General Word Sense Disambiguation Method
Based on a Full Sentential Context

Jiri Stetina•õ and Makoto Nagao•õ

This paper presents a new general supervised word sense disambiguation method based on a relatively small syntactically parsed and semantically tagged training corpus.The method exploits a full sentential context and all the explicit semantic relations in a sentence to identify the senses of all of that sentence's content words. It solves the sparse data problem of a small training corpus by substituting the words by their semantic classes.In spite of a very small training corpus,we report an over-all accuracy of 80.3%(85.7,63.9,83.6 and 86.5%,for nouns,verbs,adjectives and adverbs,respectively),which exceeds the accuracy of a statistical sense-frequency based semantic tagging,the only really applicable general disambiguating technique. Because the method uses the sentential syntactic structure it is particularly suitable
for integration with a probabilistic syntactic analyser.

KeyWords:word sense disambiguation,conceptual hierarchy,semantic distance, semantic relations,relational probability,sentential context,discourse
context

1 Introduction

Identification of the right sense of a word in a sentence is crucial to any successful Natural Language Processing system.The same word can have different meanings in different con-texts.The task of Word Sense Disambiguation is to determine the correct sense of a word in a given context.Consider,for example,the following sentences:
(a)The central bank raised the interest rate.
(b)John has no interest in biology.
The word interest has a different meaning in each of these sentences.In sentence(a)it refers to a financial percentage,while in(b)to John's state of mind.As in other examples,the intended meaning depends on the words which co-occur with the ambiguous word,on their semantic relations and on our knowledge about the co-occurring concepts.In most cases the correct word sense can be identified using only the words co-occurring in the same sentence. However,very often we also need to use the context of words that appear outside the given sentence.For this reason we distinguish two types of contexts:the sentential context and the discourse context.The sentential context is given by the words which co-occur with the word

†Kyoto University,Department of Electronics and Communications

47

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

in a sentence and by their relations to this word,while the discourse context is given by the words outside the sentence and their relations to the word.The problem that arises here is that most of the co-occurring words are also polysemous,and unless disambiguated they can-not fully contribute to the process of disambiguation.The senses of these words,however,also depend on the sense of the disambiguated word and therefore there is a reciprocal dependency which we will try to resolve by the algorithm described in this paper.

2 The Task Specification

2.1 Sense definitions

Dictionaries differ widely in the number of sense distinctions they draw;pocket dictionar-ies offer few,and unabridged dictionaries offer many alternative senses.For our work,we used the word sense definitions as given in WordNet(Miller 1990),which is comparable to a good printed dictionary in its coverage and distinction of senses.It provides about the same semantic granularity as a good desk dictionary(Miller 1994).Since WordNet only provides definitions for content words(nouns,verbs,adjectives and adverbs),we are only concerned with identifying the correct senses of the content words.

2.2 Training and Testing Sentences

Both for the training and for the testing of our algorithm,we used the syntactically anal-ysed sentences of the Brown Corpus(Marcus 1993),which have been manually semantically tagged(Miller,Leacock,and Tengi 1993)into semantic concordance files(SemCor)1.These

files combine 103 passages of the Brown Corpus with the WordNet lexical database in such a way that every content word in the text carries both a syntactic tag and a semantic tag pointing to the appropriate sense of that word in WordNet.Passages in the Brown Corpus are approximately 2,000 words long,and each contains approximately 1,000 content words.The
fact that the sentences are also syntactically analysed,enables a straightforward extraction of relations among the words.We randomly selected 15 files for testing,always removing each tested file from the corpus,using the remaining 102 files for training,thus avoiding the training on the testing data.

1 Available via anonymous ftp from clarity.princeton.edu.
48

Stetina,J.et al. General Word Sense Disambiguation Method

Table 1 Percentage of nouns,verbs,adjectives and adverbs and average number of senses

Table 2 Percentage of polysemous word in the corpus

2.3 Sense distribution

The portion of Brown Corpus we used contained 23,934 different content words,adding
up to a total of 106,755 content words.The percentages of the nouns,verbs,adjectives and adverbs in the semantically tagged corpus,together with their average number of WordNet senses,are given in Table 1.
Although most of the words in a dictionary are monosemous,it is the polysemous words that occur most frequently in speech and text.For example,over 80% of words in WordNet are monosemous,but almost 78% of the content words in the tested corpus had more than one sense2,as shown in Table 2.
It is generally recognised that systems for automatic word sense disambiguation should be evaluated against the null hypothesis.(Gale,Church,and Yarowsky 1992)suggest that the appropriate basis for comparison would be a system that assumes that each word is being used in its most frequently occurring sense.They estimate that the most frequent sense would be correct 75% of the time,which they propose as a lower boundary for automatic word sense disambiguation performance.The WordNet lexical database contains the word senses ordered

2 We counted words as polysemous,if their lemmas had more than one sense in WordNet.Lemma,defined in SemCor,is either the word's baseform(if found in WordNet)or its hand-made redefinition.For example,the
proper noun Fulton,a county name not found in WordNet,is redefined in SemCor into location.Because lemma location has four senses in WordNet,its ambiguity was counted as four.We extracted all such redefinitions
from SemCor,whose coverage in WordNet is almost perfect.

49

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

Table 3 Sense frequenies

according to their frequency.However,it does not provide information on how frequently each sense occurs.Using the semantic concordance files,we determined the frequencies of each sense number occurring in the used corpus.
As Table 3 suggests,assigning the most frequent sense(as defined by WordNet)to every content word in the used corpus would result in an accuracy of 75.2 %.Our aim is to create a word sense disambiguation system for identifying the correct senses of all content words in a given sentence,with an accuracy higher than would be achieved solely by a use of the most frequent sense.

3 General Word Sense Disambiguation

The aim of the system described here is to take any syntactically analysed sentence on the input and assign each of its content words a pointer to an appropriate sense in WordNet. Because the words in a sentence are bound by their syntactic relations,all the word's senses are determined by their most probable combination in all the syntactic relations derived from the parse structure of the given sentence.It is assumed here that each phrase has one central constituent(head),and all other constituents in the phrase modify the head(modifiers).It is also assumed that there is no relation between the modifiers.The relations are explicitly present in the parse tree,where head words propagate up through the tree,each parent re-ceiving its head word from its head-child.Every syntactic relation can be also viewed as a semantic relationship between the concepts represented by the participating words.Consider, for example,the sentence(1)whose syntactic structure is given in figure 1.

(1)The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced no evidence that any irregularities took place.

50

Stetina,J.et al. General Word Sense Disambiguation Method

Fig.1 Example parse tree

Each word in the above sentence is bound by a number of syntactic relations which deter-mine the correct sense of the word.For example,the sense of the verb produced is constrained by the subject-verb relation with the noun investigation,by the verb-object relation with the noun evidence and by the subordinate clause relation with the verb said.Similarly,the verb said is constrained by its relations with the words Jury,Friday and produced;the sense of the noun investigation is constrained by the relation with the head of its prepositional phrase-election,and by the subject-verb relation with the verb produced,and so on.
The key to extraction of the relations is that any phrase can be substituted by the corre-sponding tree head-word(links marked bold in Figure 1).To determine the tree head-word we used a set of rules similar to that described by(Magerman 1995)(Jelinek,Lafferty,Magerman, Mercer,Rathnaparkhi,and Roukos 1994)and also used by(Collins 1996)3,which we modified in the following way:

・ The head of a prepositional phrase(PP->INNP)was substituted by a function the

name of which corresponds to the preposition,and its sole argument corresponds to the
head of the noun phrase NP.

・ The head of a subordinate clause was changed to a function named after the head of the first element in the subordinate clause(usually 'that' or a 'NULL' elelnent)and its sole
3 We would like to thank Mr Collins for providing these rules.

51

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

argument corresponds to the head of its second element(usually head of a sentence). Because we assumed that the relations within the same phrase are independent,all the relations are between the modifier constituents and the head of a phrase only.This is not necessarily true in some situations,but for the sake of simplicity we took the liberty to assume so.A complete list of applicable relations for sentence(1)is given in(2)4.

(2) NP(NNP(County),NNP(Jury)) VP(VBD(took),NP(place))
NP(NNP(Grand),NNP(Jury)) NP(NN(evidence),SBAR(that(took))
NP(NP(Atlanta),NP(election)) S(NP(investigation),VP(produced))
NP(JJ(recent),NP(election)) VP(VBD(produced),NP(evidence))
NP(JJ(primary),NN(election)) VP(VBD(said),NP(Friday))
NP(NN(investigation),PP(of(election))) VP(VBD(said),SBAR(O(produced)))
S(NP(irregularities),VP(took)) S(NP(Jury),VP(said))

Unfortunately,the syntactic structures of the Brown Corpus sentences do not distinguish
the internal structure of the noun phrases and,therefore,two noun phrases of the example sentence(1)are rather simplified in the actual corpus,

leading to a slightly different set of relations(3):

4 We deliberately ignore relations between non-content constituents,like determiners,punctuation,etc.

52

Stetina,J.et al. General Word Sense Disambiguation Method

(3)NP(NNP(Fulton),NNP(Jury))
NP(NNP(County),NNP(Jury))VP(VBD(took),NP(place)) NP(NNP(Grand),NNP(Jury))NP(NN(evidence),SBAR(that(took)) NP(NP(Atlanta),NN(election))S(NP(investigation),VP(produced)) NP(JJ(recent),NP(election))VP(VBD(produced),NP(evidence)) NP(JJ(primary),NN(election))VP(VBD(said),NP(Friday)) NP(NN(investigation),PP(of(election)))VP(VBD(said),SBAR(O(produced))) S(NP(irregularities),VP(took))S(NP(Jury),VP(said))

Each of the extracted syntactic relations has a certain probability for each combination of the senses of its arguments.This probability is derived from the probability of the semantic relation of each combination of the sense candidates of the related content words.Therefore,
the approach described here consists of two phases:1.learning the semantic relations,and 2.disambiguation through the probability evaluation of relations.Section 4 describes the learning phase,Section 5 the process of disambiguation.

4 Learning

At first,every content word in every sentence in the training set was tagged by an ap-propriate pointer to a sense in WordNet.These pointers were extracted from SemCor.If several words in a phrase formed a collocation,only the phrase-head was assigned a tag.For example,because the adjective primary in primary election is not tagged separately in the semantic concordance file,this adjective remained untagged,while the noun election received a tag corresponding to the WordNet sense of the collocation primary_election.
Secondly,using the parse trees of all the corpus sentences,all the syntactic relations present
in the training corpus were extracted and converted into the following form:

(4)rel(PNT,MNT,HNT,MS,HS,RP).
where PNT is the phrase parent non-terminal,MNT the modifier non-terminal,HNT the head non-terminal,MS the semantic content(see below)of the modifier constituent,HS the semantic content of the head constituent and RP the relative position of the modifier and the head(RP=1 indicates that the modifier precedes the head,while for RP=2 the head precedes the modifier).Relations involving non-content modifiers were ignored.Synsets of the words not present in WordNet were substituted by the words themselves.
The semantic content was either a WordNet sense identificator(synset)or,in the case of prepositional and subordinate phrases,a function of the preposition(or a null element)

53

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

and the sense identificator of the second phrase constituent.(Stetina and Nagao 1997) has shown the importance of word sense disambiguation for the resolution of a prepositional phrase attachment ambiguity.The preposition is a particularly important element in deciding the attachment of prepositional phrases.Therefore.assuming prepositions are also very important for specifying the relation between the phrase head and the head of the noun phrase which follows the preposition.we have implemented this functional exception.For example.in the case of a prepositional phrase taken from the Brown Corpus structure

[NP,[DT'an],[NN,investigation], [PP,[IN,of],
[NP,[NP,['NNP',Atlanta]],
[POS,-s],
[JJ,recent],
[JJ,Primary],
[NN,election]]]],

the extracted relations were rel(NP.NN.PP.103935809.of(100103176).1) for NP->NN PP, and rel(NP.NP.NN.105608324.100103176.1) and rel(NP.JJ.NN.300610062.100103176,
1) for NP— >NP POS JJ JJ NN.where 103935809 is the WordNet synset of the appropriate sense of the noun investigation.100103176 the synset of primary_election.and 105608324 the synset for Atlanta.We have further collapsed all the non-terminals that belong to the same syntactic category into the same group.i.e.the Brown Corpus part-of-speech tags NN.NNS, NNP.NNPS were collapsed into NOUN; VB.VBD.VBG.VBN.VBP.VBZ into VERB; JJ, JJR and JJS into ADJ and RB.RBR and RBS into ADV.

Also.in order to avoid the incorrect extraction of relations with verbs in the passive voice, we have converted all passive verb phrases into an appropriate active voice.e.g.the incorrect subject-verb relation in

…the election was conducted

was replaced by a corresponding verb-object relation

…conducted the election5

During the learning phase we extracted 76,426 relations.32,949 with RP=1 and 43,477 with RP=2.The following table shows a few of the most common types of extracted relations:
5 We did not attempt to resolve the anaphoric link concerning the possible subject of the verb.leaving the subject candidate open.

54

Stetina,J.et al. General Word Sense Disambiguation Method

Table 4 Few of the most common semantic relations

5 Disambiguation Algorithm

As mentioned above,we assumed that all the content words in a sentence are bound by
a number of syntactic relations.Every content word can have several meanings,but each of these meanings has a different probability,which is given by the set of semantic relations in which the word participates.Because every relation has two arguments (head and its modi-fier),the probability of each sense also depends on the probability of the sense of the other

participant in the relation.The task is to select such a combination of senses for all the content words,that the overall relational probability is maximal.If,for any given sentence,we had extracted N syntactic relations Ri,the overall relational probability for the combination of senses X would be:

(5)

where p(RilX) is the probability of the i-th relation given the combination of senses X.If we consider,that an average word sense ambiguity in the used corpus is 5.8 senses,a sentence with 10 content words would have 5.810 possible sense combinations,leading to a combinatorial ex-plosion of over 43,080,420 overall probability combinations,which is not feasible.Also,with a very small training corpus,it is not possible to estimate the sense probabilities very accurately. Therefore,we have opted for a hierarchical disambiguation approach based on similarity mea-sures between the tested and the training relations,which we will describe in Section 5.2.At

first,however,we will describe the part of the probabilistic model which assigns probability

estimates to the individual sense combinations based on the semantic relations acquired in the learning phase.

55

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

5.1 Relational Probability Estimate

Consider,for example,the syntactic relation between a head noun and its adjectival mod-ifier derived from NP->JJ NN.Let us assume that the number of senses in WordNet is k

for the adjective and 1 for the noun.The number of possible sense combinations is therefore m=k*1.The probability estimate of a sense combination(i,j)in the relation R,where i is the sense of the modifier(adjective in this example)and j is the sense of the head(noun in this example),is calculated as follows:

(6)

where fR(i,j)is a score of co-occurrences of a modifier sense x with a head word sense y, among the same semantic relations R extracted during the learning phase.Please note,that because fR(i,j)is not a count but rather a score of co-occurrences(defined below),pR(i,j)is not a real probability but rather its approximation.Because the occurrence count is replaced by a similarity score,the sparse data problem of a small training corpus is substantially re-duced.The score of co-occurrences is defined as a sum of hits of similar pairs,where a hit is a multiplication of the similarity measures,sim(i,x)and sim(j,y),between both participants,

i.e.:

(7)

where x,y E R;r is the number of relations of the same type(for the above example R=rel(NP,ADJ,NOUN,x,y,1))found in the training corpus.To emphasise the sense-restricting contribution of each example found,every pair(x,y)is restricted to contributing to only one sense combination(i,j):every example pair(x,y)contributes only to such a combination for which sim(i,x)* sim(j,y)is maximal.
fR(i,j)represents a sum of all hits in the training corpus for the sense combination(i,j)6. Because the similarity measure(see below)has a value between 0 and 1 and each hit is a multiplication of two similarities,its value is also between 0 and 1.The reason why we used a multiplication of similarities was to eliminate the contributions of examples in which one participant belonged to a completely different semantic class.For example,the training pair new airport,makes no contribution to the probability estimate of any sense combination of new management,because none of the two senses of the noun management(group or human

6 In the actual implementation,we only used the twenty best hits for each relation.This was to compromise between the number of training examples and their quality(in terms of semantic distance).

56

Stetina,J.et al. General Word Sense Disambiguation Method

activity)belongs to the same semantic class as airport(entity).On the other hand,new air-port would contribute to the probability estimate of the sense combination of modern building because one sense of the adjective modern is synonymous to one sense of the adjective new, and one sense of the noun building belongs to the same conceptual class(entity)as the noun airport.The situation is analogous for all other relations.The reason why we used a count modified by the semantic distances,rather than a count of exact matches only,was to avoid situations where no match would be found due to the sparse data,a problem of many small training corpora.
Every semantic relation can be represented by a relational matrix,which is a matrix whose first coordinate represents the sense of the modifier,the second coordinate represents the sense of the head and the value at the coordinate position(i,j)is the estimate of the prob-ability of the sense combination(i,j)computed by(6).An example of a relational matrix for an adjective-noun relation modern building based on two training examples(new airport and classical music)is given in Figure 3.Naturally,the more the examples,the more fields of the matrix get filled.The training examples have an accumulative effect on the matrix,because the sense probabilities in the matrix are calculated as a sum of 'similarity based frequency scores' of all examples(7)divided by the sum of all matrix entries,(6).The most likely sense combination scores the highest value in the matrix.Each semantic relation has its own matrix. The way all the relations are combined is described in Section 5.2.

Semantic Similarity

We base the definition of the semantic similarity between two concepts7 on their semantic distance,as follows:

(8)sim(a,b)=1- sd(a,b)2,

The semantic distance sd(a,b)is squared in the above formula in order to give a bigger weight to closer matches.Because sd(a,b)E<0,1>,also the sim(a,b)E<0,1>.The similarity of identical concepts is 1 and the similarity of two concepts which do not have a common ancestor in the semantic hierarchy is 0.

Semantic Distance

A.Semantic Distance for Nouns and Verbs
As in(Stetina and Nagao 1997)(Sussna 1993)we define the semantic distance between two
7 Concepts are defined by their WordNet synsets a,b.
57

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

Fig.2 Semantic distance example

senses of nouns and verbs in the WordNet hierarchy as the average length of the path of each sense to its nearest common ancestor divided by its depth in the hierarchy:

If a and b have no common ancestor,sd(a,b)=1.See Figure 2 for an example.
If any of the participants in the semantic distance calculation is a function(derived from
a prepositional phrase or subordinate clause),the distance is equal to the distance of the function arguments for the same functor,or equals 1 for different functors.For example, sd(of(sensel),of(sense2))=sd(sensel,sense2),while sd(of(sensel),about(sense2))=1,no matter what sensel and sense2 are.
Because only nouns and verbs form a hierarchy in WordNet,whereas adjectives and adverbs do not,we have to treat the latter differently.
B.Semantic Distance for Adjectives
The semantic contribution of adjectives is secondary to,and dependent on,the head nouns they modify.(Sapir 1944)seems to have been the first linguist to point out explicitly that many adjectives take on different meanings when they modify different nouns.WordNet di-vides adjectives into two major classes:descriptive and relational.Descriptive adjectives ascribe to their head nouns values of typically bipolar attributes and consequently are organ-

58

Stetina,J.et al. General Word Sense Disambiguation Method

Example to disambiguate:MODERN(X)BUILDING(Y):rel(NP,ADJ,NOUN,X,Y,1)
Example training set:NEW(9)AIRPORT:rel(NP,ADJ,NOUN,3006112602,102055456,1) CLASSICAL(1)MUSIC(3):rel(NP,ADJ,NOUN,300306289,100313161,1)

Fig.3 Relational matrix based on two training examples

ised in terms of antonymy and similarity of meaning(synonymy).Descriptive adjectives that do not have direct antonyms are said to have indirect antonyms by virtue of their semantic similarity to adjectives that do have direct antonyms.Relational adjectives are assumed to be stylistic variants of modifying nouns and so are cross-referenced to the noun files(Miller 1990).
Adjectives and their antonymy,the salient feature of descriptive adjectives,exhibit a mu-tuality of association(Deese 1964)(good associates with bad,long with short,etc.).This mutuality seems to be acquired as a consequence of co-occurrence of adjective-antonymy pairs in the same sentences(Justeton and Katz 1991).Since adjectives and their antonyms usually modify the same noun,we can use this information in calculating the probabilities of semantic relations which involve adjectives.Because WordNet contains the pointers between adjectives, their synonymy,antonymy and similar adjectives,we define semantic distance for descriptive

59

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

adjectives in the following way:
sd(a,b)=0 for the same adjectival synsets(incl.synonymy)
sd(a,b)=0 for the synsets in antonymy relations,i.e.for ant(a,b)
sd(a,b)=0.5 for the synsets in the same similarity cluster
sd(a,b)=0.5 if a belongs to the same similarity cluster as c and
b is the antonymy of c(indirect antonymy)
sd(a,b)=1 for all other synsets
In the case of relational adjectives,which pertain to a noun whose modifying function they take(atomic bomb=atom bomb),we calculate the semantic similarity as a similarity of the nouns they pertain to.If,for example,we are to calculate the semantic distance be-tween the adjectives atomic and chemical,as in atomic bomb and chemical weapons,we first access the pertinent information in WordNet and calculate the distance as a distance between corresponding senses of the noun atom and the noun chemical.
C.Semantic Distance for Adverbs

The only two relations in WordNet that involve adverbs are synonymy and antonymy.We therefore define semantic distance between two adverbial synsets as:
sd(a,b)=0 for the same synsets(incl.synonymy).
sd(a,b)=0 for the synsets in antonymy relation ant(a,b)
sd(a,b)=1 for all other synsets

5.2 Hierarchical Disambiguation

This section describes the main part of the algorithm,i.e.the disambiguation process based on the overall probability estimate of sentential relations.As we have outlined above, for computational reasons,it is not feasible to evaluate overall probabilities for all the sense combinations.Instead,we take advantage of the hierarchical structure of each sentence and arrive at the optimum combination of its word senses,in a process which has two parts:1. bottom-up propagation of the head word sense scores and 2.top-down disambigua-tion.The head words propagate from the leaves of the parse tree to their topmost parents and at each level of the tree participate in semantic relations with their modifiers.The relation with each modifier thus changes the likeliness of each sense of the head word.Similarly,the likeliness of each sense of every modifier is changed by the semantic relation with the head (according to our assumption that there are no relations between modifiers within a phrase, each modifier participates only in the relation with its head).Because the likeliness of each sense of the head is to be further modified at higher levels of the tree,we store the likeliness of

60

Stetina,J.et al. General Word Sense Disambiguation Method

each sense of the modifier in a matrix,where each line represents a vector of the sense scores of the modifier corresponding to one sense of the head.This matrix will be used to determine the sense of a modifier once its governing head word sense has been identified.As we will describe below,the matrix is calculated as a product of the relational matrix of the modifier and its head,and of the propagated sense scores of the modifier.The heads propagate up the tree until the root which represents the main head of the sentence is reached.At the root of the tree we disambiguate the main head of the sentence by choosing its sense which corre-sponds to the highest value among its sense scores,i.e.we choose the sense with the highest likeliness.This is plausible,because by the time the tree root is reached,all the semantic relations of the sentence have been taken into account,and have participated in the calcula-tion of the sense scores of the main head.Under the assumption that there are no relations between modifiers,the sense scores of the main head word thus approximate the probabilities of the head word senses and serve as an intermediate to approximate the highest overall sense combination probability of the sentence's words as given by(5).Once the main head has been disambiguated,its sense travels down the hierarchy and facilitates the disambiguation of other words,based on their previously stored sense score matrices.The disambiguation terminates after all the leaves have been reached and all the words of the sentence disambiguated.The process is described in detail below.

Bottom-up head word sense score propagation

In compliance with our assumption that all the semantic relations are only between a head word and its modifiers at any syntactic level,the modifiers do not participate in any relation with an element outside their parent phrase.As depicted in the example in figure 1,it is only the head word concepts that propagate through the parse tree and that participate in semantic relations with concepts on other levels of the parse tree.The modifiers(which are heads themselves at lower tree levels),however,play an important role in constraining the head-word senses.The number of relations derived at each level of the tree depends on the number of concepts that modify the head.Each of these relations contributes to the score of each sense of the head word.We define the sense score vector of a word w as a vector of scores of each WordNet sense of the word w.The initial sense score vector of the word

w is given by its contextually independent sense distribution in the whole training corpus. Because the training corpus is relatively small,and because it always excludes the tested file, an appropriate sense of the word w may not be present in it at all.Therefore,each sense i of the word w is always given a non-zero initial score pi(w)(9a):

61

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

(9a)

where count(w)2 is the number of occurrences of the sense i of the word w in the entire training corpus,and n is the number of different WordNet senses of the word w.
The sense score vectors of head words propagate up the tree.At each level,they are mod-ified by all the semantic relations with their modifiers which occur at that level.Also,the sense score vectors of head words are used to calculate the matrices of the sense score vectors of the modifiers.This is done as follows:
Let H=[hl,h2,...,hk] be the sense score vector of the head word h.Let T=[Ri,R2,...Rn] be a set of relations between the head word h and its modifiers.
1. For each semantic relation Ri E T between the head word h and a modifier mi with sense score vector Mi=[oil,oi2,...oil],do:
1.1 Using(6),calculate the relational matrix Ri(m,h)of the relation Ri
1.2 For each of E Mi multiply all the elements of the Ri(m,h)for which
m=oi by oi,yielding Qi-the sense score matrix of the modifier mi
2. The new sense score vector of the head word h is now G=[gl,g2,...,gkl,where
(10)
Lj/L represents the score of the head word sense j based on the matrices Q calcu-
lated in the step 1.,i.e.:
(11)
where xi(j,u)e Qi and max(xi(j,u))is the highest score in the line of the matrix
Qi which corresponds to the head word sense j.n is the number of modifiers of the
head word h at the current tree level,and

where k is the number of senses of the head word h.

The reason why gj(10)is calculated as a sum of the best scores(11),rather than by using the traditional maximum likelihood estimate(Berger,Pietra,and Pietra 1996)(Gale, Church,and Yarowsky 1993),is to minimise the effect of the sparse data problem.Imagine, for example,the phrase VP->VB NP PP,where the head verb VB is in the object relation with the head of the noun phrase NP and also in the modifying relation with the head of the prepositional phrase PP.Let us also assume that the correct sense of the verb VB is a. Even if the verb-object relation provided a strong selectional support for the sense a,if there was no example in the training set for the second relation(between VB and PP)which would

62

Stetina,J.et al. General Word Sense Disambiguation Method

score a hit for the sense a,multiplying the scores of that sense derived from the first and from the second relation respectively,would gain a zero probability for this sense and thus prevent its correct assignment.Although such a situation is quite unlikely because of the use of the semantic distance in the hit scoring(7),we experimentally verified that adding the scores rather that their multiplication provides slightly better results.During the implemen-tation of the algorithm we tested several modifications,including the traditional smoothing techniques.Most of the methods provided similar or marginally worse results,ranging from 79.1% accurate sense tags for multiplication,to 79.4% of accurate sense tags for addition(see further discussion in the Evaluation section).The problem with traditional smoothing tech-niques,e.g.flooring,is that these are designed for domains which deal with real probabilities, rather than with scores as in our case.Although the initial score vectors(9a)contain good approximations of the contextually independent probabilities of each sense of the given word, after their first propagation through the bottom-up phase,these probabilities are replaced by scores or sense evaluations based on the products of the semantic distances to the samples in the training corpus.This substantially reduces the sparse data problem,because instead of counting exact word-pair occurrences,we base the score estimate on the semantic distance measure.Therefore,for almost every word-pair there is a non-zero sense combination score. This can be view as a kind of smoothing over the sparse data.
The newly created head word sense score vector G propagates upwards in the parse tree and the same process repeats at the next syntactic level.Note that at the higher level,de-pending on the head extraction rules described in section 3,the roles may be changed and the former head word may become a modifier of a new head(and participate in the above calculation as a modifier).The process repeats itself until the root of the tree is reached.The word sense score vector which has reached the root,represents a vector of scores of the senses of the main head word of the sentence(verb said in the example in figure 1),which is based on the whole syntactic structure of that sentence.The sense with the highest score is selected and the sentence head disambiguated.

Top-down Disambiguation

Having ascertained the sense of the sentence head,the process of top-down disambiguation begins.The top-down disambiguation algorithm,which starts with the sentence head,can be described recursively as follows:
Let 1 be the sense of the head word h on the input.Let M=[ml,m2,...,mx] be the set of the modifiers of the head word h.For every modifier mi E M,do:

63

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

(1) In the sense score matrix Qi of the modifier mi(calculated in step 1.2 of the bottom-
up phase)find all the elements x(ki,1),where 1 is the sense of the head h
(2) Assign the modifier mi such a sense k=ki for which the value x(ki,1)is maximum.
In the case of a draw,choose the sense which is listed as more frequent in WordNet.
(3) If the modifier mi has descendants in the parse tree,call the same algorithm again
with mi being the head and k being its sense,else end.
The disambiguation of the modifiers(which become heads at lower levels of the parse tree), is based solely on those lines of their sense score matrices which correspond to the sense of the head they are in relation with.This is possible because of our assumption that the mod-ifiers are related only to their head words,and that there is no relation among the modifiers themselves.To what extent this assumption holds in real life sentences,however,has yet to be investigated.

6 DISCOURSE CONTEXT

The objective of any human communication is to convey meaning.This objective would be obliterated if we could not identify the senses of the used words.An important factor in doing so is the context in which the words occur.This context consists of the physical,psychological and cultural environment of the participants in the communication(Benjamin 1986).The sentences of natural language are structured so that the listener(or reader)understands the conveyed message without doubts.If the context in which a message occurs does not provide sufficient clues for the listener to understand,the speaker may choose to specify the meaning in the preceding sentences.(Yarowsky 1995)pointed out that the sense of a target word is highly consistent within any given document(one sense per discourse).Because our algorithm does not consider the context given by the preceding sentences,we have conducted the following experiment to see to what extent the discourse context could improve the performance of the word-sense disambiguation:
Using the semantic concordance files(Miller et al.1993),we have counted the occurrences of content words which previously appear in the same discourse file.The experiment indicated that the "one sense per discourse" hypothesis works fairly well for nouns,however,the evi-dence is much weaker for verbs,adverbs and adjectives.Table 5 shows the numbers of content words which appear previously in the same discourse with the same meaning(same synset), and those which appear previously with a different meaning.The experiment also confirmed our expectation that the ratio of words with the same sense to those with a different sense, depends on the distance of sentences in which the same words appear(distance 1 indicates

64

Stetina,J.et al. General Word Sense Disambiguation Method

Table 5 Number of words with the same and different sense as its previous occurrence
in the same discourse(shortened)

that the same word appeared in the previous sentence,distance 2 that the same word was present 2 sentences before,etc.).
We have modified the disambiguation algorithm to make use of the information gained by the above experiment in the following way:All the disambiguated words and their senses are stored.The words of all the input sentences are first compared with the set of these stored word-sense pairs.If the same word is found in the set,the initial sense score assigned to it by (9a)is modified using Table 5,so that the sense,which has been previously assigned to the word,gets higher priority8.The calculation of the initial sense score(9a)is thus replaced by (9b):

(9b)

where e(POS,SN)is the probability that the word with syntactic category POS which al-ready occurred SN sentences before,has the same sense as its previous occurrence.If,for example,the same noun has occurred in the previous sentence(SN=1)where it was assigned sense n,the probability of sense n of the same noun in the current sentence is multiplied by e(NOUN,1)=3,039/(3,039+103)=0.967,while all the probabilities of its remaining senses are multiplied by 1-0.967=0.033.If no match is found,i.e.the word has not previously occurred in the discourse,e(POS,SN)is set to 1 for all senses.

7 EVALUATION

To evaluate the algorithm,we randomly selected 15 files(with a total of 18,413 content
words tagged in SemCor)from the set of 103 files of the sense tagged section of the Brown
8 Word-sense pairs from closer sentences are checked first.

65

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

Table 6 Result Accuracy[%]

Corpus.Each tested file was removed from the set and the remaining 102 files were used for learning(Chapter 4).Every sense assigned by the hierarchical disambiguation algorithm (Chapter 5)was compared with the sense from the corresponding semantic concordance file. Table 6 shows the achieved accuracy compared with the accuracy which would be achieved by a simple use of the most frequent sense9.
As the above table shows,the accuracy of the word sense disambiguation achieved by our method was better than using the first sense for all lexical categories.In spite of a very small training corpus,the overall word sense accuracy exceeds 80%.To our knowledge,there is no current method which attempts to identify the senses of all words in whole sentences,so we cannot make a practical comparison.
An interesting aspect of the result was the relatively small contribution of the discourse context to the overall accuracy.As was expected,the discourse context had a relatively higher contribution to the accuracy of nouns and adjectives,than to verbs and adverbs.This is be-cause nouns and adjectives more often share the same sense in multiple occurrences within the same discourse,than verbs and adverbs(Table 5).It is possible to further improve the discourse context accuracy contribution by searching the preceding sentences,not only for same words but also for words expressing semantically close concepts.If,for example,one sentence contains the noun car,disambiguated to its 'vehicle' sense,and one of the subse-quent sentences includes the as yet ambiguous noun train,we feel that the semantic similarity between the two should be incorporated in the estimate of the initial sense probability(9b) of the noun train.Another improvement could be made by the use of the semantic relations extracted from the previously located sentences in the same discourse.In the learning process, only the sentences from other files are used in order to avoid learning on the tested data.We believe,however,that in any natural discourse,it is not merely the presence of entities that establishes a context,but also relations among them.If,for example,a word wl is in semantic 9 Collocations were treated as one word,e.g.primary election was treated as a one word term primary_election. Words,which were not found in WordNet,were disambiguated against,the sense candidates of their lemma from the semantic concordance.If an unknown word did not have an equivalent in the concordance file,it was not counted.Words,which constitute names,e.g.;Mayor William B.Hartsfield and which are joined into a single lemma in the semantic concordance,e.g.person,were disambigauted against the possible sense candidates of the lemma.

66

Stetina,J.et al. General Word Sense Disambiguation Method

relation r with word w2 in one sentence(r(wl,w2)),and a word w3 is the same relation with word w4 in one of the subsequent sentences(r(w3,w4)),we hypothesise that the use of the semantic similarity of the arguments would yield a further increase of accuracy.
Another way to improve the accuracy would be to consider each discourse file as a whole, where,not only preceding,but also subsequent sentences influence the probabilities of word senses.This would,however,require the thorough investigation of many discourse context phenomena,and the implementation of a much more complex probabilistic model,which is beyond the scope of this work.
The relatively low accuracy in the case of verbs is to be attributed to the fact that verbs are naturally more difficult to disambiguate and have on average more senses(Table 1).(Leacock, Towell,and Voorhees 1993)suggest that some words are harder to disambiguate than others, but they state that the overall accuracy is a function of the difficulty,rather than being strictly a function of the number of senses.In any case,however,we believe that disambiguation of verbs involves the need of more complicated inferences with general knowledge,and this is not possible to substitute by a small training corpus.
As has been already implied by many researchers(Wilks 1990)(Bruce and Wiebe 1994) etc.,determining the correct sense of a word involves subtle human judgements.The WordNet lexical hierarchy,whose sense definitions have been used in this work,contains many senses and choosing the right one for a given word in a given context is not always easy even for human annotators.The semantic concordance files,which were used for learning and evalu-ation,were tagged manually10.The manual tagging required the annotator to examine each word of the text in its context and to decide which WordNet sense is correct(Miller et al. 1993).In many situations,we believe,the decision was not the only one possible.Different annotators could choose different senses in the same situations.This is particularly true for verbs,the senses of which,we feel,are divided too finely in WordNet,making it possible for several candidates to fit the given context.The evaluation results given in Table 6 are based on exact matches between the automatically determined sense and the corresponding sense in the semantic concordance.If the sense was different,it was counted as an error,even when the automatic tag was plausible.Because it was not within our capacity to manually compare the sense errors in the entire tested set,we have conducted a smaller experiment limited only to one file11.Using the general disambiguation algorithm,a sense was assigned to every content word in the file.Every word,whose sense was different from the sense in the semantic concor-dance,was manually reviewed and counted correct if the automatically assigned sense was also 10 Error rate measured by the autors is around 10% for polysemous words.
11 We have used the first file of the Brown Corpus,i.e.a01,which contains 1025 content words in 90 sentences. 67

Journal of Natural Language Processing Vol.5 No.2 Apr.1998

Table 7 Informative manual evaluation accuracy

possible.Table 7 shows the results.It has to be noted,however,that this evaluation involved subjective judgements and should be therefore considered only informatively.Since we have tried to err on the side of accuracy strictness when considering possible sense candidates,the accuracy could possible be even higher.However,further research needs to be done in this

area.

The general disambiguation algorithm described in this paper is easy to implement and computationally very simple.Because the number of semantic relations in a sentence and the ambiguity of its words represent a constant factor,the computational complexity of the algo-rithm grows linearly with the size of the training corpus12.The sparse data problem of a small training corpus is compensated for by the use of semantic similarity.We believe that more training data would provide an even higher accuracy,but due to the lack of other semantically tagged corpora,we cannot test the performance on more data.The biggest advantage of the method is that all the words in a sentence are considered as one set of inter-related concepts. The utilisation of the syntactic tree makes the method extremely useful for integration with a probabilistic syntactic analyser into a powerful natural language processing system.

8 Related Work

The vast majority of work on Word Sense Disambiguation focuses on the identification of the correct senses of content(open-class)words,particularly nouns.Most of the methods attempt to identify the word sense by comparing the context in which the word appears with the context of the same word in example situations.Acquisition of these example situations constitutes the bottleneck of all example-based methods.Some researchers tried to get around the example acquisition problem by building large expert systems by hand(Hirst 1987)(Small and Rieger 1983)and/or focus on limited domains(Isabelle 1984)(Hirschman 1986).Others such as(Lesk 1996)(Luk 1993)have turned to machine readable dictionaries in the hope that

12 It took approximately 1 hour to disambigaute the whole testing set on a SUN SS20 workstation.
68

Stetina,J.et al. General Word Sense Disambiguation Method

they might provide a way out of the knowledge acquisition bottleneck.Another group of researchers(Dagan,Alon,and Ulrike 1991)have argued that the knowledge on senses can be gained by using two rather than one language.This approach is very promising.However, it could be argued that three,four,five languages are better than two.For example,both the English noun interest and the French equivalent interet are similarly ambiguous in both languages,so the use of equivalent examples from one other(especially similar)language does not always solve the problem.(Gale et al.1993)gave an excellent analysis of the disambigua-tion technique based on the Canadian Hansards,a corpus with parallel English and French text.Using a context of +/-fifty words,they have achieved a disambiguation performance from 86 to 90% for disambiguating two equiprobable senses of six 'difficult' nouns.
The biggest group of researchers turns to hand-annotated corpora,the strategy we have also adopted.Earlier attempts,such as(Kelly and Stone 1975)(Weiss 1973)(Black 1987)(Black 1988)used manually created rules to perform word sense disambiguation,while more recent work uses various sorts of statistical processing.The biggest challenge of the statistical ap-proaches is in the selection and probabilistic combination of contextual features.(Brown, Pietra,and Mercer 1991)(Dagan et al.1991)(Yarowsky 1993)present techniques for identi-fying optimal feature to perform disambiguation.Naive Bayes classifier has been found to perform well for(Bruce and Wiebe 1994)(Leacock et al.1993)(Mooney 1996).Recently,a considerable amount of attention has been given to maximum entropy models(Berger et al. 1996),which have been used to express the interactions among multiple contextual features. Every model,however,which,like ours,is based on maximum likelihood estimate(MLE),
has to deal with sparse data problem.Because no systematic study of interactions among the multiple variables in MLE has been carried out,researchers modify the MLE baseline to suit the specific needs of their domains.(Gale et al.1993)combine the probabilities from local context with probabilities from entire training corpus in an interpolation procedure based on the Beta function.(Bruce,Pedersen,and Wiebe 1997)propose sequential search to find the best model of interactions of multiple features.(Ng and Lee 1996)integrate multiple features (part-of-speech,surrounding words,local collocations and verb-object syntactic relation)in a specific heuristic formula to disambiguate senses of 121 most common nouns and 70 most common verbs.Similarly to our method,they use WordNet sense definitions,but their train-ing corpus has been manually annotated to be used only with these selected words.Ng and Lee tested the method on 7,119 occurrences of the selected nouns and verbs that occurred in 50 text files of the Brown Corpus(BC)and 14,139 occurrences in Wall Street Journal(WSJ). Testing only these polysemous words they report an average accuracy 54.0 and 68.6 percent,

69

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

for BC and WSJ,respectively.,
Similarly to our work,also(Resnik 1995)(Agirre and Rigau 1996)challenge the fine-grainedness of WordNet,but their work is limited to nouns only.(Agirre and Rigau 1996) report coverage 86.2%,precision 71.2% and recall 61.4% for nouns in four randomly selected semantic concordance files.From among the methods based on semantic distance,(Resnik 1993)(Sussna 1993)use a similar semantic distance measure for two concepts in WordNet, but they also focus on selected group of nouns only.(Karov and Edelman 1996)use an inter-esting iterative algorithm and attempt to solve the sparse data bottleneck by using a graded measure of contextual similarity.They achieve 90.5,92.5,94.8 and 92.3 percent accuracy in distinguishing between two senses of the noun drug,sentence,suit and player,respectively. (Yarowsky 1995),whose training corpus for the noun drug was 9 times bigger than that of Karov and Edelman,reports 91.4% correct performance improved to impressive 93.9% when using the "one sense per discourse" constraint.These methods,however,focus on only two senses of a very limited number of nouns and therefore are not comparable with our approach. All existing methods face the problem of defining context and selecting appropriate con-textual features from it.Most of the methods use a window of surrounding words without considering the syntactic relations among them.Our method,on the other hand,explores the syntactic structure of the sentence and estimates the probability of each word sense as a product of all the semantic relations in the sentence.(Collins 1996)took a similar approach in combining the probabilities of dependencies between head words in grammar rules of a parse tree to achieve statistical parsing.Because Collins also exploits bigram dependencies in a whole sentence,our method would be particularly suitable to be integrated within a statistical parser of this kind.

9 Conclusion

This paper presents a new general approach to word sense disambiguation.Unlike most
of the existing methods,it identifies the senses of all content words in a sentence based on an estimation of the overall probability of all semantic relations in that sentence.By using the semantic distance measure,our method reduces the sparse data problem since the training examples and their contexts do not have to match the disambiguated words exactly.All the semantic relations in a sentence are combined according to the syntactic structure of the sen-tence,which makes the method particularly suitable for integration with a statistical parser into a powerful Natural Language Processing system.The method is designed to work with any type of common text and is capable of distinguishing among many word senses.It has a 70

tetina,J.et al. General Word Sense Disambiguation Method

very wide scope of applicability and is not limited to only one part-of-speech.Although the testing results exhibit relatively high disambiguation accuracy,there is still space for further improvement.In the first place,the method requires a semantically tagged training corpus. However,we believe that a huge amount of untagged data would allow for an unsupervised modification of the algorithm.The relational probabilities,instead of being based on disam-biguated senses of training examples,could be based on such a selection of training senses which would be semantically nearest to the given tested pair.The algorithm would be natu-rally more complicated,however,this would avoid the need of manual tagging.To what extent the amount of data would substitute the semantic tags remains yet to be tested.Another weak point of our method is in its poor use of discourse context.Semantically close concepts in surrounding sentences,their relations(including cross-category relations)and their influence on the probability bias of word senses,needs to be investigated more thoroughly.Moreover,
we assumed that there is no dependency among the semantic relations between phrase heads and their modifiers.This is not necessarily true,but it substantially simplified the algorithm. Also,we believe that more accurate handling of the scores derived from multiple contextual features is possible and it remains an objective of our further research.

Reference

Agirre, E., and Rigau, G.(1996)."Word Sense Disambiguation Using Conceptual Density." In Proc. of COLLING, pp.16-22.
Benjamin, J.(1986). Communication: Concepts and Contexts. Harper & Row, Publishers. Berger, A., Pietra, V., and Pietra, S.(1996)."A Maximum Entropy Approach to Natural Language Processing." Computational Linguistics, 1(22), 39-72.
Black, E.(1987). Towards Computational Discrimination of English Word Senses. Ph. D. Thesis. City University of New York.
Black, E.(1988)."An Experiment in Computational Discrimination of English Word Senses." IBM Journal of Research and Development.
Brown, P., Pietra, S., and Mercer, R.(1991)."Word-sense Disambiguation Using Statistical Methods." In Proc. of the 29th Annual Meeting of the ACL.
Bruce, R., Pedersen, T., and Wiebe, J.(1997)."Sequential Model Selection for Word Sense Disambiguation." In Proc. of the 5th Annual Meeting of the ACL.
Bruce, R., and Wiebe, J.(1994)."Word-sense Disambiguation Using Decomposable Models." In Proc. of the 32nd Annual Meeting of the ACL.

71

Journal of Natural Language Processing Vol.5 No.2 Apr.1998

Collins, M.(1996)."A New Statistical Parser Based on Bigram Lexical Dependencies." In Proc. of the 34th Annual Meeting of the ACL, pp.184-191.
Dagan, I., Alon, I., and Ulrike, S.(1991)."Two Languages are More Informative than One." In Proc. of the 29th Annual Meeting of the ACL.
Deese, J.(1964)."The Associative Structure of Some Common English Adjectives." Journal of Verbal Learning and Verbal Behavior, 3(5), 347-357.
Gale, W., Church, K., and Yarowsky, D.(1992)."Estimating upper and lower bounds on the performance of word-sense disambiguation programs." In Proc. of the 30th Annual Meeting of the ACL, pp.249-256.
Gale, W., Church, K., and Yarowsky, D.(1993)."A Method for Disambiguating Word Senses in a Large Corpus." Computers and Humanities, pp.415-4397.
Hirschman, L.(1986). Discovering Sublanguage Discovery. In Analyzing Language in Re-stricted Domains. Ed. Ralph Grisham. Lawrence Erlbaum.

Hirst, G.(1987). Semantic Intrepretation and the Resolution of Ambiguity. Cambridge Uni-

versity Press.

Isabelle, P.(1984). Machine Translation at the TA UM Group. In Machine Translation Today: The State of the Art. Edinburgh University Press.

Jelinek, F., Lafferty, J., Magerman, D., Mercer, R., Rathnaparkhi, A., and Roukos, S.(1994). "Decision Tree Parsing Using a Hidden Derivation Model." In Proc. of the ARPA Human Language Technology Workshop, pp.272-277.
Justeton, J., and Katz, S.(1991)."Co-occurrences of Antonymous Adjectives and Their Contexts." Computational Linguistics, 1(17), 1-19.
Karov, Y., and Edelman, S.(1996)."Learning Similarity-based Word Sense Disambiguation from Sparse Data." In Proc. of the 3rd Workshop on Very Large Corpora, pp.42-55. Kelly, E., and Stone, P.(1975). Computer Recognition of English Word Senses. North-Holland. Leacock, C., Towell, G., and Voorhees, E.(1993)."Corpus-Based Statistical Sense Resolu-tion." In Proc. of the ARPA Human Language Technology Workshop, pp.260-265.
Lesk, M.(1996)."Automated Sense Disambiguation Using Machine-Readable Dictionaries: How to tell a pine cone from an ice cone." In Proc. of the SIGDOC Conference.
Luk, A.(1993)."Statistical Sense Disambiguation with Relatively Small Corpora Using Dic-tionary Definitions." In Proc. of the 33rd Annual Meeting of ACL, pp.181-188. Magerman, D.(1995)."Statistical Decision-Tree Models for Parsing." In Proc. of the 33rd Annual Meeting of ACL, pp.276-283.
Marcus, M.(1993)."Building a Large Annotated Corpus of English: The Penn Treebank."

72

Stetina,J.et al. General Word Sense Disambiguation Method

Computational Linguistics, 2(19), 313-330.
Miller, G.(1990)."WordNet: An On-line Lexical Database." International Journal of Lexi-cography, 3(4), 235-312.
Miller, G.(1994)."Using A Semantic Concordance for Sense Identification." In Proc. of the ARPA Human Language Technology Workshop, pp.240-243.
Miller, G., Leacock, C., and Tengi, R.(1993)."A Semantic Concordance." In Proc. of the ARPA Human Language Technology Workshop, pp. 303-308.
Mooney, R.(1996)."Comparative Experiments on Disambiguating Word Senses: An Illustra-tion of the Role of Bias in Machine Learning." In Proc. of the Conference on Empirical Methods in Natural Language Processing.
Ng, H., and Lee, H.(1996)."Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach." In Proc. of the 34th Annual Meeting of the ACL, pp.40-47.
Resnik, P.(1993)."Semantic Classses and Syntactic Ambiguity." In Proc. of the APRA Human Language Technology Workshop, pp.278-283.
Resnik, P.(1995)."Disambiguating Noun Groupings with Respect to WordNet Senses." In Proc. of the 3rd Workshop on Very Large Corpora.
Sapir, E.(1944)."Grading: A Study in Semantics." Philosophy of Science, pp. 83-116. Small, S., and Rieger, C.(1983). Parsing and Comprehending with Word Experts (A Theory and its Realization). In Strategies for NLP. Lawrence Erlbaum.
Stetina, J., and Nagao, M.(1997)."Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary." Journal of NLP.
Sussna, M.(1993)."Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network." In Proc. of Second International Conference on Information and

Knowledge Management, pp.67-74.
Weiss, D.(1973)."Learning to Disambiguate." Information Storage and Retrieval, pp. 33-41. Wilks, Y.(1990)."Providing Machine Tractable Dictionary Tools." Machine Translation, 2(5), 99-154.
Yarowsky, D.(1993)."One Sense per Collocation." In Proc. of the ARPA Human Language Technology Workshop, pp.266-271.
Yarowsky, D.(1995)."Unsupervised Word Sense Disambiguation Rivaling Supervised Meth-ods." In Proc. of the 32nd Annual Meeting of the ACL.

73

Journal of Natural Language Processing Vol.5 No.2 Apr. 1998

(Received July 25, 1997)
(Revised October 8, 1997)
(Rerevised December 15, 1997)
(Accepted January 23, 1998)

74