WSD as a Distributed Constraint Optimization Problem

Siva Reddy
IIIT Hyderabad
India
gvsreddy@students.iiit.ac.in

Abhilash Inumella
IIIT Hyderabad
India
abhilashi@students.iiit.ac.in

Abstract

This work models Word Sense Disambiguation (WSD) problem as a Distributed Constraint Optimization Problem (DCOP). To model WSD as a DCOP, we view information from various knowledge sources as constraints. DCOP algorithms have the remarkable property to jointly maximize over a wide range of utility functions associated with these constraints. We show how utility functions can be designed for various knowledge sources. For the purpose of evaluation, we modelled all words WSD as a simple DCOP problem. The results are competitive with state-of-art knowledge based systems.

1 Introduction

Words in a language may carry more than one sense. The correct sense of a word can be identified based on the context in which it occurs. In the sentence, *He took all his money from the bank*, *bank* refers to *a financial institution* sense instead of other possibilities like *the edge of river* sense. Given a word and its possible senses, as defined by a dictionary, the problem of Word Sense Disambiguation (WSD) can be defined as the task of assigning the most appropriate sense to the word within a given context.

WSD is one of the oldest problems in computational linguistics which dates back to early 1950's. A range of knowledge sources have been found to be useful for WSD. (Agirre and Stevenson, 2006; Agirre and Martínez, 2001; McRoy, 1992; Hirst, 1987) highlight the importance of various knowledge sources like part of speech, morphology, collocations, lexical knowledge base (sense taxonomy, gloss), sub-categorization, semantic word associations, selectional preferences, semantic roles, domain, topical word associations, frequency of senses, collocations, domain knowledge. etc. Methods for WSD exploit information from one or more of these knowledge sources.

Supervised approaches like (Yarowsky and Florian, 2002; Lee and Ng, 2002; Martínez et al., 2002; Stevenson and Wilks, 2001) used collective information from various knowledge sources to perform disambiguation. Information from various knowledge sources is encoded in the form of a feature vector and models were built by training on sense-tagged corpora. These approaches pose WSD as a classification problem. They crucially rely on hand-tagged sense corpora which is hard to obtain. Systems that do not need hand-tagging have also been proposed. Agirre and Martinez (Agirre and Martínez, 2001) evaluated the contribution of each knowledge source separately. However, this does not combine information from more than one knowledge source.

In any case, little effort has been made in formalizing the way in which information from various knowledge sources can be collectively used within a single framework: a framework that allows interaction of evidence from various knowledge sources to arrive at a global optimal solution.

Here we present a way for modelling information from various knowledge sources in a multi agent setting called distributed constraint optimization problem (DCOP). In DCOP, agents have constraints on their values and each constraint has a utility associated with it. The agents communicate with each other and choose values such that a global optimum solution (maximum utility) is attained. We aim to solve WSD by modelling it as a DCOP.

To the best of our knowledge, ours is the first attempt to model WSD as a DCOP. In DCOP framework, information from various knowledge sources can be used combinedly to perform WSD.

In section 2, we give a brief introduction of

Proceedings of the ACL 2010 Student Research Workshop, pages 13–18,
Uppsala, Sweden, 13 July 2010.

DCOP. Section 3 describes modelling WSD as a DCOP. Utility functions for various knowledge sources are described in section 4. In section 5, we conduct a simple experiment by modelling all-words WSD problem as a DCOP and perform disambiguation on Senseval-2 (Cotton et al., 2001) and Senseval-3 (Mihalcea and Edmonds, 2004) data-set of all-words task. Next follow the sections on related work, discussion, future work and conclusion.

2 Distributed Constraint Optimization Problem (DCOP)

A DCOP (Modi, 2003; Modi et al., 2004) consists of n variables $V = x_1, x_2, ...x_n$ each assigned to an agent, where the values of the variables are taken from finite, discrete domains $D_1, D_2, ..., D_n$ respectively. Only the agent has knowledge and control over values assigned to variables associated to it. The goal for the agents is to choose values for variables such that a given global objective function is maximized. The objective function is described as the summation over a set of utility functions.

DCOP can be formalized as a tuple (A, V, D, C, F) where

- A = $\{a_1, a_2, \ldots a_n\}$ is a set of n agents,
- V = $\{x_1, x_2, \ldots x_n\}$ is a set of n variables, each one associated to an agent,
- D = $\{D_1, D_2, \ldots D_n\}$ is a set of finite and discrete domains each one associated to the corresponding variable,
- C = $\{f_k : D_i \times D_j \times \ldots D_m \rightarrow \Re\}$ is a set of constraints described by various utility functions f_k. The utility function f_k is defined over a subset of variables V. The domain of f_k represent the constraints C_{f_k} and $f_k(c)$ represents the utility associated with the constraint c, where $c \in C_{f_k}$.
- F = $\sum_k z_k \cdot f_k$ is the objective function to be maximized where z_k is the weight of the corresponding utility function f_k

An agent is allowed to communicate only with its neighbours. Agents communicate with each other to agree upon a solution which maximizes the objective function.

3 WSD as a DCOP

Given a sequence of words W= $\{w_1, w_2, \ldots w_n\}$ with corresponding admissible senses $D_{w_i} = \{s^1_{w_i}, s^2_{w_i} \ldots\}$, we model WSD as DCOP as follows.

3.1 Agents

Each word w_i is treated as an agent. The agent (word) has knowledge and control of its values (senses).

3.2 Variables

Sense of a word varies and it is the one to be determined. We define the sense of a word as its variable. Each agent w_i is associated with the variable s_{w_i}. The value assigned to this variable indicates the sense assigned by the algorithm.

3.3 Domains

Senses of a word are finite in number. The set of senses D_{w_i}, is the domain of the variable s_{w_i}.

3.4 Constraints

A constraint specifies a particular configuration of the agents involved in its definition and has a utility associated with it. For e.g. If c_{ij} is a constraint defined on agents w_i and w_j, then c_{ij} refers to a particular instantiation of w_i and w_j, say $w_i = s^p_{w_i}$ and $w_j = s^q_{w_j}$.

A utility function $f_k : C_{f_k} \rightarrow \Re$ denote a set of constraints $C_{f_k} = \{D_{w_i} \times D_{w_j} \ldots D_{w_m}\}$, defined on the agents $w_i, w_j \ldots w_m$ and also the utilities associated with the constraints. We model information from each knowledge source as a utility function. In section 4, we describe in detail about this modelling.

3.5 Objective function

As already stated, various knowledge sources are identified to be useful for WSD. It is desirable to use information from these sources collectively, to perform disambiguation. DCOP provides such framework where an objective function is defined over all the knowledge sources (f_k) as below

$$F = \sum_k z_k \cdot f_k$$

where F denotes the total utility associated with a solution and z_k is the weight given to a knowledge source i.e. information from various sources

can be weighted. (Note: It is desirable to normalize utility functions of different knowledge sources in order to compare them.)

Every agent (word) choose its value (sense) in a such a way that the objective function (global solution) is maximized. This way an agent is assigned a best value which is the target sense in our case.

4 Modelling information from various knowledge sources

In this section, we discuss the modelling of information from various knowledge sources.

4.1 Part-of-speech (POS)

Consider the word *play*. It has 47 senses out of which only 17 senses correspond to *noun* category. Based on the POS information of a word w_i, its domain D_{w_i} is restricted accordingly.

4.2 Morphology

Noun *orange* has at least two senses, one corresponding to *a color* and other to *a fruit*. But plural form of this word *oranges* can only be used in the *fruit* sense. Depending upon the morphological information of a word w_i, its domain D_{w_i} can be restricted.

4.3 Domain information

In the sports domain, *cricket* likely refers to *a game* than *an insect*. Such information can be captured using a unary utility function defined for every word. If the sense distributions of a word w_i are known, a function $f : D_{w_i} \rightarrow \Re$ is defined which return higher utility for the senses favoured by the domain than to the other senses.

4.4 Sense Relatedness

Sense relatedness between senses of two words w_i, w_j is captured by a function $f : D_{w_i} \times D_{w_j} \rightarrow \Re$ where f returns sense relatedness (utility) between senses based on sense taxonomy and gloss overlaps.

4.5 Discourse

Discourse constraints can be modelled using a n-ary function. For instance, to the extent one sense per discourse (Gale et al., 1992) holds true, higher utility can be returned to the solutions which favour same sense to all the occurrences of a word in a given discourse. This information can be modeled as follows: If $w_i, w_j, \ldots w_m$ are the occurrences of a same word, a function $f : D_i \times D_j \times \ldots D_m \rightarrow \Re$ is defined which returns higher utility when $s_{w_i} = s_{w_j} = \ldots s_{w_m}$ and for the rest of the combinations it returns lower utility.

4.6 Collocations

Collocations of a word are known to provide strong evidence for identifying correct sense of the word. For example: if in a given context *bank* co-occur with *money*, it is likely that *bank* refers to *financial institution* sense rather than *the edge of a river* sense. The word *cancer* has at least two senses, one corresponding to the astrological sign and the other a disease. But its derived form *cancerous* can only be used in disease sense. When the words *cancer* and *cancerous* co-occur in a discourse, it is likely that the word *cancer* refers to *disease sense*.

Most supervised systems work through collocations to identify correct sense of a word. If a word w_i co-occurs with its collocate *v*, collocational information from v can be modeled by using the following function

$$coll\_infrm\_v_{w_i} : D_{w_i} \rightarrow \Re$$

where $coll\_infrm\_v_{w_i}$ returns high utility to collocationally preferred senses of w_i than other senses.

Collocations can also be modeled by assigning more than one variable to the agents or by adding a dummy agent which gives collocational information but in view of simplicity we do not go into those details.

Topical word associations, semantic word associations, selectional preferences can also be modeled similar to collocations. Complex information involving more than two entities can be modelled by using n-ary utility functions.

5 Experiment: DCOP based All Words WSD

We carried out a simple experiment to test the effectiveness of DCOP algorithm. We conducted our experiment in an all words setting and used only WordNet (Fellbaum, 1998) based relatedness measures as knowledge source so that results can be compared with earlier state-of-art knowledge-based WSD systems like (Agirre and Soroa, 2009; Sinha and Mihalcea, 2007) which used similar knowledge sources as ours.

Our method performs disambiguation on sentence by sentence basis. A utility function based on semantic relatedness is defined for every pair of words falling in a particular window size. Restricting utility functions to a window size reduces the number of constraints. An objective function is defined as sum of these restricted utility functions over the entire sentence and thus allowing information flow across all the words. Hence, a DCOP algorithm which aims to maximize this objective function leads to a globally optimal solution.

In our experiments, we used the best similarity measure settings of (Sinha and Mihalcea, 2007) which is a sum of normalized similarity measures jcn, lch and lesk. We used used Distributed Pseudotree Optimization Procedure (DPOP) algorithm (Petcu and Faltings, 2005), which solves DCOP using linear number of messages among agents. The implementation provided with the open source toolkit FRODO[1] (Léauté et al., 2009) is used.

5.1 Data

To compare our results, we ran our experiments on SENSEVAL-2 and SENSEVAL -3 English all-words data sets.

5.2 Results

Table 1 shows results of our experiments. All these results are carried out using a window size of four. Ideally, precision and recall values are expected to be equal in our setting. But in certain cases, the tool we used, FRODO, failed to find a solution with the available memory resources.

Results show that our system performs consistently better than (Sinha and Mihalcea, 2007) which uses exactly same knowledge sources as used by us (with an exception of adverbs in Senseval-2). This shows that DCOP algorithm perform better than page-rank algorithm used in their graph based setting. Thus, for knowledge-based WSD, DCOP framework is a potential alternative to graph based models.

Table 1 also shows the system (Agirre and Soroa, 2009), which obtained best results for knowledge based WSD. A direct comparison between this and our system is not quantitative since they used additional knowledge such as extended WordNet relations (Mihalcea and Moldovan, 2001) and sense disambiguated gloss present in WordNet3.0.

Senseval-2 All Words data set					
	noun	verb	adj	adv	all
P_dcop	67.85	37.37	**62.72**	56.87	58.63
R_dcop	66.44	35.47	61.28	56.65	57.09
F_dcop	67.14	36.39	**61.99**	56.76	57.85
P_Sinha07	67.73	36.05	62.21	60.47	**58.83**
R_Sinha07	65.63	32.20	**61.42**	60.23	56.37
F_Sinha07	66.24	34.07	61.81	60.35	57.57
Agirre09	**70.40**	**38.90**	58.30	**70.1**	**58.6**
MFS	71.2	39.0	61.1	75.4	60.1
Senseval-3 All Words data set					
P_dcop	62.31	43.48	57.14	**100**	54.68
R_dcop	60.97	42.81	55.17	**100**	53.51
F_dcop	61.63	43.14	56.14	**100**	54.09
P_Sinha07	61.22	45.18	54.79	**100**	54.86
R_Sinha07	60.45	40.57	54.14	**100**	52.40
F_Sinha07	60.83	42.75	54.46	**100**	53.60
Agirre09	**64.1**	**46.9**	**62.6**	92.9	**57.4**
MFS	69.3	53.6	63.7	92.9	62.3

Table 1: Evaluation results on Senseval-2 and Senseval-3 data-set of all words task.

5.3 Performance analysis

We conducted our experiment on a computer with two 2.94 GHz process and 2 GB memory. Our algorithm just took 5 minutes 31 seconds on Senseval-2 data set, and 5 minutes 19 seconds on Senseval-3 data set. This is a singable reduction compared to execution time of page rank algorithms employed in both Sinha07 and Agirre09. In Agirre09, it falls in the range 30 to 180 minutes on much powerful system with 16 GB memory having four 2.66 GHz processors. On our system, time taken by the page rank algorithm in (Sinha and Mihalcea, 2007) is 11 minutes when executed on Senseval-2 data set.

Since DCOP algorithms are truly distributed in nature the execution times can be further reduced by running them parallely on multiple processors.

6 Related work

Earlier approaches to WSD which encoded information from variety of knowledge sources can be classified as follows:

- Supervised approaches: Most of the supervised systems (Yarowsky and Florian, 2002;

[1] http://liawww.epfl.ch/frodo/

Lee and Ng, 2002; Martínez et al., 2002; Stevenson and Wilks, 2001) rely on the sense tagged data. These are mainly discriminative or aggregative models which essentially pose WSD a classification problem. Discriminative models aim to identify the most informative feature and aggregative models make their decisions by combining all features. They disambiguate word by word and do not collectively disambiguate whole context and thereby do not capture all the relationships (e.g sense relatedness) among all the words. Further, they lack the ability to directly represent constraints like one sense per discourse.

- Graph based approaches: These approaches crucially rely on lexical knowledge base. Graph-based WSD approaches (Agirre and Soroa, 2009; Sinha and Mihalcea, 2007) perform disambiguation over a graph composed of senses (nodes) and relations between pairs of senses (edges). The edge weights encode information from a lexical knowledge base but lack an efficient way of modelling information from other knowledge sources like collocational information, selectional preferences, domain information, discourse. Also, the edges represent binary utility functions defined over two entities which lacks the ability to encode ternary, and in general, any N-ary utility functions.

7 Discussion

This framework provides a convenient way of integrating information from various knowledge sources by defining their utility functions. Information from different knowledge sources can be weighed based on the setting at hand. For example, in a domain specific WSD setting, sense distributions play a crucial role. The utility function corresponding to the sense distributions can be weighed higher in order to take advantage of domain information. Also, different combination of weights can be tried out for a given setting. Thus for a given WSD setting, this framework allows us to find 1) the impact of each knowledge source individually 2) the best combination of knowledge sources.

Limitations of DCOP algorithms: Solving DCOPs is NP-hard. A variety of search algorithms have therefore been developed to solve DCOPs (Mailler and Lesser, 2004; Modi et al., 2004; Petcu and Faltings, 2005) . As the number of constraints or words increase, the search space increases thereby increasing the time and memory bounds to solve them. Also DCOP algorithms exhibit a trade-off between memory used and number of messages communicated between agents. DPOP (Petcu and Faltings, 2005) use linear number of messages but requires exponential memory whereas ADOPT (Modi et al., 2004) exhibits linear memory complexity but exchange exponential number of messages. So it is crucial to choose a suitable algorithm based on the problem at hand.

8 Future Work

In our experiment, we only used relatedness based utility functions derived from WordNet. Effect of other knowledge sources remains to be evaluated individually and in combination. The best possible combination of weights of knowledge sources is yet to be engineered. Which DCOP algorithm performs better WSD and when has to be explored.

9 Conclusion

We initiated a new line of investigation into WSD by modelling it in a distributed constraint optimization framework. We showed that this framework is powerful enough to encode information from various knowledge sources. Our experimental results show that a simple DCOP based model encoding just word similarity constraints performs comparably with the state-of-the-art knowledge based WSD systems.

Acknowledgement

We would like to thank *Prof. Rajeev Sangal* and *Asrar Ahmed* for their support in coming up with this work.

References

Eneko Agirre and David Martínez. 2001. Knowledge sources for word sense disambiguation. In *Text, Speech and Dialogue, 4th International Conference, TSD 2001, Zelezna Ruda, Czech Republic, September 11-13, 2001*, Lecture Notes in Computer Science, pages 1–10. Springer.

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41, Morristown, NJ, USA. Association for Computational Linguistics.

Eneko Agirre and Mark Stevenson. 2006. Knowledge sources for wsd. In *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*, pages 217–252. Springer, Dordrecht, The Netherlands.

Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer. 2001. Senseval-2. `http://www.sle.sharp.co.uk/senseval2`.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 233–237, Morristown, NJ, USA. Association for Computational Linguistics.

Graeme Hirst. 1987. *Semantic interpretation and the resolution of ambiguity*. Cambridge University Press, New York, NY, USA.

Thomas Léauté, Brammert Ottens, and Radoslaw Szymanek. 2009. FRODO 2.0: An open-source framework for distributed constraint optimization. In *Proceedings of the IJCAI'09 Distributed Constraint Reasoning Workshop (DCR'09)*, pages 160–164, Pasadena, California, USA, July 13. `http://liawww.epfl.ch/frodo/`.

Yoong Keok Lee and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 41–48, Morristown, NJ, USA. Association for Computational Linguistics.

Roger Mailler and Victor Lesser. 2004. Solving distributed constraint optimization problems using cooperative mediation. In *AAMAS '04: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 438–445, Washington, DC, USA. IEEE Computer Society.

David Martínez, Eneko Agirre, and Lluís Màrquez. 2002. Syntactic features for high precision word sense disambiguation. In *COLING*.

Susan W. McRoy. 1992. Using multiple knowledge sources for word sense discrimination. *COMPUTATIONAL LINGUISTICS*, 18:1–30.

Rada Mihalcea and Phil Edmonds, editors. 2004. *Proceedings Senseval-3 3rd International Workshop on Evaluating Word Sense Disambiguation Systems*. ACL, Barcelona, Spain.

Rada Mihalcea and Dan I. Moldovan. 2001. extended wordnet: progress report. In *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100.

Pragnesh Jay Modi, Wei-Min Shen, Milind Tambe, and Makoto Yokoo. 2004. Adopt: Asynchronous distributed constraint optimization with quality guarantees. *Artificial Intelligence*, 161:149–180.

Pragnesh Jay Modi. 2003. Distributed constraint optimization for multiagent systems. *PhD Thesis*.

Adrian Petcu and Boi Faltings. 2005. A scalable method for multiagent constraint optimization. In *IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence*, pages 266–271, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ravi Sinha and Rada Mihalcea. 2007. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 363–369, Washington, DC, USA. IEEE Computer Society.

Mark Stevenson and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Comput. Linguist.*, 27(3):321–349.

David Yarowsky and Radu Florian. 2002. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8:2002.