Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, TX 76019

Concept Based Information Access Using Ontologies and Latent Semantic Analysis

Rifat Ozcan, Y. Alp Aslandogan
{ozcan,alp}@cse.uta.edu

Technical Report CSE-2004-8

Concept-based Information Retrieval Using Ontologies and Latent Semantic Analysis

Rifat Ozcan
Dept. of Computer Science and Engineering
University of Texas at Arlington, U.S.A
ozcan@cse.uta.edu

Y. Alp Aslandogan
Dept. of Computer Science and Engineering
University of Texas at Arlington, U.S.A
alp@cse.uta.edu

Abstract

Concept-based access to information promises important benefits over keyword-based access. One of these benefits is the ability to take advantage of semantic relationships among concepts in finding relevant documents. Another benefit is the elimination of irrelevant documents by identifying conceptual mismatches. Concepts are mental structures. Words and phrases are the linguistic representatives of concepts. Due to the inherent conciseness of natural language, words can represent multiple concepts and different words may represent the same or very similar concepts. Word Sense Disambiguation attempts to resolve this ambiguity by pinpointing which concept is represented by a word or phrase in a context. The use of an ontology facilitates identification of related concepts and their linguistic representatives given a key concept. Latent semantic analysis, on the other hand, attempts to reveal the hidden conceptual relationships among words and phrases based on linguistic usage patterns. In this work we explore the potential of concept-based information access via these two mechanisms. We apply these techniques in three domains and examine under what circumstances concept-based access becomes feasible and improves user experience.

1 Introduction

The amount of information that is accessible to an ordinary person today is mind-boggling. While a few centuries ago people were struggling to access information, today many are struggling to eliminate the irrelevant information that reaches them through various channels. The information needs of people are in concept space. Keyword based access to information is sometimes unsatisfactory since it works in word space. Words represent concepts in human language but the mapping from words to concepts is many-to-many. That means one concept may be represented with many different words (synonym) and one word may represent many different concepts (polysemy). This mapping problem is known as Word Sense Disambiguation. Secondly, since concepts are abstract entities, representing them is another problem. In part of this study, we used WordNet 2.0 [14] as our knowledge base. It contains synsets (set of synonym words representing the same concept) and their relationships with other synsets. Two basic relationships among synsets are is-a and part-of/member-of relationships.

In this thesis, we present two alternate ways for concept identification: one is based on identifying concept through word sense disambiguation (WSD) process and the other one is based on representing concepts through neighboring words using domain specific corpus. In the first approach we combined several WSD methods using some evidence combination techniques to achieve good performance. Then we tested our method on some IR test collections against traditional word-based indexing. The results show that concept-based IR is more successful on short queries and short documents.

Secondly, we identified the concepts in a domain specific corpus based on the assumption that only one sense of a word relevant for a domain. We achieved the co-occurring words of a word through infomap-nlp program [11] that uses a variant of latent semantic analysis. Then we choose top-5 most relevant words using a measure of concept relatedness. These words are tested with query expansion process. We achieved very encouraging results.

The organization of the paper is as follows: Section 2 describes the concept-based IR process and its challenges. Section 3 presents our approach to achieve concept-based access using evidence combination of several WSD techniques and query expansion with related concepts. Section 4 describes the evaluation process of our approach. Section 5 discusses the results of WSD and concept-based indexing experiments. Section 6 gives detailed related work about concept-based indexing such as concept representation methods, several WSD techniques and building knowledge repositories.

2 Challenges of Concept-based Access

The Figure 1 shows the steps in a concept-based IR system.

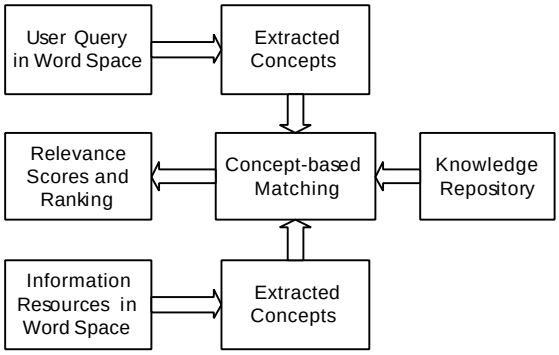

Figure 1: Concept Based IR System

Firstly, representations of information resources and user query are based on concepts. Since concepts are abstract entities, representing them is another problem. The way we choose to represent concepts affects all other parts of the system.

Secondly, we need to identify concepts in information resources and user queries. Since words can be ambiguous such that one word may represent many concepts, we need to disambiguate them. This task is known as Word Sense Disambiguation (WSD) and in the later sections some approaches to this problem are mentioned.

Finally, we need to do conceptual matching between extracted concepts. At this stage it is easy to find exact concept matching but the important part is to match remaining relevant concepts with the help of knowledge repository that is used. The knowledge repository gives information about concepts and their relationships with other concepts. At this point which relationships to choose to go to relevant concepts is a problem. So this stage requires a knowledge repository that does not miss any concepts and any relationships in the application domain. Building such a knowledge base is another challenge for concept-based access. Some approaches are mentioned in later sections.

3 Our Approach

We represent a concept as a node in ontology. Ontology is defined in [9] as follows:

Definition: Ontology

...an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. [9]

There are many ways to represent concepts and conceptual relationships in ontology. In our case, we use semantic network representation as in directed labeled graph. It is a simplified conceptual graph. We represent conceptual relationships as edges between nodes in the graph rather than representing them as nodes like in conceptual graphs. The detailed information on conceptual graphs can be found in related work section. Our ontology representation schema is defined below:

Definition: Our Ontology Representation

G = (V, E)
V = <Concept Identifier, Concept Label, [Concept Description] >
E = <u, v, relationshipName> where u,v ∈ V
Concept Identifier = <literals and numbers>
Concept Label = < string literal>
Concept Description = <character string>

G represents the semantic network graph consisting of vertices V as concepts and edges E as relationship among concepts. A concept is uniquely identified by a Concept Identifier and it has Concept Label. Concept description is a textual description of the concept like dictionary definitions. Conceptual relations are represented using two concepts and a relationship name. Relationship name is required since there can be many types of conceptual relationships among concepts such as generalization/specialization, part-of etc.

We are currently using WordNet ontology but the general architecture of the system applies to other kinds of ontologies as well. WordNet is a general domain ontology. A concept is represented as a synset that is a set of synonym words that represent the same conceptual entity in real world.

3.1 Concept Identification

We apply Word Sense Disambiguation (WSD) to identify concepts. We use WordNet sense distinctions as predefined set of senses. Our technique is based on evidence combination of supervised and unsupervised WSD methods.

Secondly, we identify the concepts in a domain with its related words. The relatedness of two words is identified using similarity of their content neighboring words using Latent Semantic Analysis technique.

3.1.1 Syntalex [15]

Syntalex is a supervised WSD system developed by Mohammad and Pedersen [15]. Its feature space consists of bigrams and POS tags of target and surrounding words. The system is evaluated with different combinations of these features with varying size of the context window. The experiments show that ensemble of bigrams and POS features gives the best precision values.

The logic behind choosing Syntalex to do evidence combination with our WSD approach is that it uses the local context information and ours focuses on the topical context information. In a way, we aimed to see an improvement by combination of two different WSD systems. This shows also that these systems complement each other.

3.1.2. WSD based on Domain Labels

WordNet lacks topical relationships among concepts that are related to each other since they are used in the same topic or domain. For example “election” (in the sense that “a vote to select the winner of a position or political office;”) and “candidate” (in the political candidate meaning) are two related concept because of topical relationship between them. WordNet cannot handle this type of relatedness. Another problem in WordNet is fine granularity of sense distinctions. This makes WSD even harder.

In order to solve these two problems WordNet Domains [13] is developed. This resource is obtained by tagging each WordNet 1.6 sense with one or more domain labels such as “ARCHITECTURE, SPORT and MEDICINE”. Since there are significant amount of word senses that are not domain-dependent, a special label, “factotum”, is used for them.

The method we use here is a modified version of baseline algorithm defined in [1]. We count the support for each domain in a determined window (in our case we use whole document size as a contextual window) for each domain. The score coming from each sense of a word (noun, verb or adjective) is calculated using the following formula:

$$score_{D(w(i))} = \frac{tfidf}{(i+1) * N_w} \qquad (1)$$

If i^{th} sense of word w is tagged with domain D in WordNet Domains [13], then the score coming from this sense is proportional to the tfidf value of word w and inversely proportional to N_w, number of senses of word w, and the ranking of sense in WordNet. The reason is that words with higher *tfidf* values have much effect on the domains mentioned in a document. On the other hand we penalize the words that has lots of senses and senses that are in low rank according to WordNet ordering because order reflects the frequency of usage of that sense of the word.

Then we simply go over the nouns in the text and check scores of domains that the senses of the noun belongs and choose the max scored one. The performance measure of this WSD method is given in results section.

3.1.3 WSD based on Contextual Weights of Hypernyms

This is an unsupervised WSD approach that is similar to the method presented in [2]. It uses hypernym relationships among synsets (concepts) in WordNet.
The method has the following steps:

Step 1. Pre-process the text. This involves tokenization and Part of speech tagging using an improved version of Brills' POS tagger [3]. This phase also encompasses a named entity extraction routine that identifies and extracts interesting patterns by deploying regular expressions.

Step 2. Identify word senses and hypernyms. Each word w_i is converted to its morphological root mw_i using the morphological analyzer of Wordnet [14]**.** All the possible senses $\{s_1,s_2,\ldots,s_n\}$ of mw_i given its POS (which was identified in the previous step) are recorded. The ISA hierarchy $\{s_{ip},s_{i(p-1)},s_{i(p-2)},\ldots,s_{10}\}$ of each sense s_i is also recorded.

Step 3. Compute contextual weights. There is no pre-set contextual window size but limiting the size to a sub section within a document yields better results. Within a sub context we compute the contextual support frequency f_{ij} for each sense s_{ij} within the hypernym hierarchy of each sense s_i.

Step 4. Calculate sense score using contextual weights
We sum the contextual weights in the hypernym path of the sense. We multiply each contextual weight with decreasing coefficient when we go through the root. Then the sense that has max score is chosen.

$$Score(s_i) = \sum_{j=0}^{p} c_{ij} * f_{ij} \qquad (2)$$

f_{ij} : contextual support frequency for j^{th} hypernym of sense s_i
c_{ij} : coefficient used for this hypernym
p : The number of hypernyms of sense s_i

3.1.4 PageRanking Soft WSD Algorithm [12]

[12] presents a soft WSD algorithm that ranks all senses of a word based on their relevance to the text. This algorithm is adapted from Google's page ranking algorithm.

WordNet has similar graph structure as web. The nodes are web pages and links are the edges between nodes in web but in WordNet, synsets are nodes and relationships among synsets (hypernym/hyponym, meronym etc.) are edges. The idea is similar to page rank algorithm such that if many synsets are pointing to the sense then that synset is more appropriate for the text. We cannot give details of the algorithm here because of space constraints in the paper. (See [12] for details)

Our aim was to use this WSD technique as additional source for doing evidence combination in order to achieve high precision.

3.1.5 Evidence Combination of WSD techniques

We apply different techniques to combine answers of WSD methods such as using uncertainty values, voting based with uncertainty and using sense rankings. Each will be explained below:

3.1.5.1 Evidence Combination based on uncertainty values

The following uncertainty formula is used to calculate uncertainty of each method that is based on class differentiation quality. [21]

$$H(U) = 1 - \frac{K}{K-1}\sum_{i=1}^{K}(m(i) - \frac{1}{K})^2 \qquad (3)$$

K: the number of senses for a word listed in WordNet
m (i): the belief for i^{th} sense of the word. In our case it is the normalized score for each possible sense that is given as output of WSD method.

So we normalize scores for each possible sense. Then uncertainty values are calculated based on the above formula. Then we rely on the WSD technique that has least uncertainty value.

3.1.5.2. Evidence Combination based on Voting with uncertainty

In this case, we use the simple voting principle to do the evidence combination in the first phase. That means each WSD method gives its choice for the sense of the word. Then the sense that has maximum vote is chosen as the sense of the word. If there is a tie in the first phase, then we compare the uncertainty values of WSD sources and choose the sense with minimum uncertainty value.

3.1.5.3. Evidence Combination based on sense rankings

This evidence combination technique considers the sense rankings given by each WSD method. Instead of using only first senses given by each method like in the Voting case, this technique takes other probable senses into account. The following formula [7] is used to do evidence combination.

$$P(s) = \frac{\sum_{k} l_k \, rank_k(s)}{\sum_{s'}\sum_{k} l_k \, rank_k(s')} \qquad (4)$$

$$rank_k(s) = \left(\left|\{s' \mid P_k(s') > P_k(s)\}\right| + 1\right)^{-1} \qquad (5)$$

l_k is the weight assigned to WSD method k. That shows the reliability of k^{th} WSD method in some way. In our case we used equal weights for each method since they had very similar performances. $rank_k(s)$ is the rank score for sense s given by k^{th} WSD technique. This score is inversely proportional to the number of senses that are strictly more probable than sense s according to k^{th} WSD method output of sense probabilities.

3.4 Concept-based Search with Knowledge Repository

3.4.1 Concept-based IR as synset-based indexing

We used Vector Space model for IR. Since we don't do disambiguation of all content words, we used word based indexing for the terms that we didn't do disambiguation on them. These terms are nouns that we couldn't find in

WordNet, verbs, adjectives and adverbs. So indexing structure consists of two vectors: one is synset (concept) based and one is word-based (stemmed version of these words). Each synset has a unique offset that is an integer value. We used this integer, that we call the "concept-ID" to refer concepts in vectors. We use our knowledge repository, WordNet, to add related concepts to the synset vectors so that even we don't have exact match of the concept we can still find the related documents.

3.4.2 Concept-based IR using Latent Semantic Analysis and Query Expansion

Considering the difficulty of WSD techniques and state-of art results as around 70% precision (according to Senseval conferences that is held to compare different WSD algorithms), we decided to try another way to identify concepts in a document. We assume that only a single sense of a word is relevant to a domain. There are some senses of words that are domain independent in the sense that it can be used in many domains but since they don't contribute much to the general meaning of the document as domain specific words do, we don't consider them at this point.

We used the Infomap-nlp software [11] developed at Computational Semantics Laboratory of Stanford University. It takes a corpus as input and produces word-word similarity measures. It constructs a model that is called "WORDSPACE" that represents each word using the content bearing co-occurring words. Each content-bearing word represents a dimension in WORDSPACE. A variant of Latent Semantic Analysis is used to reduce the number of dimensions in word vectors. Then it computes the similarity of two term based on cosine similarity of these coocurrence vectors. At the end, it gives related terms of a word with normalized similarity measures. The underlying theory behind this software can be found in [24]. Since we assume that only one sense of a word is relevant for a domain then the output of the program can be considered as related concepts of a concept in the domain. So we can use these related concepts for query expansion in order to retrieve documents. Addition of related concepts to the query eliminates the documents that use a different sense of the concept and retrieves documents that do not contain the main concept but related concepts in it.

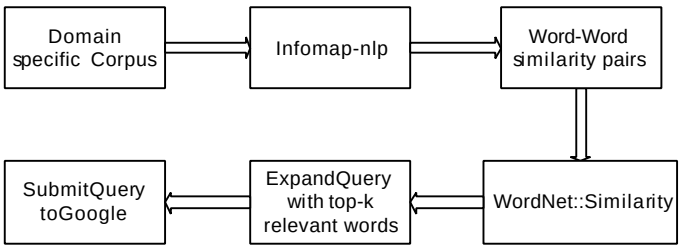

Figure 2: Query Expansion Process

The output of infomap-nlp program [11] contains some words that do not really related to the concept. So in some way we need to eliminate these words. Ted Pedersen et al. [18] developed a software called "WordNet::Similarity" that measures the similarity of concepts using several measure such as dictionary definitions, shortest paths among concepts in WordNet etc. We used this program to compute conceptual similarity of words given by infomap-nlp program. [11] Since WordNet::Similarity finds similarity among two senses of a word and we don't have senses but words, we compute all sense combinations of pairs of words and choose the sense pair that has highest similarity score. Then we eliminate words that have similarity score that is smaller than some threshold value.

4 Evaluation

4.1 Evaluation of WSD

We did experiments with some SENSEVAL-2 English All task data. These documents are sense tagged by human annotators. Choosing the first sense always is used as baseline in these experiments. We disambiguate only nouns currently.

4.2 Evaluation of Concept-based IR based on WSD

We used Cranfield test collection in order to test the effect of synset-based indexing on IR. The collection contains 1400 documents and 225 queries. In order to compare synset-based indexing, we also did experiments using traditional word based indexing (again same vector space model is used) using TFIDF values as weights.

Secondly, we used another dataset that consists of 2714 image captions and 47 queries. [25] This dataset is already used for concept based indexing by [25] but concepts (senses) are identified manually in the documents and queries. In our case we use our automatic WSD system to extract concepts in documents and queries. We used vector space model for indexing using synsets. The logic behind choosing this dataset is to see the effect of short queries in concept based indexing.

4.3 Evaluation of Concept-based IR based on Latent Semantic Indexing and Query Expansion

We used TEKS (Texas Essential Knowledge and Skills) curriculum as a corpus for infomap-nlp program. [11] This corpus consists of curriculums of following courses of Grades 5, 6, 7 and 8: Mathematics, Art, English Language Arts Reading, Health Education, Music, Physical Education, Science, Social Studies and Theatre. As a second corpus we used astronomy category of BankSearch [23] Dataset that is pre-classified dataset of 10000 html web page documents with 10 categories. So our astronomy corpus consists of 1000 web page documents.

Finally we constructed another corpus that consists of 4369 html pages downloaded from Recreation>Travel part of Google's web directory page. Before using this corpus as an input to the infomap-nlp [11] program, we preprocessed the files in order to remove html tags in them. In our experiments, we construct expanded query by adding top-5 most related concepts. For example, for query term, "reservation", for travel domain, we obtained the following most related concepts: Agent, Confirmation, Hotel, Customer, and Rates. Then the following query is formulated:
Original Query: reservation
Expanded Query: reservation AND (agent OR confirmation OR hotel OR customer OR rates)

5 Results

5.1 Results of WSD

The Table 1 shows the results that we got with SENSEVAL-2 dataset. CHW refers to WSD method based on Contextual Weights of Hypernyms and Domains refers to WSD based on Domain Labels. The coverage of our system is 92.1% only considering the nouns.

The Table 1 shows the results that we got with SENSEVAL-2 dataset. WSD based on Contextual Weights of Hypernyms method performs worse than baseline. On the other hand, WSD based on Domain Labels achieves close performance to baseline. Syntalex gives the best result if we don't consider the evidence combinations. Ranking based evidence combination gives the best result by increasing the performance of Syntalex by 3.18 %.

Table 1. WSD Experiment Results

WSD METHODS	Precision
Baseline (Choosing First Sense always)	56.33
WSD based on Contextual Weights of Hypernyms	51.64
WSD based on Domain Labels	55.87
Syntalex	58.27
PageRanking	53.67
Uncertainty (CWH + Syntalex + Baseline)	56.37
Voting (CWH+Domain+Syntalex+baseline)	58.31
Ranking (CWH+Domain+ Syntalex)	60.00
Ranking (CWH+Domain+ Syntalex+baseline)	61.45

5.2 Results of Concept based IR Experiments

5.2.1 Results of Concept based IR Experiments using WSD

Figure 3 is the graph that shows interpolated precision at standard recall points for experiments with Cranfield collection. The results show performance degradation from word based indexing when we use synset based. We think that the errors in WSD process might cause this result. One more reason might be the query sizes. Since queries consist of around 8 or 9 words, so indirectly ambiguity among words is resolved. Then there is not much need for WSD unless it achieves very good precision.

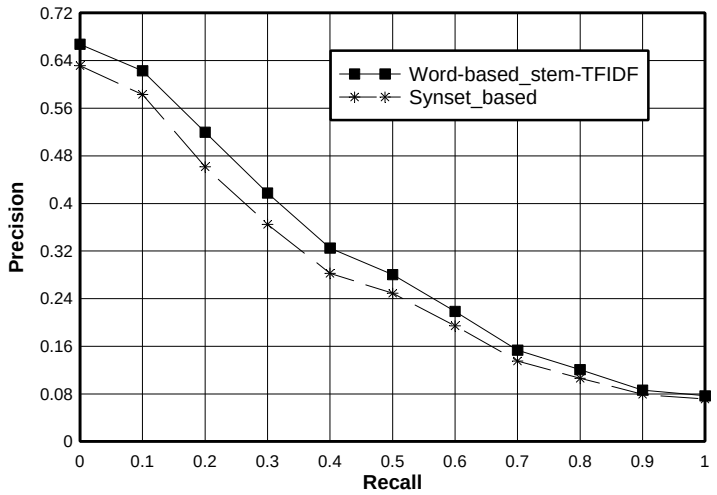

Figure 3. Synset-based Experiment Results-1

The Figure 4 shows the experiment results for image captions dataset. Word based indexing and three variants of concept based indexing interpolated precision values are shown at standard recall points. Synset based indexing clearly outperforms from word-based in this case. Adding hyponyms and hypernyms of unambiguous words also increases the overall precision values dramatically. Having short queries and documents cannot help word-based indexing but concept based indexing can enrich the content of the documents and queries by related concepts.

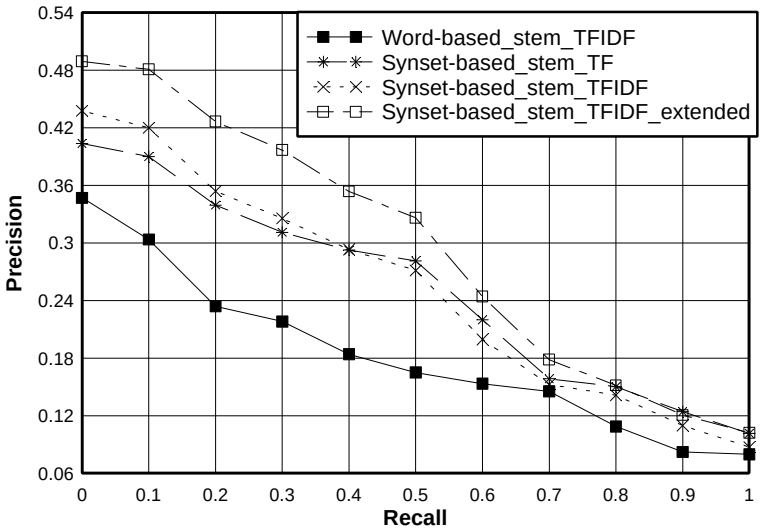

Figure 4. Synset-based Experiment Results-2

5.2.2 Results of Concept based IR Experiments using LSA and Query Expansion

The Table 2 shows the precision values for Top-50 documents retrieved for some queries using Google with original query term and expanding query using related terms determined using LSA and WordNet::Similarity programs. Early results are very encouraging.

Table 2. LSA based Query Expansion Results

Query	Domain	Google	Using LSA
Virus	Biology	6	100
Operation	Mathematics	2	26
Launch	Astronomy	68	98
Star	Astronomy	10	100
Galaxy	Astronomy	45	96
Venus	Astronomy	70	96
Historic home	Travel	88	94
Climbing	Travel	96	100
Reservation	Travel	68	100
Dolphins	Travel	80	92
Average		53.3	90.2

5.3 Discussion of Results

According to WSD experiment results, evidence combination techniques improve the performance of individual WSD methods significantly. Secondly, the results of concept based experiments show that we cannot outperform in long queries-long documents case since words in long queries implicitly disambiguates each other to some extent. So in order to make some improvement, we need to do disambiguation very accurately. On the other hand, we achieved significant improvement in image captions dataset. So, in this case word based matching is very poor since queries and documents have less number of words and simple matching of words does not suffice. On the other hand, short queries means less contextual information. So this means WSD is more difficult in short queries. The results show that the conceptual information we gained using synsets and especially by adding related concepts through relationships compensate the errors that we have in WSD process due to the lack of contextual information. Finally, we also tried concept identification through LSA. In this case we assumed that only one sense of a word is relevant for a specialized domain. The results show very significant improvement over original query. This indicates that LSA performs concept identification accurately. However, we need to do more large-scale experiments on this issue with more queries.

6 Related Work

6.1 Concept Representation Approaches

We mention two approaches here due to space constraints:

Frame-based Representation: The concepts are represented by concept nodes. A concept node is a frame-like structure that has slots that contain information about that concept such as its triggering word, patterns for extracting concept from text. [20]

Conceptual Graphs: The conceptual graph is a technique that is developed by Sowa [26], to represent knowledge. A conceptual graph g is a bipartite graph that has two kinds of nodes called, concepts and conceptual relations. [26] This graphical notation of conceptual relations is for human readability. This notation can be transformed into Knowledge Interchange Format (KIF) or predicate calculus notation for processing them in computers.

6.2 Concept Identification (Mapping words to concepts)

Word Sense Disambiguation (WSD) is already a research area by itself. Senseval conferences are being held for comparing different WSD programs. The results of these competitions show that we need more time to reach state of art in this field.

[10] presents detailed survey information on WSD techniques. The techniques used for WSD can be grouped into three: Supervised, Unsupervised and hybrid methods. Supervised WSD systems generally make use of local contextual information such as local collocation and POS information of surrounding words. [15] uses this type of WSD algorithm. Unsupervised methods generally use a lexical resource such as WordNet or LDOCE and topical contextual information surrounding the target word. They don't require any training data. As an example, [16] uses WordNet hierarchy for his WSD algorithm called "Specification Marks" and achieves 65.8% precision in some part of Semcor data. Hybrid methods use both of these techniques to improve WSD performance. [17] gives an example of this kind of method.

6.3 Building Cost of Knowledge Bases (Ontology, Dictionary)

There are some automatic and semi-automatic ontology construction techniques in the literature that can be considered as an alternative to manually building them. [19] presents a semi-automatic technique to construct semantic lexicons but the system only finds out the words for a category and the resulting lexicon does not have any relationships between words. Secondly, OntoMiner [5] automatically construct ontology from domain specific web sites by doing data mining on taxonomic structures in pages. This technique achieves a taxonomic structure of concepts but it has also lacks semantic relationships among different concepts.

If there is an initial semantic network of concepts for a domain that is previously built, there is another option of enriching this already built knowledge base by adding missing concepts and missing relationships. [4] presents an example of this case. The knowledge base is enriched by adding some new concepts and new relationships among concepts by compound name analysis.

6.4 WSD and Information Retrieval

The effect of WSD on Information Retrieval analyzed by many researches and some contradicting results achieved. [22] contains survey information on this issue. [27] used word sense indexing in their experiments but they couldn't achieved any improvement and even degradation in precision. They concluded that more accurate WSD programs could improve IR performance. On the other hand, [8] used manually disambiguated senses and analyzed the effect of perfect WSD on IR. They achieved significant improvements in IR performance. ([8] achieved 29% improvement in precision.)

6.5 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is proposed in order to overcome the polysemy and synonym problems of traditional keyword based retrieval. [6] The main goal of this technique is to reveal the underlying semantic structure of the documents by representing them in high dimensional space. LSA uses singular value decomposition in order to reduce the number of dimensions in the term-by-documents matrix. [6] tested the performance of latent semantic indexing in two test corpus: MED and CISI. The results show that LSA improves average precision of traditional term matching by 13 % in MED collection. However, they couldn't achieve any improvement in CISI experiments over classical approach. They thought that the homogenous structure of the dataset might cause this result. [6]

7 Future Work

We plan to do more WSD experiments using other datasets and also try to disambiguate verbs as well as nouns. Disambiguating verbs can increase the performance of synset based indexing.

8 References

[1] Bernardo Magnini and Carlo Strapparava. *Experiments in Word Domain Disambiguation for Parallel Texts.* Proceedings of the ACL workshop on Word Senses and Multilinguality, pag. 27-33, Hong Kong, 2000.

[2] Boppana P., Y. Alp Aslandogan, *The 3C Architecture: An XML Topic Maps-Based Framework for Integrating Content, Context and Common Knowledge About Multimedia*, IEEE International Conference on Information Integration and Reuse, Las Vegas, NV, 2003

[3] Brill, E. Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 722-727, Seattle, WA, 1994.

[4] Clark, P., J. Thompson, H. Holmback, L. Duncan. *Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search*. In Proc 12th Conf on Innovative Applications of AI (AAAI/IAAI'2000), pp 988-995, 2000

[5] Davulcu, H., S. Vadrevu, S. Nagarajan. *OntoMiner: Bootstrapping Ontologies From Overlapping Domain Specific Web Sites*. Poster presentation at the 13th International World Wide Web Conference, New York, NY., 2004

[6] Deerwester, S. & Dumais, S. T. & Furnas, G. W. & Landauer, T. K., & Harshman, R. *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, 41(6), pp.391-407, 1990

[7] Florian, R. and D. Yarowsky. ``Modeling Consensus: Classifier Combination for Word Sense Disambiguation." In *Proceedings of EMNLP'02*, pp 25--32 Philadelphia, PA, USA, 2002

[8] Gonzalo J, F. Verdejo, I. Chugur and J. Cigarran . Indexing with WordNet synsets can improve Text Retrieval, Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, Montreal., 1998

[9] Gruber, T. R., *A translation approach to portable ontologies. Knowledge Acquisition*, 5(2):199-220, 1993

[10] Ide, N., J.Véronis. *Word Sense Disambiguation: The State of the Art*. Special issue of Computational linguistics on Word Sense Disambiguation, 24:1, Pages 1-40, 1998.

[11] Infomap-nlp program, Available At: http://infomap-nlp.sourceforge.net/

[12] Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya, *Generic Text Summarization Using Wordnet*, Language Resources Engineering Conference, Barcelona, May, 2004.

[13] Magnini B. and Cavagliá G., *Integrating Subject field Codes into WordNet*. In Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, June 2000

[14] Miller, G.A., R. Beckwith, C. Felbaum, D. Gross, and K. Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235-244, 1990.

[15] Mohammad S. and Pedersen T., *Complementarity of Lexical and Simple Syntactic Features: The Syntalex Approach to Senseval-3.* In Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3), Barcelona, Spain, 2003.

[16] Montoyo, A. and Manuel Palomar: Word Sense Disambiguation with Specification Marks in Unrestricted Texts. DEXA Workshop, 103-107, 2000

[17] Montoyo, A., Suarez A. Palomar, M., *Combining supervised-unsupervised methods for Word Sense Disambiguation. 3er International conference on Intelligent Text Processing and Computational Linguistics -CICLing-2002.* Lecture Notes in Computer Science 2276: 156-164. México D.C. (México). February 2002

[18] Pedersen T, Patwardhan S. and Jason Michelizzi, *WordNet:: Similarity - Measuring the Relatedness of Concepts*. in the Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), San Jose, CA, 2004

[19] Riloff, E. and Shepherd, J. *A Corpus-Based Approach for Building Semantic Lexicons.* Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2), 1997

[20] Riloff E. and W. Lehnert, *Automated dictionary construction for information extraction from text*. In Proceedings of Ninth IEEE Conference on Artificial Intelligence for Applications, pages 93-99, 1993.

[21] Rujie, Liu. and Yuan Baozong, *A D-S Based Multi-Channel Information Fusion Method Using Classifier's Uncertainty Measurement*, Proceedings Of ICSP2000,pp. 1297-1300, 2000

[22] Sanderson, M.Retrieving with good sense. *Information Retrieval*, 2(1):49—69, 2000

[23] Sinka, M.P., Corne, D.W. A large benchmark dataset for web document clustering, in Abraham, A., Ruiz-del-Solar, J., Koeppen, M. (eds.), Soft Computing Systems: Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications, pp. 881-890, 2002

[24] Schutze, H. *Automatic word sense discrimination.* Computational Linguistics, 24(1):97–124, 1998.

[25] Smeaton, A.F. & Quigley, I. *Experiments on Using Semantic Distances Between Words in Image Caption Retrieval.* In Proceedings of ACM SIGIR Conference, 19: 174-180, 1996

[26] Sowa, J.F., Conceptual Structures: Information Processing in Minds and Machines, Addison-Wesley, Reading, Mass., 1984

[27] Voorhees E. *Using WordNet to Disambiguate Word Senses for Text Retrieval*, in Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 171-180, PA, June 1993.