semcor data; the FRGL and SPGL overlap 44% of the

time; the GRGL and SPGL data overlap 32.3%. Accordingly, it can be deduced that the different languages targeted different portions of the test set. As expected, the coverage figures are low because lexical ambiguity is preserved cross-linguistically for these three languages since they are close to English in different ways. Yet, it is interesting to see how well the approach performs given multilingual sources as filters for creating target sets. Therefore, the accuracy rates are computed for the data when the three language sources, FRGL, GRGL, and SPGL, unanimously agreed on a specific tag set for a word instance in BAE. The results are shown in table 2.

Condition	*Coverage %*	*Accuracy %*
Classim (default)	***6.6***	***49.9***
Pairsim_1	***1.7***	***67.1***
Pairsim_all	***4.5***	***90.4***

Table 2: Results of three languages voted on the same sense tag

The coverage results of the test data illustrated in table 2 are extremely low. Yet the accuracy results are promising, both experimental condition 2 and 3 yield very high results according to the defined evaluation metrics.

6. Future work

The goals for the immediate future include evaluating the sense tagging of the source corpus. The feasibility of manual evaluation needs to be assessed. In such an evaluation the granularity of the sense annotation would be an interesting factor to change, i.e. conduct different experiments with different levels of sense graininess for the human evaluation. Moreover, there is a need for assessing the value of referencing the InterLingual Index (ILI), which gives a mapping of WordNet to the various European WordNets, that accompanies EuroWordNet, for an automated method of evaluation of the quality of sense tagging on the source side. Currently, a copy of EuroWordNet is not available, but studies in the literature report it to be much smaller for the various languages than WordNet [Gonzalo et al., 2000]. Unfortunately, the ILI is currently

available only for WordNet 1.5, but there exists a mapping between WordNet 1.5 and WordNet 1.6 (the currently used version of WordNet)

A natural goal is to investigate the performance of the proposed approach on a genuine parallel corpus. As the qualitative data suggest (cf. section 5), the performance will improve significantly. Evaluation will be more challenging since large amounts of sense-tagged data for monolingual data do not really exist, except for the Semcor data, and the problem is escalated in dealing with parallel texts.

As mentioned in the general discussion section, fine-tuning the target sets will improve the results immensely. Clustering the words in the target sets is an essential step toward that goal.

Exploring other parts of speech is also a feasible goal especially verbs. It would be interesting to measure the impact of using context information on the performance of the system in particular when the work is extended to deal with verbs.

Participation in workshops organized for evaluating sense annotation is definitely ont he agenda. It is important to have a feel for where the current system stands with respect to the state of the art in data driven methods.

It would be interesting to study patterns of behavior in the annotated data and test whether information given by the semantic annotations of the nouns are sufficient to glean any interesting information about the different languages and the way they represent meanings. Questions such as “how much semantic annotation is needed before solid conclusions can be drawn? Are nouns sufficient for such a study? How important is it to integrate other parts of speech? Is this type of semantic annotation – the sense level, hence the salient characteristic level - really helpful? Furthermore, addressing questions about the level of granularity of the meaning-bearing unit in a language and if there is an underlying correlation with selection preference strength, is interesting.

Finally, it would be interesting to explore the possibility of extending this method to comparable corpora.

7. Conclusion

This paper presents an investigation into the feasibility of exploiting translations as a source of semantic annotation across languages. It addresses the issue of translations serving as a means of sense distinction. The hypothesis is that words that are translated into the same orthographic form share some dimensions of meaning exemplified by the words' respective senses. A data driven approach is proposed and evaluated on a large scale for several languages. The method annotates two languages from a parallel corpus with their senses in an ontology for one of the languages. The method yields very promising results. The results are significantly better than previous comparable methods evaluated against the same data.

Bibliography

Al-Onaizan, J. Y.Curin, M. Jahr, K. Knight, J. Laferty, D. Melamed, F. Och, D. Purdy, N. Smith, & D. Yarowsky (1999). "*Statistical Machine Translation, Final Report"*, JHU workshop. http://www.clsp.jhu.edu/ws99/projects/mt/final.report/mt-final-report.ps

Abney, S. & M. Light (1999). *"Hiding a semantic Hierarchy in a Markov model"*. **Proc. of the Workshop on Unsupervised Learning in Natural Language Processing, ACL**, Maryland, June.

Agirre, R., L. Padro & J. Atserias (2000). *"Combining Supervised and Unsupervised lexical knowledge methods for word sense disambiguation"*. Special issue on SENSEVAL**. Computers and the Humanities,** (34), pp. 103-108.

Alonge, A., N. Calzolari, P. Vossen, L. Loksma, I. Casrellon, M. A. Marti & W. Peters. (1998). *"The linguistic design of the EuroWordNet Database"*. Special issue on EuroWordNet, **Computers and the Humanities,** (32) 2-3.

Bergler, S. (1995). *"From Lexical Semantics to text Analysis"*. In P. Saint-Dizier & E. Viegas (ed.) **Computational Lexical Semantics**. Cambridge University Press, USA.

Brown, P. F., S. S. Della Pietra, V. J. Della Pietra, and R. L. Mercer (1993). "*The mathematics of statistical machine translation: Parameter estimation"*. **Computational Linguistics**, (19)-2: pp. 263-311.

Brill, Eric (1994). *"Some Advances In Rule-Based Part of Speech Tagging"*. **American Association for Artificial Intelligence** (AAAI).

Bruce, Rebecca & Janyce Wiebe (1994). "*Word-senseDisambiguation Using Decomposable Models"*. **Proc. of 32nd Association of Computational Linguistics**, Las Cruces, NM.

Ciaramita, M., & M. Johnson (2000). *"Explaining Away Ambiguity: Learning verb selectional preference with Bayesian networks"*. http://xxx.lang.gov/arXiv:cs:CL/0008020.

Cruse, D. A. (1995). *"Polysemy and related phenomenafrom a cognitive linguistic viewpoint"*. In P. Saint-Dizier & E. Viegas (ed.).**Computational Lexical Semantics**, Cambridge University Press, USA.

Dagan, Ido & Alon Itai (1994). "*Word Sense Disambiguation Using a Second Language Monolingual Corpus"*. **Computational Linguistics** (20), pp. 563-596

Diab, M. & S. Finch (2000). *"A Statistical word Level translation model for Comaprable corpora"*. **In proc. of Conference on Content based multimedia information Access (RIAO)**, Paris, April.

Dyvik, Helge (1998). *"Translations as Semantic mirrors"*. **Proc. of Workshop W13: Multilinguality in the lexicon II. The 13th biennial European Conference on Artificial Intelligence (ECAI 98)**, Brighton, UK, pp 24-44.

Fellbaum, Christiane. (ed.) (1998). **WordNet: An Electronic Lexical Database**. MIT Press, Cambridge, MA.

Fellbaum, C., J. Grabowski & S. Landes (1998). *"Performance and Confidence in a Semantic Annotation Task"*. In C. Fellbaum (ed.) **WordNet: An Electronic Lexical Database.** Chapter 9. MIT Press. Cambridge, MA.

Francis, W. & H. Ku□era (1982). **Frequency Analysis of English Usage**. Houghton Mifflin Co: New York.

Gonazalo, J., I. Chugur & F. Verdejo (2000). *"Sense Clusters for information Retrieval: Evidence from Semcor and EuroWordNet Interlingual Index"*. **Proc. of ACL-2000 Workshop: Word Senses and Multilinguality, in ACL 2000**, Hong Kong, October.

Hanks, Patrick (2000). *"Do Word Meanings Exist?"*. Special issue on SENSEVAL**. Computers and the humanities**, (34), pp. 205-215.

Ide, Nancy (2000). "*Cross-lingual sense determination: Can it work?"*. Special issue on SENSEVAL**. Computers and the Humanities**, (34), pp. 223-234.

Ide, N. & J. Veronis (1998). *"Word Sense Disambiguation: the state of the art"*. **Computational Linguistics**, (24)-1, pp. 1-40.

Kilgarriff, A. & M. Palmer (2000). *"Introduction to the Special Issue on SENSEVAL"*. Special Issue on SENSEVAL**. Computers and the Humanties,** (34) pp.1-13.

Kilgarriff, A. & J. Rosenzweig. (2000). *"Framework and Results for English SENSEVAL"*. **Special Issue on SENSEVAL. Computers and the Humanties,** (34), pp.15-48.

Kilgarriff, A. (1997). *"I don't believe in word senses"*. http://xxx.lang.gov/cmp-lg/9712006.

Lesk, Michael (1986). *"Automated Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone"*. **Proc. of the 1986 SIGDOC conference**, Toronto, Canada, June, pp. 24-26.

Lin, Dekang (1999). "*A Case-base Algorithm for Word Sense Disambiguation"*. **Pacific Association for Computational Linguistics**, Waterloo, Canada.

Lin, Dekang (2000). *"Word Sense Disambiguation with a similarity smoothed ase Library"*. Special Issue on SENSEVAL**. Computers and the Humanties,** (34), pp.147-152.

Litkowski, K. (2000). *"SENSEVAL: the CL-Research experience"*. Special Issue on SENSEVAL. **Computers and the Humanties,** (34), pp.153-158.

Melamed, Dan & P. Resnik (2000). *"Tagger Evaluation Given Hierarchical Tag Sets"*. Special Issue on SENSEVAL. **Computers and the Humanities**, (34), pp. 79-84.

Miller, G., M. Chodorow, S. Landes, C. Leacock, and R. Thomas (1994). *"Using a Semantic Concordance for Sense Identification"*. **ARPA Human Language Technology Workshop**, San Francisco, CA.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross & K. Miller (1990). "*WordNet: An on-line lexical database"*. **International Journal of Lexicography**, (3)-4, pp. 235-244.

Nie, J.Y., P. Isabelle, M. Simard, R. Durand (1999). *"Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web"*, **ACM-SIGIR conference**, Berkeley, CA, pp. 74-81.

Och, Franz J. & Hermann Ney (2000a). "*A Comparison of Alignment Models for Statistical Machine Translation"*. **8th Int. Conference on Computational Linguistics**, Saarbrücken, Germany, July.

Och, franz Josef & Hermann Ney (2000b). *"Improved Statistical Alignment Models"*. **Proc. of 38th Annual meeting of the Association for Computational Linguistics**. Hong Kong, October.

Periera, F., N. Tishby & L. Lee (1993). *"Distributional Clustering of English"*. **Proc. of 31st Annual Meeting of the Association for Computational Linguistics**, Ohio, June, pp. 183-190.

Pustejovsky, J. (1995). *"Linguistics constraints on type coercion"*. In P. Saint-Dizier & E. Viegas (ed.), **Computational Lexical Semantics**, Cambridge University Press, USA.

Resnik, Philip (1999). "*Mining the Web for Bilingual Text"*, **37th meeting of Association for Computational Linguistics**, College Park, Maryland, USA, June.

Resnik, Philip (1997). "*Selectional Preference and Sense Disambiguation"*, **SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?**, Washington, D.C., USA, April.

Resnik, Philip (1999). "*Semantic Similarity in a Taxonomy: An information-based Measure and its Application to Problems of Ambiguity in Natural Language"*. **Journal of Artificial Intelligence Research**, (11), pp. 95-130.

Resnik, P., M. Olsen & M. Diab (in press). *"Creating a Parallel Corpus from the book of 2000 Tongues"*. **Computers and the Humanities**.

Resnik, Philip & David Yarowsky (1998). "*Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation"*. **Natural Language Engineering**, (1), pp. 1-25.

Rodd, Jennifer, G. Gaskell & W. Marslen-Wilson (2000). *"The Advantages and Disadvantages of Semantic Ambiguity"*. **Proc. of 22nd conference of the Cognitive Science Society**, University of Pennsylvania, August.

Rodriguez, H., S. Climent, P. Vossen, L. Loksma, W. Peters, A. Alonge, F. Bertagna, A. Roventini. (1998). *"The Top-Down strategy for building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology"*. Special issue on EuroWordNet, **Computers and the Humanities**, (32)2-3.

Schütze, Hinrich (1992). *"Dimensions of Meaning"*. **Proc. of Supercomputing'92**. Los Alamitos, California: IEEE Computer Society Press, pp. 787-796.

Sinclair, John. (1993). **Collins Cobuild English Language Dictionary**. London: Harper Collins publications.

Vossen, P., W. Peters & J. Gonzalo (1999). *"Towards a Universal Index of Meaning"* **Proc. of Workshop SIGLEX, ACL**, Maryland, June.

Yarowsky, David (1993). *"One sense per collocation"*. **Proceedings of the ARPA Human Language technology Workshop**, New Jersey: Princton, 1993, pp. 266-271.

Yarowsky, David (1992). "*Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora"*. **Proc. of 14th International Conference on Computational Linguistics**, Nantes, France, July.

Yarowsky, David (1995). "*Unsupervised Word Sense Disambiguation Rivalling Supervised Methods"*. **33rd meeting of Association for Computational Linguistics**, Cambridge, MA.