Skimming in comparison

Relating of Concepts

Postgraduate paper in
Computer Science

Ole Torp Lassen
DIKU, Autumn 2006

Advisor:
Neil D. Jones

1 Introduction

When first I started out on my project, "Skimming for Context" - (*Lassen 2005*), I had no idea of what other work on similar topics had already been undertaken and documented.

Furthermore, as the project progressed, I deliberately avoided any and all contact with any such work because of a distinct wish to let my own reasoning work on its own and thereby avoid unintended copying or adaptation of the work of other researchers.

As was subsequently pointed out to me, I should of course have related my work to the rest of the work in the field, once my approach had matured sufficiently to distinguish itself from that of other projects.

Consequently, "Skimming for Context" does indeed represent an original piece of research and its theory is further experimented with in "Skimming for Paragrahs", (*Lassen 2006*). It does, however, remain wanting of proper orientation and positioning with in the field of **Word Sense Disambiguation(WSD)** and **Context Recognition (CR).**

This obvious lack of overview is what the present paper, "Skimming in Comparison", seeks to remedy.

A word on references. The very nature of a comparison project like this entails a lot of references to the work other people. In all cases where I have cited from or adapted such work I have tried to make clear reference to its respective source. Where ever such a reference is not present, the table, definition or figure can be assumed to be of my own devise.

2 Identifying Similarities

In this paper, I am going to focus on the general approach to **WSD**/**CR** that has been called **knowledge based** *methods*. These methods all rely on large external knowledgebases to provide basic semantic knowledge about concepts in "the world". Just like people now and then need a dictionary to pinpoint the meaning of a particular term, **Machine Readable Dictionaries** (**MRD**s) are intended to provide a static source of linguistic, semantic and pragmatic knowledge for automatic systems. The systems that I am comparing in this paper furthermore all employ the same **MRD**, namely WordNet, (*Miller 1990*).

Apart from **knowledge based** methods, however, there are several other, less related, approaches to **WSD** and **CR**. Some of these I will mention but I will not describe any of them in detail in this paper. A good general survey can be found in (*Ide & Véronis, 1998*).

2.1 Problem context - Widdows' model

Any project on **WSD** or **CR** must decide on how to analyse the interaction of words, concepts and contexts. D.Widdows attempts to facilitate straightforward comparison between such approaches by providing a "unified model of context and word-meaning", (*Widdows, 2003*). While his model needs some adaptation in order to be useful for my purposes, presenting it and its perceived shortcomings will serve as a good reintroduction to the problems and challenges of **WSD** and **CR**. It distinguishes between three spaces in the following way:

" W*, words:* *Primitive units of expression.*
Single words.
Parts of compound words.
Independent multiword expressions.

L*, lexicon:* *The available meanings to which signs refer.*
Traditional dictionaries.
Ontologies/Taxonomies.
Meanings collected from training examples.

C*, contexts:* *Pieces of linguistic data in which signs are observed.*
Sentences.
Immediate collocations.
Whole domains of knowledge. "
(source: Widdows, 2003)

Regarding Widdows' distinctions in turn, the space of W, is fairly self-explanatory. It ranges over explicit linguistic expressions. **Skimming**, however, restricts itself to analyse only nouns and that is also the case for the majority of the projects that I will describe in the present paper. Therefore, unless otherwise stated, W will in this paper range over only single English nouns - in their orthographic dictionary form.

With regard to L, it should be clear that it has to do with the meanings that the expressions in W may refer to. I do not, however, find the term *lexicon* entirely appropriate. I will instead use the term *meanings,* M, to range over word-meanings (or concepts, as it is). Furthermore, because W ranges over nouns only, it suffices for M to range over nominal concepts in this paper.

The third space , *C*, is much less straightforward. In particular, it seems clear to me that Widdows, in his description of the space, confuses two distinct meanings of the term *context*. Pseudo-formally, the two meanings can be represented as follows:

1) *context;* (~LEXICAL COLLOCATION): *context of x* : ANY TEXT THAT x, ($x \in \mathbb{W}$), OCCURS IN, or THOSE WORDS IMMEDIATELY ADJACENT TO x, ($x \in \mathbb{W}$), IN SUCH A TEXT.
2) *context;* (~CONCEPTUAL DOMAIN): *context of x* : ANY PARTICULAR DOMAIN OF KNOWLEDGE THAT EITHER
   a. x, ($x \in \mathbb{M}$), BELONGS TO
      OR
   b. IS DESCRIBED BY A TEXT THAT x, ($x \in \mathbb{W}$), OCCURS IN.

Both interpretations of the term are obviously very important to **WSD** and **CR** and they must be distinguished clearly from each other. Following the terminology that I adopted in (*Lassen 2005*), I will refer to the LEXICAL COLLOCATION of 1) using the term *sequence*, (S), and to the CONCEPTUAL DOMAIN of 2) using the term *context*, (C). This leads to the adaptation of Widdows' distinctions presented in Definition 1.

W, *words*:	Single English nouns.
M, *meanings*:	All possible distinct nominal concepts.
S, *sequences:*	Any sequence of English nouns as they occur in actual text.
C, contexts:	All possible situations, topics or domains of knowledge; real or imaginary; known, forgotten or yet un-experienced.

Definition 1 : Spaces of referential interactivity, adapted from Widdows.

With the respective extensions of the various spaces in place, I can reintroduce basic notions and relations accordingly as presented in Definition 2.

Def . : lexeme and realization:

(w, m) where
w ∈W,
m ∈M and
the interpretation w→ m is present in the **MRD**.

A concept m∈M , is said to be realizable in a text T if a word, w∈W, occurs in T that has m among its possible meanings, i.e.: if a lexeme (w, m) is possible in T. Consider an interpretation of T that involves the lexeme (w, m), then w is said to be the realization of m in T; likewise m is said to be realized in T by w.□

Def . : antonym - relation :

$(w1, m1) \leftrightarrow_{anto} (w2, m2)$, where
w1, w2 ∈ W,
m1, m2 ∈ M and
the two lexemes are recorded as antonyms in the **MRD**.

This the relation between opposites, i.e.: ((*dog*,DOG) - (*cat*,CAT)). It is important to realize that antonymy ranges over complete lexemes, rather than words or concepts in isolation. Antonymy is symmetric, horizontal and non-ordering.□

Def . : synonym - relation :

$(w1, m) \leftrightarrow_{syno} (w2, m)$, where
w1, w2 ∈ W,
m ∈ M and
the two words can map to the same sense in the **MRD**.

True synonymy is very rare and implies that two words means the exactly the same in any and all contexts and situations. The notion used in this definition is called near-synonymy, where the words may realize the same concept in some situations.□

Def . : hypernym - relation:

$m1 \rightarrow_{hyper} m2$ or its reverse
$m2 \rightarrow_{hypo} m1$ where
m1, m1∈M and
m1 is an immediate subordinate to m2 or vice versa in the **MRD**.

The hypernym-relation is the basic is-a relation between concepts (e.g.: DOG–CANINE). Hyponymy is the reverse of hypernymy. These relations are inherently transitive (i.e.: a $\rightarrow_{hyper}$ b and b $\rightarrow_{hyper}$ c entails that a $\rightarrow_{hyper}$ c). The hypernym-relation is vertically ordering.□

Def . : meronym - relation :

m1 $\rightarrow_{mero}$ m2 or its reverse
m2 $\rightarrow_{holo}$ m1 where
m1, m2 $\in$ M and
m1 is recorded as a part of m2 or vice versa in the **MRD**.

The meronym-relation ranges over concepts. It prototypically refers to the relation between a part-concept of a composite-concept and the composite-concept itself (e.g.: finger-hand). It extends naturally to cover also the relation between portion-substance, (e.g.: drop-liquid, grain-powder), member-group (e.g.: pup-litter) and other similar relations. The reverse relation is called holonymy and is like meronymy also transitive. Meronymy can be seen as horizontally ordering.□

Definition 2 : Some important general notions in WSD and CR.

2.2 Relatedness of concepts

An important subgoal in this comparison must be to choose the criteria for projects to compare.

Central to the *Skimming* method, is the idea that polysemous words can be sufficiently disambiguated by the semantic context in which they occur - that within a coherent semantic context no semantically misleading ambiguity will ideally occur.

This idea was inspired and justified by the easily observable phenomenon that people seldom realise any ambiguity in every day language use but instead un-effortly, and more or less unconsciously as well, choose sensible interpretations for ambiguous words. The phenomenon presumably relies on a fundamental and universal coherence through chains of semantic relations between concepts (meanings of words), that "go together" in the same semantic context.

In this view, tracking such conceptual coherences is a central key to the proper interpretation of polysemous words.

The projects that I chose as similar to *Skimming*, all rely on that same fundamental idea. They differ from *Skimming* - and from each other, primarily in their respective suggestive solutions to the following problem:

How to formalize conceptual coherence ?

In this paper some of these suggestions will be presented and discussed. Other topics of relevance will be touched upon as well.

3 Comparison in contrast

This chapter will describe three systems for *Context Recognition/ Word Sense Disambiguation* : My own *Skimming* - (*Lassen 2005*), *Semantic Distance* - (*Sussna 1993*) and finally *Conceptual Density* - (*Agirre & Rigau 1996*).

For each of these I will describe its theory, algorithm and results in order to discover their similarity and respective strengths and weaknesses. In order to demonstrate in proper contrast I will use the toy problem, presented in Figure 1**,** as a common example of the workings of the three systems. In this example is given a sequence of three two-way ambiguous nouns. The possible meanings of the nouns are all represented in a hierarchical semantic structure, describing their respective semantic relationships.

How the systems treat this problem will form a good basis for a discussion of their respective differences and similarities.

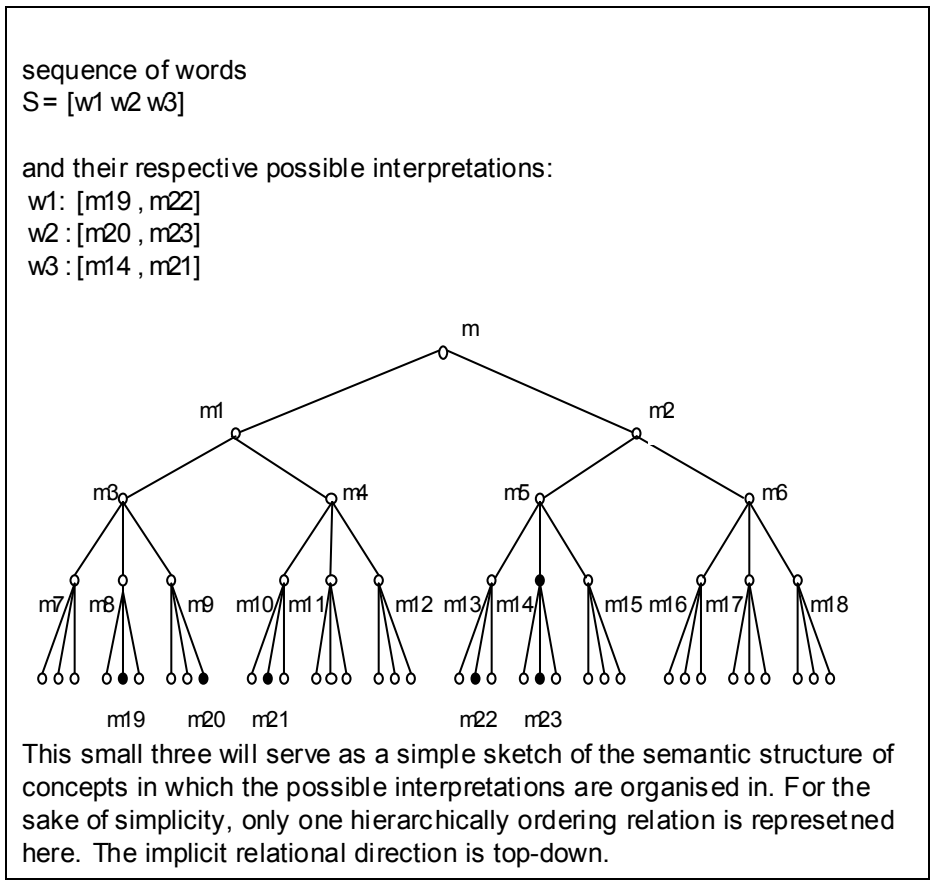


Figure 1 : Toy problem : a simple WSD/CR problem of WSD/CR be used as a common illustration for all the systems presented in this paper.

3.1 Skimming - neighbours only

I must begin by summing up what **Skimming** is, what it attempts to do and to what degree it succeeds in doing it. In addition to the definitions already presented in section 2 of this paper, the notion of (directly) connected lexemes is central to the *Skimming* algorithm and is formalised below.

Def . : *Connected lexemes :* Two lexemes,(w1,m1*)* and (w2,m2), are considered (directly) connected if and only if one of the following conditions are met : a) the lexemes are directly related via antonymy, or b) the respective concepts, m1 and m2, are directly related via either hypernymy or meronymy. □
NB : It important to realize that this definition emphasizes on the distinction between direct connections as opposed to indirect connections that are made up of chains of direct connections. The algorithm used in Skimming-prototype relies solely on direct connections, i.e.: adjacent nodes in the network.

Definition 3 : Directly connected lexemes – adjacent nodes in the network.

3.1.1 Theory – Skimming

Skimming is an experimental method for analysing arbitrary informative English text semantically and pragmatically. It attempts primarily to establish a representation of the topical context or contexts present in the source text. The proposal suggests interpreting the nouns of the source text in accordance with the joined theories of **Lexical Semantics**, see for instance (*Cruce 1986*), and **The Cooperative Principle** (*Grice 1975*). Essentially:

- Nouns in context are assumed to realize nominal concepts that are particular to that context. Concepts particular to the same contexts are assumed to be closer related than others.
- Concepts are assumed to be related to each other via one or more universal semantic relationships, presumed immediately accessible to the vast majority of observers, i.e.: human beings.
- Informative texts are assumed not to be misleading, i.e.: the author carefully chose his words in order for his text to be sufficiently clear to any intended reader.

All of which suggests that a topical context might be characterisable by

a) a set of closely related concepts particular to that context and
b) a set of "preferred" words realizing those concepts.

The more lexemes that are recognised as pertaining to a particular context, the more specific and detailed the characterization of that context will be.

An immediate consequence of this idea is that any successful **CR** of a text implies a certain degree of **WSD** of the polysemous words in the text. On the same note, any successful **WSD** of polysemous words in a text serves as a refining filter with respect to the context of that text and effectively contributes to the **CR** of it.

3.1.2 Algorithm - Skimming

In order to treat an informative English text, T, it is first divided into subsequences, S_1, S_2, ... S_k , each of which is assumed to treat involved contexts unambiguously[1].

In order to establish an representation of the topical context for a sequence S_i, the algorithm basically interprets as many of the nouns in S_i as possible in terms of the concepts and conceptual relationships represented in the **MRD** - WordNet. For a lexeme (w1,m1) to candidate for inclusion in an interpretation I of a sequence S_i, w1 must be a potential realization of m1 in S_i, and (w1,m1) must be directly connected to or from at least one other lexeme, (w2,m2), in that same interpretation, I. So if one lexeme is included in I then at least one other lexeme directly connected to the foist one, must also be included in I. The words, w1 and w2, of the included lexemes are then considered interpreted by I and thus are restrained from realizing other concepts in the interpretation I of S_i.

Instead of examining all the different lexemes possible in S_i, only relationships between lexemes, that are actually realized in S_i, have their arguments considered for inclusion. This way any connected lexeme possible in S_i is either chosen or excluded from that a interpretation. Since inclusion of one lexeme entails risk of exclusion for other lexemes, the process strongly favours lexemes that are tried first. Therefore the process is repeated, trying potential relationships in different orders. Each repetition potentially yield an alternative interpretation, I′, I″, ... of the sequence S_i.

Eventually, the interpretation of S_i that contains the most instances of connected lexemes is deemed the semantically most coherent - and therefore the most likely interpretation. This resulting set of inter-related lexemes is what I propose as a useable characterization of the context or contexts described by the respective text..

[1] Deciding on proper portions of text for analysis is no small problem in itself. The approach taken in (*Lassen 2005*) bases this decision on the original paragraphs placed in the text by its author. It must be pointed out that not all available text corpora of the day include this kind of information. In cases where no typographic information is available other means of proper portioning of the text must be sought out and applied, see also (*Lassen 2006*).

Figure 2 illustrates how our toy problem is analysed by skimming. It should be clear that *Skimming* imposes important consequences on the solution of the context recognition problem:

a) All directly connected nominal lexemes realizable in the sequence are considered for inclusion in its interpretation `I` in pairs as indicated by their respective relationships.
b) Sequence length directly reflects paragraph boundaries that are themselves assumed to reflect contextual boundaries. Because ambiguity is assumed virtually non existent within proper contextual boundaries, the Skimming system interprets sequences en bloc.
c) Any resulting interpretation `I` of a sequence `S`, is unambiguous, i.e.: inclusion of one lexeme, (`w1`, `m1`), excludes all other possible lexemes, i.e.: (`w1`, `m2`), where `m2` ≠ `m1`.
d) Only realizable lexemes that are directly connected is ever considered. Therefore, contexts extracted this way will in all but the rarest of cases only involve some of the nouns in sequence. Nouns in `S` that cannot be unambiguously interpreted as directly connected in the sequence are simply left un-interpreted.

3.1.3 Measures and results – Skimming

The Skimming prototype was applied to a small collection of sequences that varied in length from 17 to 60 nouns, with the average length being a little more than 25 nouns.

Of the words in each sequence from 0% to 63% were assigned an unambiguous meaning with the average being close to 20%.

The variation in length of these sequences bears on the fact that instead of a fixed window size the original paragraphs present in the source text were used as sequence boundaries.

20% certainly can not be considered a full interpretation of the sequence in question. It is likely that the experimental algorithm can be improved considerably. But even a small portion that is correctly interpreted may be very useful if it represents those concepts that are closest to the core of the context or contexts treated in the sequence. It may indeed serve as a very sound base for general distinction between contexts.

The remaining question must therefore be: how good are they - these system-found contexts/interpretations ? That question obviously have several facets to it. Eventually I decided to accept any interpretation that also a proficient human speaker of the language in question would consider plausible and useable in the respective context. For lack of better subjects, I referred to my own judgement in my capacity as a proficient speaker of the English language.

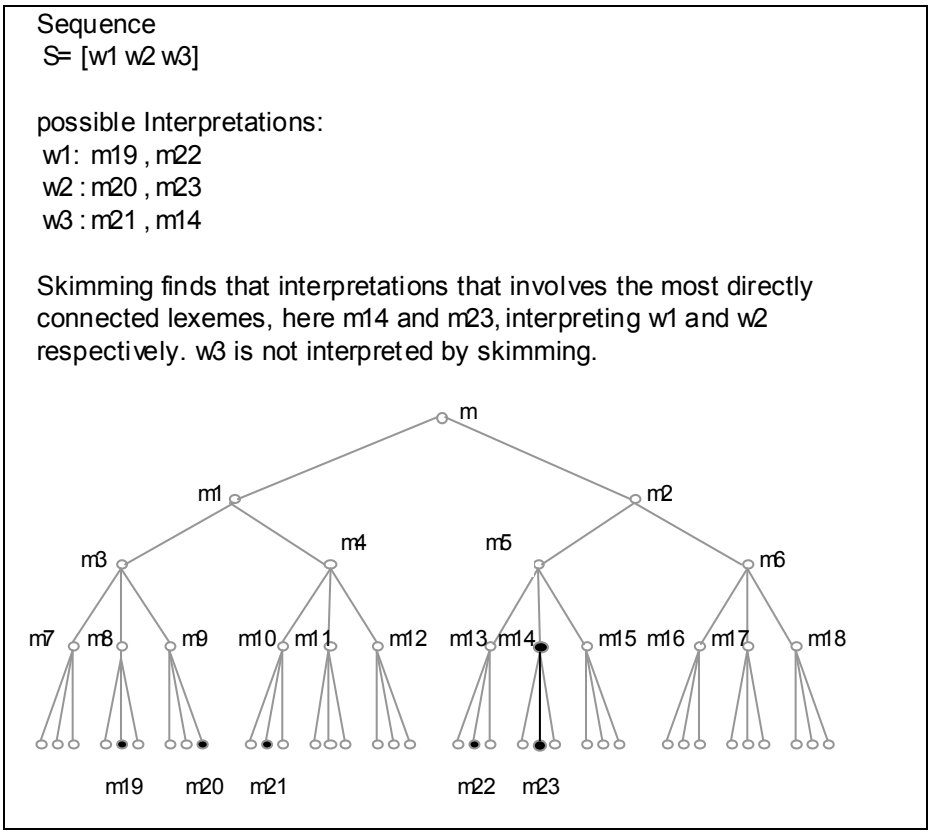


Figure 2 : Skimming considers only direct neighbours in the hierarchy as related. This clearly simplifies the problem significantly. While sometimes failing interpret all words in the sequence, we can be fairly sure of the relationship between the nouns that are interpreted.

This overtly subjective evaluation process of course is not entirely satisfactory, but at the time I simply had to make do. It simply involved looking over the lexemes found by the system and for each of them in turn decide whether or not it correlated with my own understanding of the respective text. Rather than to find plausible lexemes, it turned out to be much more straightforward to decide if a given lexeme was clearly wrong with respect to my understanding of the encompassing text.

I estimated that about 30% of the lexemes found by the system were mismatches. The remaining 70% are then estimated as fitting. Furthermore the general context of the sequences in question was deemed (equally subjectively) as reasonably well represented by about 60% of the system interpretations.

One way to increase both the relative portion of interpreted words and also the quality of the interpretation respective to a given text, could be to allow the system to consider paths of relationship rather than just single instances of relationship, and also allow for intermediate concepts, that need not be expressively realized in the text.

3.2 Semantic Distance - pursuing relationships

The first of two projects similar to Skimming, that I will include in this paper, is *Semantic Distance* by Michael Sussna, (*Sussna 1993*). His project is primarily concerned with the AI-related problem of information retrieval (IR), i.e.: given a topic or problem and a bank of documents, how can those documents that are relevant to the topic or problem be found automatically.

The motivation for employing word sense disambiguation in IR is nicely captured in the following excerpt from Sussna's paper:

> " Semantics-free, word-based information retrieval is thwarted by two complementary problems. First, search for relevant documents returns irrelevant items when all meanings of a search term is used, rather than just the intended one. This causes low *precision*. Second, relevant items are missed when they are indexed not under the actual search terms, but rather under related terms. This causes low *recall*. With semantics-free approaches there is generally no way to improve both precision and recall at the same time.
>
> Word sense disambiguation during document indexing should improve precision. ... " (source: *Sussna 1993)*

By document indexing is referred to the process of attributing to each document in the bank a content description derived from the document text itself. The later retrieval process will then basically involve comparing a particular search term to the content description of each document in the bank and returning the best matches.

The primary goal of Sussna's project, thus, is word sense disambiguation. It introduces a way of measuring semantic distance between concepts using a sophisticated edge weighting scheme. Also, a rigid method of evaluation is employed.

3.2.1 Theory – Semantic distance

Where the Skimming prototype did only consider neighbouring nodes in the semantic network, Sussna's system allows relationships between concepts that are arbitrarily far apart in the hierarchy - basically entailing that any two nominal concepts be related to each other to some degree. Therefore, where Skimming distinguished only between connected or unconnected - related or unrelated, Sussna introduces the notion of *semantic distance* to distinguish between different degrees of relatedness.

Given two words, `w1` and `w2`, the meanings of `w1` is examined for all possible relationships to the meanings of `w2`. Basically, those two concepts that are related to each other the closest are chosen as the proper interpretations of the words, `w1` and `w2`, respectively. In determining the closest relationship, Sussna simply solves the shortest path problem between two nodes in a graph. Because Sussna wants to distinguish also between different kinds of relations, the edges of the

graph are weighted accordingly. Sussna explains the weighting scheme of edges as follows:

> " Each edge consists of two inverse relations. Each relation has a weight between its own *min* and *max*, the point in this range for a particular edge depends on the number of edges of the same type leaving the same concept. This is the *type-specific fanout* (TSF) factor. TSF reflects the dilution of the *strength of connotation* between a source and target node as a function of the number of like relations that the source node has. The two inverse weights for an edge are averaged. The average is divided by the depth of the edge within the overall "tree". This process is called the *depth-relative scaling* and it is based on the observation that only-siblings deep in a tree are more closely related than only-sibling higher in the tree."
>
> (source: *Sussna 1993)*

The actual weighting function is defined in Definition 4. Basically, a given sequence T of n words, each of which may have more than one candidate meaning, is regarded. Each combination of n senses, one for each word, is then regarded as a possible interpretation, I_i. For each possible interpretation the pair wise distances between all pairs of senses are summed to arrive at an overall value, H(T). The interpretation that minimizes this overall value is the "winning" interpretation as shown in Definition 5 and its effect on the toy problem illustrated in Figure 3. We see that in the toy problem, *Semantic Distance* finds an interpretation for all the words in the sequence.

Sussna's describes the approach sketched in Definition 5 and refers to the technique calling it *mutual constraint* among terms. Operating with a fixed window size, n, the mutual constraint algorithm starts by simultaneously assigning senses to all n terms in the initial window. The algorithm then proceeds by moving the window, term by term, so that :

> "... just the middle term is assigned its sense. Record is kept of the winning sense, but when that term plays a role other than "middle term", its senses are allowed to fully wary. This gives the middle term full benefit of both previous and subsequent context. All senses of surrounding terms are considered, not just their winning senses."
>
> (source: *Sussna 1993)*

He describes also an alternate method where meanings are bound and frozen to words sequentially - first words first. In this approach, called *frozen past*, the (n+1)'st word is interpreted only with respect to the n previous words in the window, that have already been interpreted and have had their senses determined. Discussing the two approaches, he states that :

> "Mutual constraint is more appealing conceptually than frozen past but is exponential in the number of combinations of term senses that need to be tried. Frozen past avoids this combinatorial explosion by reducing the problem to essentially linear time processing time, since there are only as many "combinations" to try as there are sense of the single term being disambiguated." (source: *Sussna 1993)*

Def . : Weight $w(\texttt{m1}, \texttt{m2})$, (of relationship between directly connected concepts) :

Consider two concepts m1 and m2, directly connected via relation *r* from m1 to m2. Taking into regard that the inverse relation, *r'*, must also hold from m2 to m1, the relationship is weighted as follows :

$$w(\texttt{m1}, \texttt{m2}) = \frac{w(\texttt{m1} \rightarrow_r \texttt{m2}) + w(\texttt{m2} \rightarrow_{r'} \texttt{m1})}{2d}$$

given that

$$w(X \rightarrow_r Y) = max_r - \frac{max_r - min_r}{n_r(X)}$$

where $\rightarrow_r$ is a relation of type *r*, $\rightarrow_{r'}$ is its inverse, *d* is the depth of the deeper of the two nodes, max_r and min_r are the maximum and minimum weights possible for a relation of type *r* respectively, and $n_r(X)$ is the number of relations of type *r* leaving node *X*. □

Base weight of various relations :

A synonym relation gets a weight of 0,
Any hyponym or meronym variant or inverse gets a weight range from 1 to 2.
An antonym relationship is weighted invariantly 2.5 .

Definition 4 : Sussna's weighting of edges in the semantic network

3.2.2 Measures and results - Semantic distance

The semantic distance software was applied to a series of documents from the general *Time Magazine article collection*. As was also the case in the skimming project, Sussna filters the original answers found by the semantic distance software. In this process several observations were made. For instance, some instances were found to have no good interpretations in their respective contexts for various corpus data in order to arrive at the noun sequences of the respective documents. To measure the system disambiguation performance, the first five of the *Time* documents was manually disambiguated. This provided researchers with a *golden standard* to which to compare the reasons[2]. These instances was marked and ignored in the experiments. Furthermore, distinction was made between trivial and non-trivial success with regard to the disambiguation task:

Trivial success : An instance that cannot be interpreted incorrectly because all its senses are "good" with respect to the given context.
Non-trivial success : An instance that may realize both correct and incorrect concepts with respect to the context.

[2] Noun-like terms that were not really nouns in the context, (e.g.: *prime* in *prime minister*), or not the right noun (e.g.: *cent* in *per cent*) or proper nouns. Also some words were used in a sense that simply wasn't included in the MRD.

Def . : Overall distance minimization :

For a sequence of neighbouring words, T = $\{w_1, w_2, \ldots w_n\}$, let S be the set of all combinations of term senses, which has cardinality:

$$\prod_{i=1}^{n} |t_i|$$

where $|t_i|$ is the number of senses of $term_i$, and let $I \in S$ be a particular combination of senses $\{m_1, m_2, \ldots m_n\}$ where m_j is a sense of w_j.
The winning interpretation is the $I \in S$ which produces the minimal "energy" :

$$H_{min}(T) = \min_{S} \sum \text{distance}(x,y) \; \forall x,y \in S.$$

where
distance(x,y) = distance(y,x) =

$$\frac{\text{distance}(x \rightarrow y) + \text{distance}(y \rightarrow x)}{2}$$

and
distance(x,x) = 0

NB: This definition is adapted from (*Sussna 1993*) to correlate as closely as possible to the chosen formalism.

Definition 5 : Overall distance minimization - all shortest paths, unambiguously.

In the evaluation was focused on non-trivial successes only. Furthermore to different measures of success was used, firstly the hit/miss ratio with respect to the list of correct answers. Also a measure was devised to take into regard the actual difficulty of disambiguating each instance. Each disambiguated lexeme in a sequence is awarded a number of hit points and these are summed to get a hit score for the sequence which is then compared to the maximum hit score, as summarized in Definition 6.

Def . : Hit score for interpretations :

For each instance s in a sequence S, let p be its degree of polysemy, i.e.: the number of concepts it may realize (in any context), and let g be the number of good realizations (in the given context). The hit points awarded for a system hit, i.e.: a good system interpretation, equals p/g –1. Misses get zero points.
The actual hit points are summed and this sum divided by the sum of the maximum number of hit points possible, derived by treating all non-trivial instances as having been disambiguated correctly and their hit points awarded accordingly. Formally, the hit score over n instances equals:

$$\frac{\sum_{i=1}^{n} \text{hit points}_i \text{ where } s_i \text{ is a hit}}{\sum_{i=1}^{n} \text{hit points}_i}$$

Definition 6 : Hitpoints calculation for interpreted sequences relative to the manually derived golden standard.

This sophisticated evaluation scheme combined with two baselines for comparison, one resulting from a group of human disambiguators, the other chosen by assigning senses randomly to the instances in the test corpus, provided the researchers with a very robust framework for measuring system performance.

Several experiments were carried out, varying for instance the window size and the weighting scheme. The best results indicated a 50% system precision compared to chance precision, 39%, and the performance of human test subjects, 78%. The scoring scheme follows the tendency nicely with hit scores 0.447, 0.259 and 0.706 respectively.

Of the remaining conclusions from those experiments, the most important are :

- While the *mutual constraint* approach outperforms the *frozen past* method with regard to precision, the basically linear *frozen past* algorithm is, by far, the better with regard to computational complexity. The *frozen past* is still significantly more precise than chance.
- Furthermore, the experimental results suggest that *depth relative scaling*(DRS) and the combination of multiple different relations, both have noticeable impact on performance (in fact, disregarding either of these caused precision to drop 20-30%).
- Finally, neither *type specific fanout*(TSF) nor uniform vs. different weighting of different relations had significant consequence for the performance in these experiments.

All, in all *semantic distance* confirms the importance of semantic relationships between concepts. we see that :

a) All realizable lexemes in the sequence are considered for inclusion in its interpretation `I,` regardless of the length of the path connecting them.

b) Sequence length (window size) is arbitrary and thus does not contribute information in itself. Window moves along as analysis progresses, focussing on the word in the middle. The **frozen past** version decides the meaning of that word once and for all with regard to the interpretation of preceding words in the window. In **mutual constraint** the interpretation of that word is allowed to change, as the possible interpretations of words later in the sequence may provide for a better context.

c) Any resulting interpretation `I` of a sequence `S`, is unambiguous

d) All words in the sequence are interpreted, i.e.: all interpretations are complete

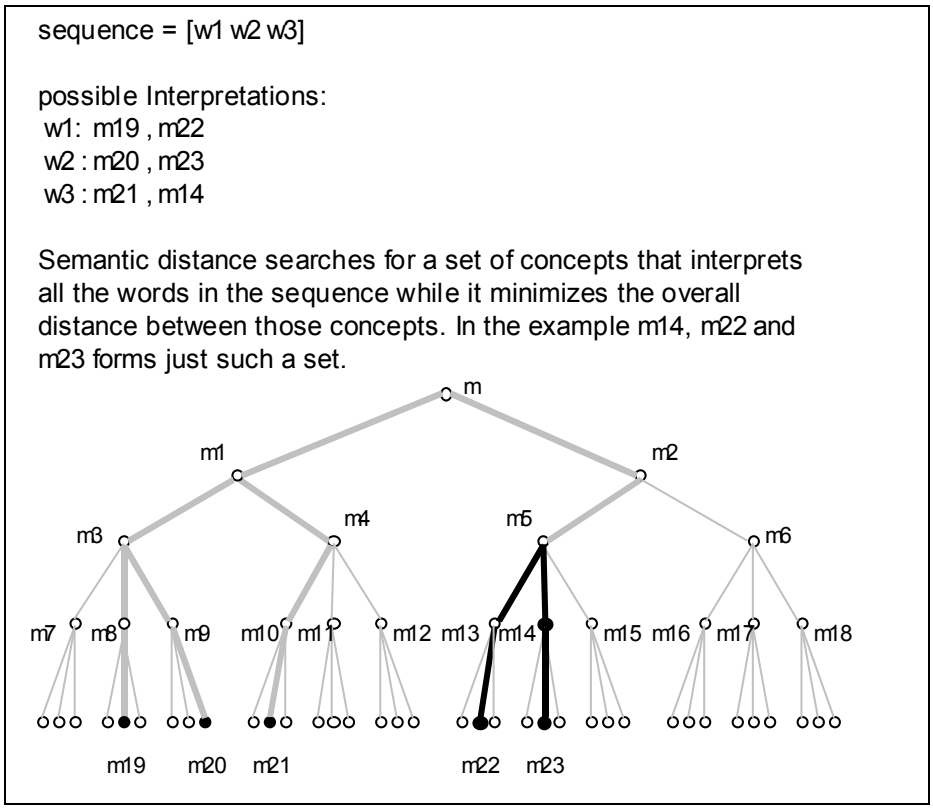


Figure 3: Semantic Distance does not consider any distance a hindrance for relationship, but deems the closest the most relevant. As a result, all words in the example sequence are interpreted unambiguously. In this simplified figure edges have equal weights.

Pursuing any line of relation to its full extent and comparing the weights of all possible paths between concepts is however extremely consuming and may not be necessary or even desirable, even though it succeeds in assigning one unambiguous meaning to all words in the sequence.

3.3 Conceptual Density - comparing subhierarchies

The third project that I will describe in detail is that of *Conceptual Density* by researchers Eneko Agirre and German Rigau, (*Agirre & Rigau 1996*). Their errand is that of word sense disambiguation. This project also regards nouns of general informative texts. Four texts was chosen randomly from the *Semantic Concordance* corpus or *SemCor*, for short. *SemCor* is a portion of the public domain general Brown Corpus, that was manually tagged with *WordNet* senses and made publicly available (*Miller et al. 1993*).

Like the two other projects described in this paper, *Conceptual Density* relies on the basic relational information of *WordNet* to detect semantically related interpretations of polysemous words. The focus, however, is abstracted away from the individual semantic relationships

themselves to regard instead the overall structure of the *WordNet* hierarchy.

3.3.1 Theory - Conceptual Density

The main hierarchy of *WordNet* consists of the directed acyclic graph (DAG) made up of nominal concepts in hypernymic relation to each other. While not every nominal concept in WordNet is necessarily involved in any relations of meronymy or antonymy, they are all part of the of hypernym-DAG. Since the hypernym-relation is the basic *is-a* relation, there is a certain amount of inheritance from ancestral concepts to their respective descendants, i.e.: if we know that a *B* is an *A* and that a *C* is a *B*, then we may assume that *C* inherits everything *B-ish*, including all of its *A-n*ess. Because new generations develop their own traits, inheritance go in one direction only, from ancestor to descendants. So nominal concepts are related to each-other via a common ancestor, they are part of the "family" represented by the subhierarchy dominated by their closest common ancestor in the hypernym DAG.

The main theory underlying *Conceptual Density* is that in the significant majority of cases, an informative text will involve only concepts from a particular subhierarchy of the WordNet concepts. If such a subhierarchy can be recognised for an arbitrary text, it is in essence proposed as the semantic context of the text. So in this theory, the most likely interpretation of the words in a given text is that which involves concepts in the smallest possible subhierarchy of the hypernym DAG. Here we see that relatedness of concepts is not determined through specific semantic relationships, but rather through membership of a particularly dense subhierarchy, represented by its dominating concept. The system :

> " - tries to resolve the lexical ambiguity of nouns by finding the combination of senses from a set of contiguous nouns that maximizes the *Conceptual Density* among senses. " (source: *Agirre & Rigau 1996*)

Def . : Conceptual Density measure :

Given a concept m, at the top of a subhierarchy, and *nhyp* (mean number of hyponyms per node), the Conceptual Density for m when its subhierarchy contains a number *n* (marks) of senses of the words to disambiguate is given by :

$$CD(m,n) = \frac{\sum_{i=0}^{n-1} nhyp^{i\ 0.20}}{descendants_m}$$

NB: The authors included the parameter 0.20 to try to smooth the exponential *i*, as m ranges between 1 and the total number of senses in *WordNet*. everal values were tried for the parameter and they found that the best performance was attained consistently when the parameter was near 0.20

Definition 7: Conceptual Density for a node, m, with regard to the relative concentration of possible interpretations in the subhierarchy it dominates.

3.3.2 Algorithm - Conceptual Density

The measure, *Conceptual Density*, itself is intended as an improvement over the semantic distance employed by for instance Sussna, described in the previous section of the present paper:

> " The measure of conceptual distance among concepts we are looking for should be sensitive to :
>
> - the length of the shortest path that connects the concepts involved.
> - the depth in the hierarchy: concepts in a deeper part of the hierarchy should be ranked closer.
> - the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a more sparse region.
> - the measure should be independent of the number of concepts we are measuring. " (source: *Agirre & Rigau 1996*)

The formal definition of the conceptual density for a given concept, in its capacity as a node in the hierarchy, is shown in Definition 7. The respective implication of this measure on the toy problem is sketched in Figure 4. Here we see that the formula seeks the subhierachy that maximizes the ratio between the number of potential interpretation to the number of concepts in general, while retaining at least one interpretation for each word in the sequence. As a consequence, the subhierarchy may involve more than one sense for some words in the sequence, in which case the algorithm is deemed unable to disambiguate the words in question

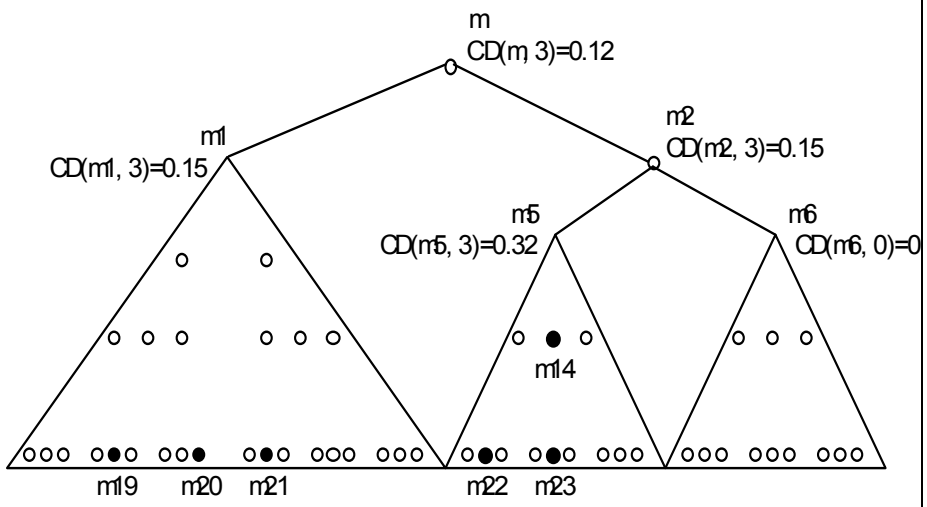
.

3.3.3 Measures and results - Conceptual Density

Agirre and Rigau applied their system to four texts, randomly chosen from the *SemCor*, (i.e.: portion of the Brown Corpus that was hand tagged with WordNet senses). They decided on three rather rigorous measures of success, summarized in Definition 8. The golden standard for the experiments is readily available via the manually assigned meanings of the **SemCor**. The researchers performed several experiments on this data among others :

- varying the window size – different texts performed very differently depending on their respective structure but best window size was found to be around 30 instances.
- incorporating meronym relations, no improvement was seen.
- global vs. local calculation of the nhyp factor – they were found to be equally good (no need for calculating the local nhyp over and over, one unique average factor for the entire hierarchy suffices.)

(source: Agirre&Rigau, 1996)



Considering the organisation of the possible interpretations (filled nodes) in the respective subhierarchies, the concptual density for each dominating node can be computed. We see that of all the subhierarchies offering an interpretation (at least one) for each, the conceptual density of the sub-hierarchy dominated by m5 is highest. Therefore the interpretations in that subhierarchy are deemed most likely.

Figure 4 Comparing the relative density of candidate concepts in various subhierarchies allows Conceptual Density may capture the proper interpretations effectively.

Def . : *Precision*, *Recall* and *Cover*:

Given
- a sequence S of instances of *polysemous* nouns,
- a sequence of *actual* (algorithmically found) interpretations of the nouns in S, and
- a golden standard of *correct* interpretations of the nouns in S,

Precision of the algorithm on that particular sequence is defined as the percentage of actual interpretations that are correct.

Recall is defined as the percentage of possible interpretations that are correct, i.e.: the number of possible interpretations equals the number of polysemous nouns in the sequence that have correct interpretations in the **MRD**.

Cover is defined as the percentage ratio of actual interpretations to possible interpretations

Definition 8: Three rigorous measures of success, precision recall and cover.

The overall results for the best window size is be summarized as shown in Table 1.

	Cover	Precision	Recall
all nouns	86.2	64.5	55.5
polysemous nouns only	79.6	53.9	42.8

Table 1: Overall results for Conceptual Density (source: Agirre&Rigau, 1996)[3]

Most interestingly for our purposes, however is a tentative comparison to the **frozen past** version of Sussna's S*emantic Distance* algorithm. Both algorithms was implemented and applied to the same data. It is expressed that a more thorough comparison between the algorithms is desirable but not possible within their framework. In this experiment, *conceptual Density* was modified to make random choces in cases where it would otherwise record a failure to disambiguate, mimicking the functionality of *Semantic Distance* in this respect. The results of the experiment are presented in Table 2 :

	Cover	Precision
Conceptual Density	100	60.1
Semantic Distance	100	52.3

Table 2: Comparison with Sussna (source: Agirre&Rigau, 1996)

All in all, considering subhierarchies instead of specific chains of relationships certainly seems to capture to very nature of contexts as conceptual neighbourhoods. We see that with the *Conceptual Density* algorithm :

a) Lexemes are considered entirely via their conceptual position within the hierarchy. The subhierarchy containing the largest concentration of possible interpretations decides wins. The actual relationships are not accessible in the method.

b) Sequence length (window size) is arbitrary and thus does not contribute information in itself. Window moves along as analysis progresses, focussing on the word in the middle

c) Any resulting interpretation `I` of a sequence `S`, is unambiguous

d) It does occur that a given subhierarchy offers more than one interpretation of a given noun. Here one of them can be chosen randomly or the noun is accepted as not disambiguated.

[3] The researchers did also experiment with using a set of "lexicographer's files" instead of the sense databases of WordNet. These files are part of the WordNet documentation on how the database distinctions were made. For the sake of clarity I have decided to omit these distinctions in this paper. Agirre&Rigau does however report finding evidence that these files can provide for a less fine grained interpretation with somewhat better cover-, precision- and recall-results.

3.4 Forces and drawbacks in comparison

The proper way to compare the three methods would of course be to establish a corpus of relevant data, subject it to of each of the three methods and compare the results. Such data could very well be portions of **SemCor** as was used in the evaluation of *Conceptual Density*. Since the framework for this paper did not allow for the replication of the experiments of Agirre and Rigau, this important analysis will have to be postponed to a later occasion.

It is possible, however, to compare the three methods on a more informal level. If the estimated results presented in section 3.1.3 is converted to the measures of **Cover**, **Precision** and **Recall**, we can put them up against the much more rigorous comparison between *Conceptual Density* and *Semantic Distance* presented in Table 2. Bearing in mind, that *Skimming* was subjected to different data than the two others and that the values for *Skimming* are the result of a very subjective evaluation procedure, the result can be regarded in Table 3.

	Cover	**Precision**
Conceptual Density	100	60.1
Semantic Distance	100	52.3
Skimming	~20	~70

Table 3: Tentative comparison of the three methods.

Even with the shortcomings of the *Skimming* evaluation in mind, these figures seem very reasonable. While *Conceptual Density* and *Semantic Distance* both offers interpretations for a much larger portion of the nouns in the sequence, they do so less precisely. That clearly confirms the intuition that considering only neighbours in the hierarchy results in "correct" interpretations with a high probability, but not nearly enough of them.

It is also clear that because *Semantic Distance* goes to the other extreme and considers paths of arbitrary length in the disambiguation process, it does of course find more interpretations but many of them are incorrect.

The focus on specific portions of the hierarchy, like the subhierarchies in *Conceptual Density,* does represent a very attractive middle road between the other two extremes. There are however consequences to viewing subhierarchies as abstractions over topical context.

One such consequence is that all concepts in a subhierarchy belongs to the respective context - even if only concepts in a top fraction of the subhierarchy are actually realized in the data. The result is a potentially very general and coarse grained characterization of contexts. For example a subhierarchy where VEHICLE is the dominating concept

also houses concepts like AEROPLANE, BICYCLE, CAR, SLEGIH, BOAT and all their respective subtypes as well, and also their constituting parts respective subhierarchies (if meronyms are included).

Also, counterintuitively, the association doesn't go in the opposite direction, so a context dominated by a concept like FORD (THE CAR BRAND) will not include VEHICLE

Furthermore it is not entirely clear how *Conceptual Density* performs in cases where several distinct contexts are in focus at the same time in the data. If for instance we imagine a text about car racing - will the resulting context be a complex one involving both distinct contexts represented here by SPORT and CAR, respectively, or by some closest common ancestor?

These consequences does not affect the fact that *Conceptual Density* clearly captures very important parts of the nature of topical contexts. It furthermore presents some very decent experimental results and a readily applicable algorithm. The rigorous measures of success and evaluation methods provides for sound comparison between methods.

The challenge remains to further delimit the borders of good contextual representation to that of confined local neighbourhoods, that increases both **cover** and **precision**.

4 Summing up

Skimming places itself solidly among the knowledge based methods to **Word Sense Disambiguation** and **Context Recognition**.

The same fundamental ideas inspired all the projects in the comparison. *Semantic Distance* and *Conceptual Density* both concerned themselves with **WSD** and therefore wanted to maximize both **Cover** and **Precision**. *Skimming* is intended primarily as **CR** and was therefore less concerned with **Cover**. The main difference between **WSD** and **CR** in this respect, is that not all words in a sequence needs to be interpreted in order to achieve a good result in **CR,** as long as those that are interpreted are interpreted correctly and sufficiently representative of the context(s) involved. All the same, it is clear that the more words, important to the context, that are correctly interpreted the more detailed and useful the representation.

Consequently, the tentative comparison suggests that while *Skimming* lacks behind woefully with regard to **Cover,** it presents itself very comparably with regard to the quality of the interpretations it does find. Though this needs to be formally verified through a proper comparison, it is an indication that maybe a refinement of *Skimming* can increase **Cover** while retaining a high **Precision** – perhaps even better than *Conceptual Density*.

4.1 Perspectives

First and foremost, it would be very interesting to see an objective rigorous comparison of the respective performances of the methods on large amounts of data. A first step in this direction would be to repeat subject *Skimming* to the same portions of the **SEMCOR**, that is freely available as already suggested elsewhere in this paper. Applying the methods to other pieces of data would of course imply implementing the respective algorithms, acquiring the source code from the developer's or conduct the respective experiments in collaboration.

Secondly, a particular direction for stepwise refinement of the experimental S*kimming* algorithm is presenting itself. Allowing the flexibility of chains of relationship of varying length - one relationship at a time - between realized concepts, and measuring the performance along the way, is likely to suggest a crossover **Cover** can no longer be improved without sacrificing **Precision –** and at the same time cater for the likely possibility that different problems may require different degrees of cover and or precision, trading off computational complexity. A properly refined S*kimming* algorithm might allow contexts to *crystallize* as one or more **generic, coherent components** of the hierarchy, restricted on all sides, rather than on the just 2/3. The CD formula may even be adapted to express the density of interpretations in such components.

The different skimming configurations could be incorporated in the paragraph recognising algorithm documented in (*Lassen 2006*). Applying the paragraph recognition algorithm to SemCor documents and skimming accordingly should result in immediate feedback about the quality of the context found by skimming.

Finally, this project has shown how important reliable sense tagged corpora are to researchers, providing for golden standards for algorithm performance. To me, much of the typographic information like headlines and paragraph and chapter structure of written data is also very important. Because of the extreme tediousness of tagging large corpora by hand, providing automated assistance for as much of that important work as possible should have high priority.

4.2 Other approaches

During the research for this paper I came across a good many projects of word sense disambiguation, context recognition, concepts mapping for artificial intelligence and cognitive science models and other interesting projects. While present circumstances does not allow for a full thorough survey, those that were considered, and ultimately rejected, for inclusion in the comparison have been included among the references.

Most importantly:

(*Resnik 1995*), who describes a method that disambiguates nouns that are assumed to be already sorted with regard to semantically similar behaviour, like concordance or collocation. Even though Resnik does not disambiguate nouns from running textual data, the work he does present many interesting and insights.

(*Richardson et al. 1995*) describes how WordNet may be transformed into a veritable Knowledge Base for tracking conceptual similarity.

(*Voorhees 1993*), whose errand is IR and describes an automatic indexing procedure that uses the "IS-A" relations contained within WordNet and the set of nouns in a text to select a sense for each polysemous noun in the text.

(*Yarowsky, 1992*) uses Roget's Thesaurus as MRD. He exploits the context/domain indicators of this well know thesaurus in order to make guesses at the proper interpretation of polysemous words.

Finally (*Ide & Véronis 1998*) is worth mentioning for it s very good historical survey of **WSD** approaches in general. The paper presents the state of the art anno 1998 with regard to many important issues in **WSD** and **CR**. They distinguished the following general methods:

AI-based: these methods in general attempts to model a theory of human language understanding. These systems relied on detailed information about syntax and semantic, being developed in the context of much larger system intended to combat full language understanding.

Knowledge-based methods, among which all the projects in this paper should be counted, seek to bypass overcome the so called "knowledge acquisition bottleneck" by extracting semantic and pragmatic knowledge from pre-existing sources - MRD's like WordNet, thessauri like Roget's and so on.

Corpus-based: These methods attacks word sense disambiguation via collocation and concordance analysis, where the frequency of a word occurring near other specific words is basically compared to empirically established statistics and interpreted accordingly.

While all of this concerns **word sense disambiguation** and thus require the correct interpretation of all the ambiguous words in the data, my view remains that for the vast majority of cases the much smaller problem of **context recognition** suffices in distinguishing the general contents of the data and, subsequently, forming a sound basis for the minute analysis of the likely very few cases where complete understanding is in demand.

References

Agirre,E. & Lopes de Lacalle,O. 2003
"Clustering WordNet Word Senses".
RANLP 03.

Agirre,E. & Rigau,G. 1996
"Word Sense Disambiguation using Conceptual Density".
COLING-96.

Cañas,A.J., Valerio,A., Lalinde-Pulido,J. Carvalho,M. & Arquedas,M. 2003
"Using WordNet for Word Sense Disambiguation to support Concept Map Construction".
SPIRE 2003, Springer.

Cowie,J., Cuthrie,J. & Cuthrie,L. 1992
"Lexical Disambiguation using Simulated Annealing".
COLING-92.

Cruse, D. A. 1986
"Lexical Semantics"
Cambridge University Press.

Grice, H. P. 1975
"Logic and Conversation"
In Cole Morgan 1975 pp.41-58.

Ide,N. & Véronis,J. 1998
"Introduction to the Special issue on Word Sense Disambiguation: The State of the Art".
Computational Linguistics 24.

Kim, Y.W., and Kim J.H. 1990
"A model of knowledge based information retrieval with hierarchical concept graph"
Journal of Documentation 46(2).

Lassen,O.T. 2005
"Skimming for Context"
Master's Thesis, Dept. of Computer Science, University of Copenhagen, Denmark.

Lassen,O.T. 2006
"Skimming for Paragraphs".
Post Graduate Paper, Dept. of Computer Science, University of Copenhagen, Denmark.

Miller G. 1990
"Five papers on WordNet".
Special Issue of International Journal of Lexicography 3(4).

Miller,G.L., Randee,T. & Bunker,R. 1993
"A Semantic Corcordance".
3rd DARPA Workshop on Human Language Technology.

Pustejowsky,J. 1995
"The Generative Lexicon".
MIT Press

Rada,R., Mili,H., Bicknell,E. & Blettner,M. 1989
"Development and application of a metric on semantic nets".
IEEE Transactions on systems, man and cybernetics 19(1).

Resnik,P. 1995
"Disambiguating Noun Groupings with Respect to WordNet Senses".
3rd Workshop on Very Large Corpora, MIT.
Richardson,R., Smeaton,A.F. & Murphy, J. 1994
"Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words".
In Working Paper CA-1294, School of Computer Applications, Dublin City University. Dublin, Ireland.
Romacker,M. & Hahn, U. 2000
"Coping with different Types of Ambiguity Using a Uniform Context Handling Mechanism".
NLDB 2000,Springer.
Rosso,P., Masulli,F., Buscaldi,D., Pla,F. & Molina,A. 2003
"Automatic Noun Sense Disambiguation".
CICLing 2003, Springer.
Sussna,M. 1993
"Word Sense Disambiguation for Free-text Indexing Using a massive Semantic Network".
Proceedings of 2nd International Conference on Information and Knowledge Management, ACM.
Voorhees,E. 1993
"Using WordNet to Disambiguate Word Senses for Text Retrieval".
16th annual international ACM SIGIR conference. ACM.
Widdows,D. & Dorow,B. 2002
"A graph model for unsupervised lexical acquisition".
19th International Conference on Computational Lingustics.
Widdows,D. 2003
"A Mathematical Model for Context and Word-Meaning".
Context, Springer.
Wilks,Y., Fass,G., Guo,C., McDonal,J., Plate,T. & Slator,B. 1990
"Providing Machine Tractable Dictionary Tools".
Machine Translations 5, Kluwer.
Yarowsky, D. 1992
"Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora".
COLING-92.