Soft Word Sense Disambiguation

Ganesh Ramakrishnan, B. P. Prithviraj, A. Deepa, Pushpak Bhattacharyya, and Soumen Chakrabarti

Department of Computer Science and Engineering, Indian Institute of Technology, Mumbai, India
Email: hare@cse.iitb.ac.in, prithvir@cse.iitb.ac.in, adeepa@cse.iitb.ac.in,
pb@cse.iitb.ac.in, soumen@cse.iitb.ac.in

Abstract. Word sense disambiguation is a core problem in many tasks related to language processing. In this paper, we introduce the notion of *soft word sense disambiguation* which states that *given a word, the sense disambiguation system should not commit to a particular sense, but rather, to a set of senses which are not necessarily orthogonal or mutually exclusive*. The senses of a word are expressed by its WordNet synsets, arranged according to their relevance. The relevance of these senses are probabilistically determined through a Bayesian Belief Network. The main contribution of the work is a completely probabilistic framework for word-sense disambiguation with a semi-supervised learning technique utilising WordNet. WordNet can be customized to a domain using corpora from that domain. This idea applied to question answering has been evaluated on TREC data and the results are promising.

Keywords: Soft Sense Disambiguation, Synset-Ranking, Bayesian Belief Networks, Semi-supervised learning

1 Introduction

Word sense disambiguation is defined as the task of finding *the* sense of a word in a context. In this paper, we explore the idea that one should not commit to a particular sense of the word, but rather, to a *set of its senses* which are not necessarily orthogonal or mutually exclusive. Very often, WordNet gives for a word multiple senses which are related and which *help connect* other words in the text. We refer to this observation as the relevance of the sense in that context. Therefore, instead of picking a single sense, we rank the senses according to their relevance to the text. As an example, consider the usage of the word *bank* in fig. 1. In WordNet, *bank* has 10 noun senses. The senses which are relevant to the text are shown in figure 2.

A passage about some bank A Western Colorado bank with over $320 Million in assets, was formed in 1990 by combining the deposits of two of the largest and oldest financial institutions in Mesa County

Fig. 1. One possible usage of *bank* as a *financial_institution*

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 291–298.

Relevant senses

1. *depository financial institution, bank, banking concern, banking company:* a financial institution that accepts deposits and channels the money into lending activities; “he cashed a check at the bank”; “that bank holds the mortgage on my home”
2. *bank, bank building:* a building in which commercial banking is transacted; “the bank is on the corner of Nassau and Witherspoon”
3. *bank:* (a supply or stock held in reserve for future use (especially in emergencies))
4. *savings bank, coin bank, money box, bank:* (a container (usually with a slot in the top) for keeping money at home; “the coin bank was empty”)

Fig. 2. Some relevant senses for *bank*

These senses are ordered according to their relevance in this context. It is apparent that the first two senses have equal relevance. The applicability of the senses tapers off as we move down the list. This example motivates soft sense disambiguation. We define *soft sense disambiguation* as the process of enumerating the senses of a word in a ranked order. This could be an end in itself or an interim process in an IR task like question answering.

1.1 Related Work

[Yarowsky 1992] proposes a solution to the problem of WSD using a thesaurus in a supervised learning setting. Word associations are recorded and for an unseen text, the senses of words are detected from the learnt associations. [Agirre and Rigau 1996] uses a measure based on the proximity of the text words in WordNet (*conceptual density*) to disambiguate the words. The idea that translation presupposes word sense disambiguation is leveraged by [Nancy 1999] to disambiguate words using bi-lingual corpora. The design of the well-known work-bench for sense disambiguation *WASP* is given in [Kilgarriff 1998]. The idea of constructing a BBN from WordNet has been proposed earlier by [Wiebe, Janyce, et al. 1998] and forms a motivation for the present work. However, unlike [Wiebe, Janyce, et al. 1998] we particularly emphasise the need for soft sense disambiguation, *i.e.* synsets are considered to probabilistically cause their constituent words to appear in the texts. Also we describe a comprehensive training methodology and integrate soft WSD into an interesting application, *viz.*, QA. Bayesian Balief Network (BBN) is used as the machine for this probabilistic framework. It is also demonstrated, how the BBN can be customized to a domain using corpora from that domain.

2 Our Approach to Soft WSD

We describe how to induce a Bayesian Belief Network (BBN) from a lexical network of relations. Specifically, we propose a semi-supervised learning mechanism which simultaneously trains the BBN and associates text tokens, which are words, to synsets in WordNet in a probabilistic manner (“soft WSD”).

In general, there could be multiple words in the document that are caused to occur together by multiple hidden concepts. This scenario is depicted in figure 3. The causes themselves may have hidden causes.

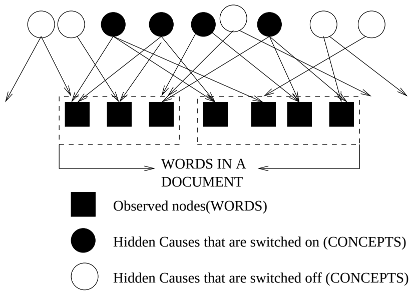

WORDS IN A DOCUMENT
Observed nodes(WORDS)
Hidden Causes that are switched on (CONCEPTS)
Hidden Causes that are switched off (CONCEPTS)

Fig. 3. Motivation

These causal relationships are represented in WordNet which encodes relations between words and concepts (synsets). For instance WordNet gives the *hypernymy* relation between the concepts { animal} and { bear}.

2.1 Inferencing on Lexical Relations

It is difficult to link words to appropriate synsets in a lexical network in a principled manner. On the example of *animal* and *bear*, the English WordNet has five synsets on the path from *bear* to *animal*: {carnivore...}, {placental_mammal...}, {mammal...}, {vertebrate..}, {chordate...}. Some of these intervening synsets would be extremely unlikely to be associated with a corpus that is not about zoology; a common person would more naturally think of a *bear* as a kind of animal, skipping through the intervening nodes.

Clearly, any scoring algorithm that seeks to utilize WordNet link information must also *discriminate* between them based (at least) on usage statistics of the connected synsets. Also required is an estimate of the likelihood of instantiating a synset into a token because it was *activated* by a closely related synset. We find a Bayesian belief network (BBN) a natural structure to encode such combined knowledge from WordNet and corpus (for training).

2.2 Building a BBN from WordNet

Our model of the BBN is that each synset from WordNet is a boolean *event* associated with a word. Textual tokens are also events. Each event is a node in the BBN. Events can *cause* other events to happen in a probabilistic manner, which is encoded in Conditional Probabiity Tabless. The specific form of CPT we use is the well-known *noisy-OR* for the words and *noisy-AND* for the synsets. This is because a word is *exclusively* instantiated by a cluster of parent synsets in the BBN, whereas a synset is compositionally instantiated by its parent synsets. The noisy-OR and noisy-AND models are described in [J. Pearl 1998].

We introduce a node in the BBN for each noun, verb, and adjective synset in WordNet. We also introduce a node for each token in the corpus. Hyponymy, meronymy, and attribute links are introduced from WordNet. *Sense links* are used to attach tokens to potentially matching synsets. For example, the string “flag” may be attached to synset nodes {sag, droop, swag, flag} and {a conspicuously marked or shaped tail}. (The purpose of probabilistic disambiguation is to estimate the probability that the string “flag” was *caused* by each connected synset node.)

This process creates a hierarchy in which the parent-child relationship is defined by the semantic relations in WordNet. A is a parent of B iff A is the *hypernym* or *holonym* or *attribute-of* or A is a synset containing the word B. The process by which the BBN is built from WordNet graph of synsets and from the mapping between words and synsets is depicted in figure 4. We define *going-up* the hierarchy as the traversal from child to parent.

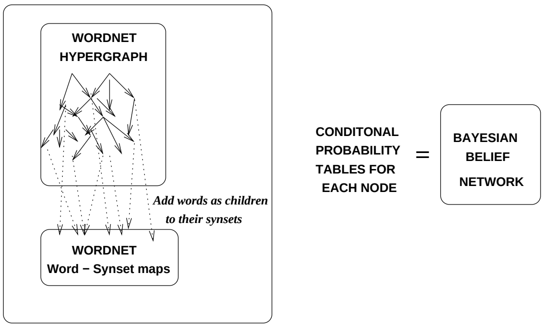


Fig. 4. Building a BBN from WordNet and associated text tokens.

2.3 Training the Belief Network

The figure 5 describes the algorithm for training the BBN obtained from the WordNet. We initialize the CPTs as described in the previous section. The instances we use for training are windows of length M each from the untagged corpus. Since the corpus is not tagged with WordNet senses, all variables, other than the words observed in the window (i.e. the synset nodes in the BBN) are hidden or unobserved. Hence we use the Expectation Maximization algorithm [Dempster 1977] for parameter learning. For each instance, we find the expected values of the hidden variables, given the “present” state of each of the observed variables. These expected values are used after each pass through the corpus to update the CPT of each node. The iterations through the corpus are done till the sum of the squares of Kullback-Liebler divergences between CPTs in successive iterations do not differ more than a small threshold. In this way we customize the BBN CPTs to a particular corpus by learning the local CPTs.

3 The WSD Algorithm: Ranking Word Senses

Given a passage, we clamp the BBN nodes corresponding to words, to a state of ‘present’ and infer using the network, the score of each of its senses which is the probability of the corresponding synset node being in a state of “present”. For each word, we rank its senses in decreasing order of its score. In other words, the synset given the highest rank (probability) by this algorithm becomes the most probable sense of the Word.

```
while CPTs do not converge do
    for each window of $M$ words in the text do
        Clamp the word nodes in the Bayesian Network to a state of ‘present’
        for each node in Bayesian network do
            find its joint probabilities with all configurations of its parent nodes (E Step)
        end for
    end for
    Update the conditional probability tables for all random variables (M Step)
end while
```

Fig. 5. Training the Bayesian Network for a corpus

```
Load the Bayesian Network parameters
for each passage p do
    clamp the variables (nodes) corresponding to the passage words ($w_1$, $w_2 ... w_n$) in network to
    a state of ‘present’
    Find the probability of each sense of each word, being in state ‘present’ i.e., $\Pr(s|w_1, w_2..w_n)$
end for
Report the word senses of each word, in decreasing order of ranks.
```

Fig. 6. Ranking word senses

4 Evaluation

We use documents from *Semcor 1.7.1 corpus* [Semcor] for disambiguation. Semcor corpus is a subset of the famous Brown corpus [Brown Corpus] sense-tagged with WordNet 1.7.1 synsets. Our soft WSD system produces rank ordered synsets on the semcor words (at most two senses). We show below in figure 7 the output of the system for the word *study*. Both semcor’s tag and our system’s first tag are correct, though they differ. The second tag from our system has low weightage and is wrong in this context. The synsets marked with ** represent the correct meaning.

Passage from Semcor It recommended that Fulton legislators act to have these laws *studied* and revised to the end of modernizing and improving them.

Semcor tag: [Synset: [Offset: 513626] [POS: verb] Words: analyze, analyse, study, examine, canvass – (consider in detail and subject to an analysis in order to discover essential features or meaning; “analyze a sonnet by Shakespeare”; “analyze the evidence in a criminal trial”; “analyze your real motives”)]

soft WSD tags: **[Synset: study 0 consider 0 [Gloss =]: give careful consideration to; “consider the possibility of moving” [Score = 0.62514]]

[Synset: study 4 meditate 2 contemplate 0 [Gloss =]: think intently and at length, as for spiritual purposes; “He is meditating in his study” [Score = 0.621583]]

Fig. 7. Example of *first match* with Semcor’s marking

Next we present an example of the second marking of the sense being correct. The word in question is the verb *urge* (figure 8).

Passage from Semcor It *urged* that the city take steps to remedy this problem.

Semcor tag: Synset: [Offset: 609547] [POS: verb] Words: urge, urge_on, press, exhort – (force or impel in an indicated direction; "I urged him to finish his studies")

soft WSD tags: [Synset: cheer 1 inspire 1 urge 1 barrack 1 urge_on 1 exhort 1 pep_up 0 [Gloss =]: urge on or encourage esp. by shouts; "The crowd cheered the demonstrating strikers" [Score = 0.652361]]

**[Synset: recommend 1 urge 3 advocate 0 [Gloss =]: push for something; "The travel agent recommended strongly that we not travel on Thanksgiving Day" [Score = 0.651725]]

Fig. 8. Example of the *second match* being correct

Table 1 summarizes soft WSD results obtained by us. If the first meaning given by the soft WSD system is correct then it is counted towards the *first match*; similarly for the *second match*.

Table 1. Results of soft WSD

Total ambiguous nouns	139
Nouns first match	66
Nouns second match	46
Total ambigous verbs	67
verbs first match	24
verbs second match	23

5 An Application: Question Answering

In this section, we mention our work on the extension of ideas presented in the previous sections to the problem of question answering, which inherently requires WSD to connect question words to answer words. The BBN is trained using the algorithm in figure 5 on the corpus to be queried. The trained BBN is used to rank passages (windows of N consecutive words) from the corpus using the algorithm presented in figure 9.

We performed QA experiments on the TREC-9 question-set and the corresponding corpus. The Mean Reciprocal Rank (MRR) figures for the different experiments are presented in table 2. Clearly, inferencing with trained BBN outperforms inferencing with untrained BBN while both inferencing procedures, outperform the baseline algorithm, the standard TFIDF retrieval system.

*The effect of WSD:*It is interesting to note that training does not substantially affect disambiguation accuracy (which stays at about 75%), and MRR improves *despite* this

```
Load the Bayesian Network parameters
for each question q do
  for each candidate passage p do
    clamp the variables (nodes) corresponding to the passage words in network to a state of 'present'
    Find the joint probability of all question words being in state 'present' i.e., $\Pr(q\,|\,p)$
  end for
end for
Report the passages in decreasing order of $\Pr(q\,|\,p)$
```

Fig. 9. Ranking candidate answer passages for given question

Table 2. MRRs for baseline, untrained and trained BBNs

System	MRR
Asymmetric TFIDF	0.314
Untrained BBN	0.429
Trained BBN	0.467

fact. This seems to indicate that learning joint distributions between query and candidate answer keywords (via synset nodes, which are "bottleneck" variables in BBN parlance) is as important for QA as is WSD. Furthermore, we conjecture that "soft" WSD is key to maintaining QA MRR in the face of modest WSD accuracy.

6 Conclusions

In this paper a robust, semi-supervised method for sense disambiguation using WordNet (*soft sense disambiguation*) was described. The WordNet graph was exploited extensively. Also, the task of soft WSD was integrated into an application *viz.* question answering.

The future work consists in exploring the use of links others than the hypernymy-hyponymy. Also WordNet 2.0 provides derivational morphology links between verb and noun synsets, the use of which needs to be investigated. Adjectives and adverbs too have to be tackled in the system. The intervention of human experts at critical steps to improve accuracy is a very interesting issue meriting attention.

The paradigm of *active learning* is highly promising in such problems as are the concern of the present work. With human help the system can tune itself for sense disambiguation using a relatively small number of examples.

References

Fellbaum 1998. Fellbaum Christiane, ed. the WordNet: An Electronic Lexical Database. *MIT Press*, Map 1998.

Nancy 1999. Nancy Ide. Parallel Translations as Sense Discriminators. In *Proceedings of SIGLEX99, Washington D.C, USA, 1999.*

Kilgarriff 1998. Adam Kilgarriff. Gold Standard Data-sets for Evaluating Word Sense Disambiguation Programs. In *Computer Speech and Language 12 (4), Special Issue on Evaluation, 1998.*

Yarowsky 1992. David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the* 14^{th} *International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France, 1992.

Agirre and Rigau 1996. Agirre, E. and Rigau, G. Word sense disambiguation using conceptual density. In *Proceedings of COLING '96*.

J. Pearl 1998. J. Pearl. In *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann Publishers, Inc.

Wiebe, Janyce, et al. 1998. Wiebe, Janyce, O'Hara, Tom, Rebecca Bruce. Constructing bayesian networks from WordNet for word sense disambiguation: representation and processing issues In *Proc. COLING-ACL '98 Workshop on the Usage of WordNet in Natural Language Processing Systems*.

Dempster 1977. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Data via The EM Algorithm. In *Journal of Royal Statistical Society*, Vol. 39, pp. 1–38, 1977.

Semcor. `http://www.cs.unt.edu/~rada/downloads.html#semcor`.

Brown Corpus.
http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

TREC. TREC `http://trec.nist.gov`.

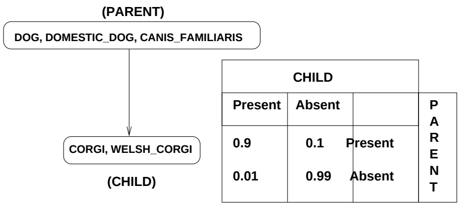
Appendix I: Bayesian Belief Network

A Bayesian Network for a set of random variables $X = \{X_1, X_2, \ldots, X_n\}$ consists of a directed acyclic graph (DAG) that encodes a set of conditional independence assertions about variables in X and a set of local probability distributions associated with each variable. Let $\mathbf{Pa}_i$ denote the set of immediate parents of X_i in the DAG, and $\mathbf{pa}_i$ a specific instantiation of these random variables.

The BBN encodes the joint distribution $\Pr(x_1, x_2, \ldots, x_n)$ as

$$\Pr(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \Pr(x_i | \mathbf{pa}_i) \tag{1}$$

Each node in the DAG encodes $\Pr(x_i | \mathbf{pa}_i)$ as a "conditional probability table" (CPT). Figure §10 shows a Bayesian belief network interpretation for a part of WordNet. The synset {*corgi, welsh_corgi*} has a causal relation from {*dog, domestic_dog, canis_familiaris*}. A possible conditional probability table for the network is shown to the right of the structure.



CHILD			
Present	Absent		PARENT
0.9	0.1	Present	
0.01	0.99	Absent	

Fig. 10. Causal relations between two synsets.