DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

A SYSTEMATIC APPROACH TOWARDS THE SOLUTION OF THE POLYSEMY PROBLEM IN NATURAL LANGUAGE PROCESSING

Abed Alhakim Freihat

April 2011

Technical Report # DISI-11-463

Submitted as qualifying exam to the ICT PhD School of the Dept. of Information Engineering and Computer Science.

A systematic Approach towards the Solution of the Polysemy Problem in Natural Language Processing[1]

Abed Alhakim Freihat
Dipartimento Ingegneria e Scienza dell'Informazione
University of Trento
Povo, 38100 Trento, Italy

ABSTRACT

WordNet has been used widely in NLP and semantic applications. Despite the reputation of WordNet, it still suffers from many problems that make it hard to be usable by NLP and semantic applications. The major problem that has been extensively researched last decades is polysemy. Solving the polysemy problem is indispensable because the high polysemous nature of WordNet leads to insufficient quality of NLP and semantic applications results. In this proposal, we describe the polysemy problem, report the state of the art approaches, and introduce a novel approach for solving polysemy.

General Terms

Languages, Theory, Algorithms, Performance, Management

Keywords

Lexical databases, WordNet, Homophony, Polysemy, Systematic Polysemy, Polysemy Reduction, Lexical semantics, Semantic Search, Knowledge Engineering.

1. INTRODUCTION

From linguistics, a word is polysemous if it has more than one meaning [12]. Linguists differentiate between ***contrastive polysemy***, i.e. words with completely different and unrelated meanings - also called ***homonyms*** or ***homographs*** - and ***complementary polysemy***, i.e. words with different but related meanings. Synthetically, we can define:

- **Homographs**: words that have the same spelling and different unrelated meanings
- **Complementary words**: words that have the same spelling and related meanings

The polysemy in WordNet [18] is considered to be the main reason that makes it hard usable by NLP and semantic applications since a real distinction between homographs and complementary words is not given [6]. The complexity of the problem becomes more difficult in the cases, where the meanings of a word are contrastive and complementary as in (1), where (1) a refers to a completely different meaning from the related meanings in the senses (2)b and 2(c). Differentiating between the two types of polysemy should be possible through the semantic relations between the senses of polysemous words. Unfortunately, relations between complementary words are not systematically provided in WordNet. As a consequence, the fact that there are no relations between two polysemous synsets does not necessarily mean that they are homographs.

(1) a: *He sat on the bank of the river and watched the currents.*

(1) b: *He cashed a check at the bank.*

(1) c: *The bank is on the corner of Nassau and Witherspoon.*

In the last decades many approaches have been introduced to solve the polysemy problem through merging the similar meanings of polysemous words. These approaches in fact are helpful in cases, where words have closed meanings. However, polysemous words with closed meanings represent a small portion of the polysemy problem only. In fact, a significant portion of the polysemous senses should not be merged, as they are just similar in meaning [13] and not redundant. Consider for example the meanings of the word *bank* in (1) b and (1) c or the meanings of the word *post office* in (2). While the meaning of *post office* refers to a building or more general *a physical entity* in (2) a, the meaning refers to an institution as *an abstract entity* in (2) b. The generative lexicon theory [8] states that the meanings of complementary words are systematic and predictable. The difference between contrastive and complementary words in the generative lexicon theory is that the meanings of the complementary words do not contrast with each other, regular (systematic) and predictable. On the other hand, the meanings of contrastive words are not related, not regular, and not predictable [8] [15]. Following this theory, systematic polysemy approaches, such as CORELEX [15], examine the regularity between polysemous words and organize complementary words in systematic polysemy classes.

(2) a: *I met john near the post office.*

(2) b: *The post office delivers mail on Saturdays.*

In CORELEX, the first systematic polysemy approach, Paul Puitelaar has analyzed the complementary nouns in WordNet 1.5 and organized them in 126 systematic polysemy classes. Organizing systematic polysemous words as has been established in CORELEX is the first major step towards solving the polysemy

[1] This paper has been submitted as qualifying exam for the admission to the second year of the ICT PhD School of the Dept. of Information Engineering and Computer Science. I thank my advisor Prof. Fausto Giunchiglia for his teaching, guidance and feedback in maturing the ideas presented here.

problem in WordNet. To solve the problem, CORELEX should be refined and further steps are needed, especially cleaning process that detects and deals with redundancy and inconsistency in organizing the meanings of polysemous words in WordNet, are essential to get to more satisfactory precision and recall in word sense disambiguation approaches and consequently in applications requiring it, such as semantic search. In this proposal, we introduce an extension of CORELEX and explain the steps needed towards solving polysemy in WordNet.

This proposal is organized as follows: In section two, we describe the polysemy problem. In section three we discuss the state of the art approaches. In section four we describe CORELEX, introduce our extension of CORELEX and our methodology for solving polysemy in WordNet.

2. Polysemy in WordNet

WordNet is a lexical database that organizes synonyms of English words into sets called synsets where each synset is described through a gloss. For example the words *happiness* and *felicity* are considered to be synonyms and grouped into one synset {*happiness, felicity*} that is described through the gloss*: state of well-being characterized by emotions ranging from contentment to intense joy*.

WordNet organizes the relations between synsets through semantic relations where each word category has a number of relations that are used to organize the relations between the synsets of that grammatical category. For example the hyponymy relation (X is a type of Y) is used to organize the ontological structure of nouns. In WordNet, the synset {*happiness, felicity*} for example is a hyponym of the synset {*blessedness, beatitude, beatification*} with the gloss: *a state of supreme happiness.* Words in WordNet on the other hand are organized via lexical relations such as the antonymy relation (opposites of). For example the word *male* is antonym of the word *female*.

A word is polysemous (e.g. has more than one meaning) means that this word participates in more than one synset. In such cases, the synsets of the polysemous word are ordered in numbers (#1, #2 ,…). This order reflects the familiarity of the senses of a polysemous word. To compute the familiarity of the senses, WordNet performs statistics on sample tagged texts to calculate the frequency score of the senses. The sense number 1 is the most familiar or common sense.

WordNet 2.1 contains 147,257 words, 117,597 synsets and 207,019 word-sense pairs. Among these words there are 27,006 polysemous words distributed as shown in the following table.

In this proposal, we are concerned with polysemy at the conceptual level only and we do not consider entities (e. g. proper names) like the proper name *java* (*an island in Indonesia south of Borneo; one of the world's most densely populated regions*).

Table 1:Distribution of polysemous words in WordNet 2.1

	Nouns	**Verbs**	**Adjectives**	**Adverbs**
#Words	15776	5227	5252	751

The number of senses a polysemous word may range from 2 senses to more than 50 senses in some rare cases such as the verb break which has 59 senses. Table 2 shows the distribution of polysemous words according to the number of senses they have. As we can see in Table 2, the most occurring numbers of senses are two, three, and four respectively.

Table 2:Distribution of the polysemous words according to the number of senses

#senses	Nouns	Verbs	Adjectives
2	10186 ≈ 64%	2534 ≈ 48.5%	3408 ≈ 65%
3	2968 ≈ 19%	1090 ≈ 20%	1041 ≈ 20%
4	1186 ≈ 7%	607 ≈ 11.5%	388 ≈ 7.5%
5	594 ≈ 3.5%	356 ≈ 7%	173 ≈ 3%
6	297 ≈ 2%	203 ≈ 4%	92 ≈ 1.8%
7	207 < 1.5%	128 < 2.5%	52 < 1%
8	100 < 0.7%	72 <1.5	20 < 0.5%
9	89 < 0.6%	50 < 1%	13 < 0.3%
10	57 < 0.4%	41 % 0.8%	18 < 0.4%
>10	110 ≈ 0.7%	146 ≈ 2.8%	47 ≈ 0.9%

A polysemous word in WordNet can be contrastive (e. g. its meanings are homonyms) as in (3), complementary (e.g. its meanings are related) as in (4), or a combination of both as in (5).

(3) **saki – contrastive polysemy:**

#1: Japanese alcoholic beverage made from fermented rice; usually served hot.

#2: small arboreal monkey of tropical South America with long hair and bushy nonprehensile tail.

(4) **mitzvah - complementary polysemy:**

#1: (Judaism) a good deed performed out of religious duty

#2: (Judaism) a precept or commandment of the Jewish law

(5) : **bass: combination of contrastive and polysemy, where the senses (1,2,3,6,7) and (4,5,8) are complementary that can be grouped into two contrastive meanings**.

#1: the lowest part of the musical range

#2: the lowest part in polyphonic music

#3: an adult male singer with the lowest voice

#4: the lean flesh of a saltwater fish of the family Serranidae

#5: any of various North American freshwater fish with lean flesh (especially of the genus Micropterus)

#6: the lowest adult male singing voice

#7: the member with the lowest range of a family of musical instruments

#8: nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes

Whether a word is contrastive, complementary, or systematic polysemous is not provided directly in WordNet. According to [12], around 95% of the words in WordNet are complementary

words, while only 5% of them are really homonyms. Nevertheless, identifying and separating contrastive words from complementary ones do not necessarily mean that the identified complementary words are systematic polysemous and hence can be organized into systematic polysemy classes. Analyzing complementary words shows that WordNet does not contain systematic and contrastive polysemous words only. In some cases, some meanings are redundant. The meanings of the word *drawers* in (6) for example have the same semantic structure and could be replaced by one sense that reflects the fact that *drawers* are worn by men and women. In other cases, we can detect missing relation between the meanings of a polysemous word as in (7). It is clear that the second meaning of *shield* is a special case of the first. We can also detect cases in which the meanings of a polysemous word are special meanings of a missing parent such as in (8), where it is clear that the senses describe two types of *green snake***,** the African and the American ones, respectively.

(6) **drawers – an example for redundant meanings:**

#1: underpants worn by men

#2: underpants worn by women

(7**) shield – an example for missing relations between meanings:**

#1: aquatic plant with floating oval leaves and purple flowers; in lakes and slow-moving streams;

suitable for aquariums

#2: common aquatic plant of eastern North America having floating and submerged leaves and white yellow-spotted flowers).

(8) **green snake – an example for missing parent that generalizes meanings:**

#1: any of numerous African colubrid snakes

#2: either of two North American chiefly insectivorous snakes that are green in color

To understand the problematic of redundancy and missing information in WordNet, let's consider the bass example again. The meanings #5 and #8 are similar and could be collapsed to a single meaning or we can at least consider the meaning #5 as a special case of #8. Let's assume that we collapsed the two meanings #5 and #8 into a single meaning #8' for example. In this case, the relation between #8' and #4 becomes clearer. The meaning #8' describes *bass* as a *fish* (*animal*), while the meaning #4 describes *bass* as *the flesh of the fish bass* (*food*). That *bass* is *fish* in #8' predicts that the word *bass* has another meaning, namely the meaning of food which is described by the sense #4. That means the polysemy type of meanings #8' and #4 is systematic. If we organize the relation between #5 and #8 such that #5 is a special case of #8 (and so #5 inherits all relations and features of #8), the same observation above holds between the senses #8 and #4.

Cleaning WordNet from redundancy and establishing the missing relations between the senses as in the previous examples is the first step towards organizing complementary words in systematic polysemy classes. Until here, the polysemy problem is not solved. In systematic polysemy classes, we can find also cases, where the distinction between the meanings is difficult, even for experienced text annotators as in the following example [20].

(9) a – **an example sentence for the meaning of pressure as the state of being under pressure :**

It wasn't just the pressure of work, although that was the excuse I often used, even to myself.

(9) b – an **example sentence for the meaning of pressure as an attribute:**

As a strike continues, these parties increase their pressure on the industry to reach an agreement.

In addition to systematic polysemy as a subset of complementary polysemy, we should be aware of other complementary polysemy types that are not systematic. These types of complementary polysemy can be explained through various linguistic phenomena such as the semantic shift (known also as semantic change, semantic progression) [15]. Other types can be explained through the cultural diversity of the speakers of the same language, where some cultures may use the broad meaning of a word while other cultures use a narrow meaning of that word. It is also possible to have slightly different meanings of the same word in different domains. For example, the word *polysemy* itself is polysemous. While linguistics uses the general meaning of the word *polysemy* (word with many meanings), in the polysemy reduction community, the term is widely used to refer to words that have related meanings (complementary polysemy).

Analyzing the previous examples, we can say that any approach for solving the polysemy problem should involve the following steps:

- Cleaning process to determine and treat the cases of redundancy, missing relations between synsets, or missing parents that connect the special meanings of the word.
- Identifying homographs and separating them from complementary ones.
- Identifying the systematic polysemy classes and assigning systematic polysemous words to their corresponding systematic polysemy classes.
- Identifying other (not systematic) complementary words.
- Analyzing the resulting polysemy classes in order to determine how to organize terms belonging to these classes so that applications requiring word sense disambiguation, such as semantic search, get sound results in terms of precision and recall.

3. Polysemy Reduction Approaches

Polysemy has been addressed in two main fields: in Information Retrieval (IR), to increase effectiveness of IR systems [6], and in word sense disambiguation (WSD), where the focus is on complementary polysemy and on how to identify the meaning of polysemous words in a given context. IR approaches aim to produce more coarse-grained lexical resources of existing fine-grained ones such as WordNet, i.e. polysemy reduction. WSD approaches focus on the recognition and identification of the intended meaning of ambiguous polysemous words using the surrounding context. In polysemy reduction, the senses are clustered such that each group contains related polysemous words

[2][10]. They are called homograph clusters. Once the clusters have been identified, the senses in each cluster are merged. Applying this approach should organize the meanings of the word bass in section two in two polysemy clusters as illustrated in figure 3. To achieve this task, several strategies have been introduced. These strategies can be mainly categorized in semantic-based and statistical-based strategies [6]. Some approaches combine both strategies [10]. Although results of applications of these approaches are reported, these results are taken usually from applying them on sample data sets and there is no way to verify these results independently.

Polysemy reduction approaches typically rely on the application of some detection rules such as: If S1 and S2 are two synsets containing at least two words, and if S1 and S2 contain the same words, then S1 and S2 can be collapsed together into one single synset [10]. However, applying this rule may wrongly result in merging two different senses as in the following example:

#2 smoke, smoking -- a hot vapor containing fine particles of carbon being produced by combustion; "the fire produced a tower of black smoke that could be seen for miles"

#7 smoke, smoking -- the act of smoking tobacco or other substances; "he went outside for a smoke"; "smoking stinks"

To enhance the quality of polysemy reduction, some approaches combine semantic rules and semantic similarity between the glosses of the synsets [6]. Identifying the semantic similarity between synsets requires disambiguating the glosses of the synsets. In [18] some heuristics are used to disambiguate the words in WordNet glosses such that at each word the corresponding synset is associated. They reached a precision of around 87% (in gloss disambiguation). This lead to the generation of a new resource called extended WordNet (XWN)[2].

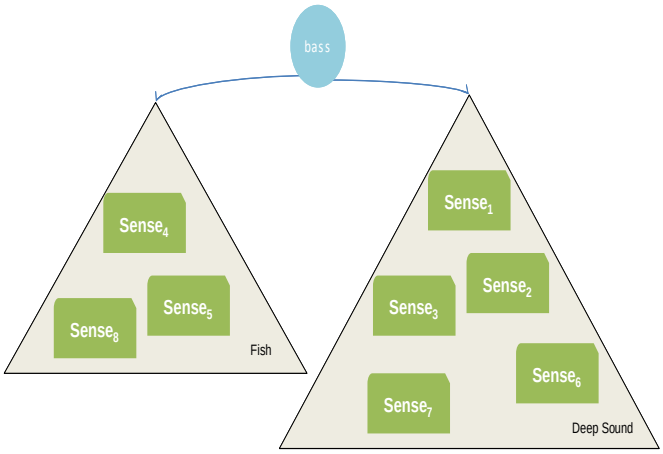

Figure 1: Polysemy clusters for the word bass

In the frame work of our study, we have implemented a polysemy reduction algorithm for testing and evaluation purposes. In this algorithm, we have implemented 9 polysemy reduction rules that are usually used in the polysemy reduction approaches [2], [10] since they encode sufficient conditions to reduce the meanings of polysemous words. Although the rules can identify and merge synsets with closed meanings in some cases, manual validation on sample results has shown very poor quality of the results.

In the following, we outline the applied rules and give statistics, in how many cases each rule was applicable.

Polysemy Reduction Rules:

Let S1 and S2 be two synsets in WordNet, then S1 and S2 can be merged if they fulfill at least one of the following rules:

Rule 1: If S1 and S2 are two synsets containing at least two words, and if S1 and S2 contain the same words.

Rule 2: If S1 and S2 are two synsets with the same hypernym or one of them is a direct hypernym of the other.

Rule 3: if S1 and S2 have the same direct hyponym synset or one is a direct hyponym of the other.

Rule 4: If S1 and S2 have the same coordinate terms (i.e., there exist a synset S3 such that S1 and S3 share a direct hypernym, and S2 and S3 share a direct hypernym).

Rule 5: If S1 and S2 are two synsets with at least K words in common (for example K= 1/2 of the words of the smaller synset).

Rule 6: If S1 and S2 have the same antonym.

Rule 7: S1 and S2 a have the same pertainym.

Rule 8: If S1 and S2 have similar to terms in common (i.e., there exist a synset S3 such that S1 is similar to S3, and S2 is similar to S3)

Rule 9: If S1 and S2 have related to terms in common (i.e., there exist a synset S3 such that S1 is related to S3, and S2 is related to S3)

The following table gives statistic, how many cases each rule was applicable.

Table 3: statistics of polysemy reduction rules

	Nouns	Verbs	Adjectives
Rule1	2037	664	365
Rule2	1125	0	0
Rule3	153	0	0
Rule4	4846	13402	9161
Rule5	3245	1938	782
Rule6	165	177	688
Rule7	89	0	203
Rule8	0	0	1686
Rule9	1171	861	293

By manual validation of the results, we have found that Rule 2 which is applicable on synsets that share the same parent (have the same path in the hierarchy) was the most successful rule. The other rules were less successful.

The other rules were successful in merging closed meanings such as the word *landing* in the following example but the error rate was high.

[2] http://xwn.hlt.utdallas.edu/

landing:

#1: the act of coming to land after a voyage

#2: the act of coming down to the earth (or other surface); "the plane made a smooth landing"; "his landing on his feet was catlike"

The poor quality of the results can be observed in two folds:

- Many synsets have not been merged although they have closed meanings such as the word *implementation* in the following example:

  implementation:

  #1: the act of implementing (providing a practical means for accomplishing something); carrying into effect

  #2: the act of accomplishing some aim or executing some order

- The algorithm merged homonyms such as the word *cakewalk* in the following example:

  cakewalk:

  #1:an easy accomplishment; "winning the tournament was a cakewalk for him"; "invading Iraq won't be a cakewalk"

  #2:a strutting dance based on a march; was performed in minstrel shows; originated as a competition among Black dancers to win a cake

In general, polysemy reduction has at least the following shortcomings:

1. Merging complementary senses may reduce the effectiveness of semantic search to the degree of syntactic search effectiveness.
2. The semantic relations among the senses in WordNet are in many cases not established correctly (missing or incorrect relations). This leads to an approximate result when detection rules are applied (insufficient precision and recall)
3. The identification of the homographs is as difficult as the identification of the complementary words.
4. These approaches can neither predict the semantic relations among the senses of polysemous words nor detect missing relations between such senses.

The shortcomings described above indicate that polysemy reduction does not solve the polysemy problem in linguistic resource. Nevertheless, it can be potentially used to solve part of the problem, namely the identification and merging of genuine redundant synsets.

4. XCORELEX: Systematic Polysemy Approach for Solving Polysemy

CORELEX[3], the first systematic polysemy lexical data base, follows the generative lexicon theory that distinguishes between systematic (also known as regular or logic) polysemy and homographs. Systematic polysemous words are systematic and predictable while homonyms are not regular and not predictable. The type of polysemy of the word *fish* for example is systematic since the meaning *food* can be predicted from the *animal* meaning and so the word *fish* belongs to the systematic class ***animal food***. The two meanings of fish describe two related aspects of *fish*: fish is an animal and fish is food.

That a word is systematic polysemous means: the meanings of this word are not homonyms and they describe different aspects of the same term. Following this distinction, CORELEX organizes polysemous nouns of WordNet 1.5 into 126 systematic polysemy classes. In order to use CORELEX in solving polysemy in WordNet, we should be aware of the following two points:

- Is CORELEX ready to be used?
  - Are there other polysemy classes not covered in CORELEX?
  - Do not these classes contain words that do not belong to them?
- How to use CORELEX?
  - How to deal with the very fine grained senses of WordNet?
  - How to organize the words in such a way that optimizes the precision and recall of semantic applications?

The systematic polysemy classes in CORELEX have been determined in a top down fashion. According to our experiments, a bottom up approach identifies new classes that are not detected by CORELEX or sub classes of CORELEX classes such as the classes ***food substance***, ***food chemical***, ***artifact quantity*** … . On the other hand, since there was no cleaning process carried out on WordNet by CORELEX construction, we assume that CORELEX classes may contain noisy words that do not belong to them due to the redundancy and inconsistency of WordNet mentioned in section two. The second point is related to the fine grained nature of WordNet. As we have seen in section two, the meanings of some CORELEX classes are very difficult to disambiguate and indistinguishable even for humans and hence we consider collapsing the meanings of the words belonging to such classes as appropriate. We plan to organize words of other polysemy classes in a light weight ontology structure and our hypothesis is that organizing the systematic polysemous words in this way shall optimize the precision and recall of semantic applications that will use the resulting lexical resource. In the following we describe the details of establishing XCORELEX as an extension of CORELEX for solving polysemy in WordNet:

WordNet Cleaning:

Here we identify all words whose senses have the same path modulo the last one or two nodes. According to the polysemy reduction approaches and our experiments, these words are most

[3] http://pages.cs.brandeis.edu/~paulb/CoreLex/corelex.html

likely to be redundant or have missing relations or parents. This process involves:

- identifying redundancy, or missing information.
- merging redundant senses, and
- establishing the missing relations and/or add missing parents.

XCORELEX Construction:

Here we build XCORELEX starting from CORELEX in bottom up fashion. This process involves:

- identifying CORELEX classes and populating these classes with the corresponding items, and
- Identifying and populating new classes and subclasses of the identified classes.
- Identifying Homonyms: Here we analyze the words that are not connected to systematic polysemy classes in order to identify homonyms.
- Identifying other (not systematic) complementary words.

XCORELEX Cleaning:

Here we examine the resulting systematic polysemy classes. This process involves:

- identify the classes whose senses are very fine grained,
- test and treat redundancy and missing information if there any, and
- test and identify homonyms or not systematic polysemy words: We do not exclude the possibility of detecting hyponyms or not systematic complementary words whose patterns occasionally match the patterns of some polysemy class.

XCORELEX Organization:

For each class, test the most appropriate way to organize the senses of the words of that class. Our hypothesis here is that most classes will be organized in a light weight ontology structure [21], but we do not exclude other possibilities such as adding missing relations between the senses in some cases.

XCORELEX Testing and evaluation:

Here, we examine the resulting systematic polysemy classes in order to evaluate of correctness of the applied procedure, were we test the quality of the classes and detect any words that do not belong to them. On the other hand, we will experiment the new WordNet in semantic search [22] where we examine the quality of the results obtained and measure how much we will improve the semantic search efficiency.

We start in our approach by addressing polysemous nouns. Other POS categories will be processed in subsequent phases.

In the following, we describe the procedure that we are going to apply in order to reconstruct CORLEX from WordNet 2.1 which was originally constructed from WordNet 1.5.

CORELEX Reconstruction - Identification of the polysemy patterns and homographs:

In the following procedure, we assign nouns to their corresponding systematic polysemy class, also called pattern. Note that each noun can potentially belong to more than one class. The residual nouns not falling in any polysemy class will be manually processed to identify homographs.

CORELEX Reconstruction Procedure

CORELEX is based on a set of polysemy patterns, such as animal food, determined in a top down fashion similarly to the approach in [3], but going one level down. In this approach, synsets (for a given word) falling in a pattern are systematic polysemous and therefore will be assigned to their corresponding systematic polysemy class.

In our approach, we start from the patterns in CORELEX, but we will extend them taking into account similarities between synsets as explained below:

- **Distinct Paths:**

  Search for all words that belong to a CORELEX systematic polysemy class and have distinct paths. For each systematic polysemy class we identify the words that:

  a. Have the same number of senses as the length of the pattern
  b. Each sense is subsumed exactly by one category in the pattern

For instance, if the pattern is animal food (length two), we look at all the words having exactly one synset subsumed by animal and one subsumed by food.

New patterns (w.r.t. CORELEX) may emerge from the analysis of the words having (or not having, to determine exceptions or sub-patterns – see partial matching below) the properties above.

- **Overlapped Paths:**

  Search for all words having overlapped paths, i.e. sharing a large portion of the ancestors from the root. This task is new w.r.t. CORELEX.

First identify words whose synsets have exactly the same path. This means that the synsets are siblings. We think that these synsets can be merged, optionally with manual validation. For instance, the following two synsets can be safely merged:

1. duke -- (a British peer of the highest rank)

2. duke -- (a nobleman (in various countries) of high rank)

For the words whose synsets have the same path modulo the last one or two nodes, compute their frequency to eventually determine new patterns, as follows:

a. Create an empty list of patterns and compute the frequency of the new patterns. For each word whose synsets have overlapped paths:
   i. Find the first distinguishing nodes
   ii. Create a pattern that corresponds to the synsets of the distinguishing nodes
   iii. If the pattern is already in the list then increment its frequency; otherwise add it to the list with frequency 1.

b. Exclude all patterns under a given frequency threshold (otherwise it is not a pattern)

c. Add identified patterns to the polysemous patterns and manually validate them

- **Partial Matching:**

Search for all words that have a partial match with a systematic polysemy pattern. For partial match we mean that the number of synsets for the word is greater o lower than the length of the pattern. These words may indicate the presence of a new patterns, sub-patterns or homonymy. It is also possible such words to indicate gaps in WordNet. For instance, the word museum has one sense only, while it should be part of the polysemous class artifact physical-object social-group of length three (the same of university and bank, even if bank has other senses also) For example, assume we are analyzing the CORELEX pattern act time (length two), and we find many words having three senses where two of them are subsumed by act and time but the third one is subsumed by another common synset event. This might indicate that act time event is a new pattern of length three.

Conversely, assume we are analyzing a pattern AAA BBB CCC and that not all the words turn out to belong to this pattern, but a significant subset of them are subsumed by only AAA and BBB. This might be an indication of either that AAA BBB is a new pattern (or sub-pattern) or that there are some missing synsets (such as the museum example).

- **Remaining words:**

Try to identify possible new patterns that are not discovered in previous steps. New patterns may correspond to new classes that are not covered in the CORELEX polysemous classes such as the patterns communication measure, animal material, quality trait (discovered thanks to some preliminary tests). Some other patterns may correspond to subclasses of CORELEX polysemous classes such as the pattern food-fish fish that can be seen as a sub-pattern of the pattern food animal. Note that fish is more specific than animal and food-fish is more specific that food.

The residual words will be manually analyzed to identify homographs. From some early experiments they should be less than 1000.

5. Conclusion and further Work

In this document, we proposed how to solve polysemy based on CORELEX approach. Our approach is different from the state of the art polysemy reduction approaches in that it deals and covers all polysemy cases rather than redundant cases only. Furthermore it covers cases of missing information. It is also different from CORELEX in such that it guarantees the quality of the resulting polysemy classes and organizes these classes in the best way that maximizes the quality of semantic search results. As a next step, we will test our procedures on lexical resources in other languages such as the Italian MultiWordNet [9], and examine how we can generalize our approach on multilingual lexical resources.

The main contributions of this work are at two levels:

At the conceptual level, we are providing a new foundation towards the problem of polysemy. At the implementation level, we aim to improve the quality the NLP and knowledge-based applications, especially in the field of the semantic search.

6. REFERENCES

[1] Martha Palmer, Hoa Trang Dang, Christiane Fellbaum. *Making fine-grained and coarse-grained sense distinctions, both manually and automatically*

[2] Reza Hemayati , Weiyi Meng , and Clement Yu. *Semantic-Based Grouping of Search Engine Results Using WordNet*

[3] Massimiliano Ciaramita, Mark Johnson. *Supersense Tagging of Unknown Nouns in WordNet*

[4] George A. Miller Florentina Hristea. *WordNet Nouns: Classes and Instances*

[5] Kanjana Jiamjitvanich and Mikalai Yatskevich. *REDUCING POLYSEMY IN WORDNET*

[6] ROBERTO NAVIGLI. *Word Sense Disambiguation: A Survey*

[7] Giunchiglia, Fausto and Zaihrayeu, Ilya. Lightweight Ontologies. Technical Report DIT-07-071, Department of Information Engineering and Computer Science, University of Trento. The Encyclopedia of Database Systems.

[8] J. Pustejovsky, The Generative Lexicon, MIT Press, Cambridge, MA, (1995)

[9] Luisa Bentivogli and Emanuele Pianta. *Extending WordNet with Syntagmatic Information*

[10] Rada Mihalcea**,** Dan I. Moldovan, *EZ.WordNet: principles for automatic generation of a coarse grained WordNet*

[11] Iraide Ibarretxe-Antuñano, *PREDICTABLE VS. UNPREDICTABLE POLYSEMY*

[12] Paul Buitelaar, CORELEX: *An Ontology of Systematic Polysemous Classes*

[13] NERLICH, B. / D.D. CLARKE. *Polysemy and flexibility: introduction and overview.* I*n: Nerlich, B./Z. Todd/V. Herman/D.D. Clarke (Hg.), Polysemy. Flexible Patterns of meaning in Mind and Language. Berlin, New York: Mouton de Gruyter 2003, 3-29.*

[14] Giunchiglia, Fausto and Maltese, Vincenzo and Farazi, Feroz and Dutta, Biswanath. *GeoWordNet: a resource for geo-spatial applications*. In the Proceedings of the 7th Extended Semantic Web Conference (ESWC), 2010.

[15] P. P. Buitelaar, CORELEX: Systematic Polysemy and Underspecification, PhD thesis, Brandeis University, Department of Computer Science, (1998)

[16] R. Snow, S. Prakash, D. Jurafsky, and A. Ng. *Learning to merge word senses*.In Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.

[17] R. Navigli. Meaningful clustering of senses helps boost word sense disambiguation performance. In COLING-ACL 2006, 2006.

[18] G. A. Miller, R. Beckwith, Ch. Fellbaum, D. Gross and K. Miller, Introduction to wordnet: An on-line lexical database, International Journal of Lexicography, (1990).

[19] Verdezoto, N. and Vieu, L. (2011). "Towards semi-automatic methods for improving WordNet", to be published in Proceedings of the 9th International Conference on Computational Semantics - Oxford, UK - January 2011.

[20] Noriko Tomuro, *Systematic Polysemy and Inter-Annotator Disagreement: emirecal Examinations*, DePaul University

[21] Giunchiglia, Fausto and Dutta, Biswanath and Maltese, Vincenzo. Faceted lightweight ontologies. Technical Report ,In: "Conceptual Modeling: Foundations and Applications", Alex Borgida, Vinay Chaudhri, Paolo Giorgini, Eric Yu (Eds.) LNCS 5600 Springer

[22] Fausto Giunchiglia, Uladzimir Kharkevich, Ilya Zaihrayeu: Concept Search. ESWC 2009: 429-444