Neighbors Help: Bilingual Unsupervised WSD Using Context

Sudha Bhingardive Samiulla Shaikh Pushpak Bhattacharyya
Department of Computer Science and Engineering,
IIT Bombay, Powai,
Mumbai, 400076.
{sudha,samiulla,pb}@cse.iitb.ac.in

Abstract

Word Sense Disambiguation (WSD) is one of the toughest problems in NLP, and in WSD, verb disambiguation has proved to be extremely difficult, because of high degree of polysemy, too fine grained senses, absence of deep verb hierarchy and low inter annotator agreement in verb sense annotation. Unsupervised WSD has received widespread attention, but has performed poorly, specially on verbs. Recently an unsupervised bilingual EM based algorithm has been proposed, which makes use only of the raw counts of the translations in comparable corpora (Marathi and Hindi). But the performance of this approach is poor on verbs with accuracy level at 25-38%. We suggest a modification to this mentioned formulation, using context and semantic relatedness of neighboring words. An improvement of 17% - 35% in the accuracy of verb WSD is obtained compared to the existing EM based approach. On a general note, the work can be looked upon as contributing to the framework of unsupervised WSD through context aware expectation maximization.

1 Introduction

The importance of unsupervised approaches in WSD is well known, because they do not need sense tagged corpus. In multilingual unsupervised scenario, either comparable or parallel corpora have been used by past researchers for disambiguation (Dagan et al., 1991; Diab and Resnik, 2002; Kaji and Morimoto, 2002; Specia et al., 2005; Lefever and Hoste, 2010; Khapra et al., 2011). Recent work by Khapra et al., (2011) has shown that, in comparable corpora, sense distribution of a word in one language can be estimated using the raw counts of translations of the target words in the other language; such sense distributions contribute to the ranking of senses. Since translations can themselves be ambiguous, Expectation Maximization based formulation is used to determine the sense frequencies. Using this approach every instance of a word is tagged with the most probable sense according to the algorithm.

In the above formulation, no importance is given to the context. That would do, had the accuracy of disambiguation on verbs not been poor 25-35%. This motivated us to propose and investigate use of context in the formulation by Khapra et al. (2011).

For example consider the sentence in chemistry domain, *"Keep the beaker on the flat table."* In this sentence, the target word '*table*' will be tagged as 'the tabular array' sense since it is dominant in the chemistry domain by their algorithm. But its actual sense is 'a piece of furniture' which can be captured only if context is taken into consideration. In our approach we tackle this problem by taking into account the words from the context of the target word. We use semantic relatedness between translations of the target word and those of its context words to determine its sense.

Verb disambiguation has proved to be extremely difficult (Jean, 2004), because of high degree of polysemy (Khapra et al., 2010), too fine grained senses, absence of deep verb hierarchy and low inter annotator agreement in verb sense annotation. On the other hand, verb disambiguation is very important for NLP applications like MT and IR. Our approach has shown significant improvement in verb accuracy as compared to Khapra's (2011) approach.

The roadmap of the paper is as follows. Section 2 presents related work. Section 3 covers the background work. Section 4 explains the modified EM formulation using context and semantic relatedness. Section 5 presents the experimental setup.

Results are presented in section 6. Section 7 covers phenomena study and error analysis. Conclusions and future work are given in the last section, section 8.

2 Related work

Word Sense Disambiguation is one of the hardest problems in NLP. Successful supervised WSD approaches (Lee et al., 2004; Ng and Lee, 1996) are restricted to resource rich languages and domains. They are directly dependent on availability of good amount of sense tagged data. Creating such a costly resource for all language-domain pairs is impracticable looking at the amount of time and money required. Hence, unsupervised WSD approaches (Diab and Resnik, 2002; Kaji and Morimoto, 2002; Mihalcea et al., 2004; Jean, 2004; Khapra et al., 2011) attract most of the researchers.

3 Background

Khapra et al. (2011) dealt with bilingual unsupervised WSD. It uses EM algorithm for estimating sense distributions in comparable corpora. Every polysemous word is disambiguated using the raw counts of its translations in different senses. Synset aligned multilingual dictionary (Mohanty et al., 2008) is used for finding its translations. In this dictionary, synsets are linked, and after that the words inside the synsets are also linked. For example, for the concept of *'boy'*, the Hindi synset {*ladakaa, balak, bachhaa*} is linked with the Marathi synset {*mulagaa, poragaa, por*}. The Marathi word '*mulagaa*' is linked to the Hindi word '*ladakaa*' which is its exact lexical substitution.

Suppose words u in language L_1 and v in language L_2 are translations of each other and their senses are required. The EM based formulation is as follows:

E-Step:

$$P(S^{L_1}|u) = \frac{\sum_{v} P(\pi_{L_2}(S^{L_1})|v) \cdot \#(v)}{\sum_{S_i^{L_1}} \sum_{x} P(\pi_{L_2}(S_i^{L_1})|x) \cdot \#(x)}$$

$$where, S_i^{L_1} \in synsets_{L_1}(u)$$
$$v \in crosslinks_{L_2}(u, S^{L_1})$$
$$x \in crosslinks_{L_2}(u, S_i^{L_1})$$

M-Step:

$$P(S^{L_2}|v) = \frac{\sum_{u} P(\pi_{L_1}(S^{L_2})|u) \cdot \#(u)}{\sum_{S_i^{L_2}} \sum_{y} P(\pi_{L_1}(S_i^{L_2})|y) \cdot \#(y)}$$

$$where, S_i^{L_2} \in synsets_{L_2}(v)$$
$$u \in crosslinks_{L_1}(v, S^{L_2})$$
$$y \in crosslinks_{L_1}(v, S_i^{L_2})$$

Here,

- '#' indicates the raw count.
- $crosslinks_{L_1}(a, S^{L_2})$ is the set of possible translations of the word 'a' from language L_1 to L_2 in the sense S^{L_2}.
- $\pi_{L_2}(S^{L_1})$ means the linked synset of the sense S^{L1} in L_2.

E and M steps are symmetric except for the change in language. In both the steps, we estimate sense distribution in one language using raw counts of translations in another language. But this approach has following limitations:

Poor performance on verbs: This approach gives poor performance on verbs (25%-38%). See section 6.

Same sense throughout the corpus: Every occurrence of a word is tagged with the single sense found by the algorithm, throughout the corpus.

Closed loop of translations: This formulation does not work for some common words which have the same translations in all senses. For example, the verb '*karna*' in Hindi has two different senses in the corpus *viz.*, '*to do*' (S_1) and '*to make*' (S_2). In both these senses, it gets translated as '*karne*' in Marathi. The word '*karne*' also back translates to '*karna*' in Hindi through both its senses. In this case, the formulation works out as follows:

The probabilities are initialized uniformly. Hence, $P(S_1|karna) = P(S2|karna) = 0.5$. Now, in first iteration the sense of '*karne*' will be estimated as follows (E-step):

$$P(S_1|karne) = \frac{P(S_1|karna) * \#(karna)}{\#(karna)}$$
$$= 0.5,$$

$$P(S_2|karne) = \frac{P(S_2|karna) * \#(karna)}{\#(karna)}$$
$$= 0.5$$

Similarly, in M-step, we will get $P(S_1|karna) = P(S_2|karna) = 0.5$. Eventually, it will end up with initial probabilities and no strong decision can be made.

To address these problems we have introduced contextual clues in their formulation by using semantic relatedness.

4 Modified Bilingual EM approach

We introduce context in the EM formulation stated above and treat the context as a bag of words. We assume that each word in the context influences the sense of the target word independently. Hence,

$$p(S|w, C) = \prod_{c_i \in C} p(S|w, c_i)$$

where, w is the target word, S is one of the candidate synsets of w, C is the set of words in context (sentence in our case) and c_i is one of the context words.

Suppose we would have sense tagged data, $p(S|w, c)$ could have been computed as:

$$p(S|w, c) = \frac{\#(S, w, c)}{\#(w, c)}$$

But since the sense tagged corpus is not available, we cannot find $\#(S, w, c)$ from the corpus directly. However, we can estimate it using the comparable corpus in other language. Here, we assume that given a word and its context word in language L_1, the sense distribution in L_1 will be same as that in L_2 given the translation of a word and the translation of its context word in L_2. But these translations can be ambiguous, hence we can use Expectation Maximization approach similar to (Khapra et al., 2011) as follows:

E-Step:

$$P(S^{L_1}|u, a) = \frac{\sum_{v,b} P(\pi_{L_2}(S^{L_1})|v, b) \cdot \sigma(v, b)}{\sum_{S_i^{L_1}} \sum_{x,b} P(\pi_{L_2}(S_i^{L_1})|x, b) \cdot \sigma(x, b)}$$

where, $S_i^{L_1} \in synsets_{L_1}(u)$
$a \in context(u)$
$v \in crosslinks_{L_2}(u, S^{L_1})$
$b \in crosslinks_{L_2}(a)$
$x \in crosslinks_{L_2}(u, S_i^{L_1})$

$crosslinks_{L_1}(a)$ is the set of all possible translations of the word 'a' from L_1 to L_2 in all its senses.

$\sigma(v, b)$ is the semantic relatedness between the senses of v and senses of b. Since, v and b go over all possible translations of u and a respectively. $\sigma(v, b)$ has the effect of indirectly capturing the semantic similarity between the senses of u and a. A symetric formulation in the M-step below takes the computation back from language L_2 to language L_1. The semantic relatedness comes as an additional weighing factor, capturing context, in the probablistic score.

M-Step:

$$P(S^{L_2}|v, b) = \frac{\sum_{u,a} P(\pi_{L_1}(S^{L_2})|u, a) \cdot \sigma(u, a)}{\sum_{S_i^{L_2}} \sum_{y,b} P(\pi_{L_1}(S_i^{L_2})|y, a) \cdot \sigma(y, a)}$$

where, $S_i^{L_2} \in synsets_{L_2}(v)$
$b \in context(v)$
$u \in crosslinks_{L_1}(v, S^{L_2})$
$a \in crosslinks_{L_1}(b)$
$y \in crosslinks_{L_1}(v, S_i^{L_2})$

$\sigma(u, a)$ is the semantic relatedness between the senses of u and senses of a and contributes to the score like $\sigma(v, b)$.

Note how the computation moves back and forth between L_1 and L_2 considering translations of both target words and their context words.

In the above formulation, we could have considered the term $\#(word, context\_word)$ (*i.e.*, the co-occurrence count of the translations of the word and the context word) instead of $\sigma(word, context\_word)$. But it is very unlikely that every translation of a word will co-occur with

Algorithm	HIN-HEALTH					MAR-HEALTH				
	NOUN	ADV	ADJ	VERB	Overall	NOUN	ADV	ADJ	VERB	Overall
EM-C	59.82	67.80	56.66	**60.38**	59.63	62.90	62.54	53.63	**52.49**	59.77
EM	60.68	67.48	55.54	**25.29**	58.16	63.88	58.88	55.71	**35.60**	58.03
WFS	53.49	73.24	55.16	**38.64**	54.46	59.35	67.32	38.12	**34.91**	52.57
RB	32.52	45.08	35.42	**17.93**	33.31	33.83	38.76	37.68	**18.49**	32.45

Table 1: Comparison(F-Score) of EM-C and EM for Health domain

Algorithm	HIN-TOURISM					MAR-TOURISM				
	NOUN	ADV	ADJ	VERB	Overall	NOUN	ADV	ADJ	VERB	Overall
EM-C	62.78	65.10	54.67	**55.24**	60.70	59.08	63.66	58.02	**55.23**	58.67
EM	61.16	62.31	56.02	**31.85**	57.92	59.66	62.15	58.42	**38.33**	56.90
WFS	63.98	75.94	52.72	**36.29**	60.22	61.95	62.39	48.29	**46.56**	57.47
RB	32.46	42.56	36.35	**18.29**	32.68	33.93	39.30	37.49	**15.99**	32.65

Table 2: Comparison(F-Score) of EM-C and EM for Tourism domain

every translation of its context word considerable number of times. This term may make sense only if we have arbitrarily large comparable corpus in the other language.

4.1 Computation of semantic relatedness

The semantic relatedness is computed by taking the inverse of the length of the shortest path among two senses in the wordnet graph (Pedersen et al., 2005). All the semantic relations (including cross-part-of-speech links) *viz.*, hypernymy, hyponymy, meronymy, entailment, attribute *etc.*, are used for computing the semantic relatedness.

Sense scores thus obtained are used to disambiguate all words in the corpus. We consider all the content words from the context for disambiguation of a word. The winner sense is the one with the highest probability.

5 Experimental setup

We have used freely available in-domain comparable corpora[1] in Hindi and Marathi languages. These corpora are available for health and tourism domains. The dataset is same as that used in (Khapra et al., 2011) in order to compare the performance.

6 Results

Table 1 and Table 2 compare the performance of the following two approaches:

1. **EM-C** (EM with Context): Our modified approach explained in section 4.
2. **EM**: Basic EM based approach by Khapra et al., (2011).
3. **WFS**: Wordnet First Sense baseline.
4. **RB**: Random baseline.

Results clearly show that EM-C outperforms EM especially in case of verbs in all language-domain pairs. In health domain, verb accuracy is increased by 35% for Hindi and 17% for Marathi, while in tourism domain, it is increased by 23% for Hindi and 17% for Marathi. The overall accuracy is increased by (1.8-2.8%) for health domain and (1.5-1.7%) for tourism domain. Since there are less number of verbs, the improved accuracy is not directly reflected in the overall performance.

7 Error analysis and phenomena study

Our approach tags all the instances of a word depending on its context as apposed to basic EM approach. For example, consider the following sentence from the tourism domain:

वह पत्ते खेल रहे थे।
(vaha patte khel rahe the)
(They were playing cards/leaves)

Here, the word पत्ते (plural form of पत्ता) has two senses *viz.*, *'leaf'* and *'playing_card'*. In tourism domain, the *'leaf'* sense is more dominant. Hence, basic EM will tag पत्ते with *'leaf'* sense. But it's true sense is *'playing_card'*. The true sense is captured only if context is considered. Here, the word खेलना (to play) (root form of खेल) endorses the *'playing_card'* sense of the word पत्ता. This phenomenon is captured by our approach through semantic relatedness.

But there are certain cases where our algorithm fails. For example, consider the following sentence:

[1] http://www.cfilt.iitb.ac.in/wsd/annotated_corpus/

वह पेड के निचे पत्ते खेल रहे थे।
(vaha ped ke niche patte khel rahe the)
(They were playing cards/leaves below the tree)

Here, two strong context words पेड (tree) and खेल (play) are influencing the sense of the word पत्ते. Semantic relatedness between पेड (tree) and पत्ता (leaf) is more than that of खेल (play) and पत्ता (playing_card). Hence, the '*leaf* sense' is assigned to पत्ता.

This problem occurred because we considered the context as a bag of words. This problem can be solved by considering the semantic structure of the sentence. In this example, the word पत्ता (leaf/playing_card) is the subject of the verb खेलना (to play) while पेड (tree) is not even in the same clause with पत्ता (leaf/playing_cards). Thus we could consider खेलना (to play) as the stronger clue for its disambiguation.

8 Conclusion and Future Work

We have presented a context aware EM formulation building on the framework of Khapra et al (2011). Our formulation solves the problems of "inhibited progress due to lack of translation diversity" and "uniform sense assignment, irrespective of context" that the previous EM based formulation of Khapra et al. suffers from. More importantly our accuracy on verbs is much higher and more than the state of the art, to the best of our knowledge. Improving the performance on other parts of speech is the primary future work. Future directions also point to usage of semantic role clues, investigation of familialy apart pair of languages and effect of variation of measures of semantic relatedness.

References

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In Douglas E. Appelt, editor, *ACL*, pages 130–137. ACL.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Morristown, NJ, USA. Association for Computational Linguistics.

Véronis Jean. 2004. Hyperlex: Lexical cartography for information retrieval. In *Computer Speech and Language*, pages 18(3):223–252.

Hiroyuki Kaji and Yasutsugu Morimoto. 2002. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mitesh M. Khapra, Anup Kulkarni, Saurabh Sohoney, and Pushpak Bhattacharyya. 2010. All words domain adapted wsd: Finding a middle ground between supervision and unsupervision. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, *ACL*, pages 1532–1541. The Association for Computer Linguistics.

Mitesh M Khapra, Salil Joshi, and Pushpak Bhattacharyya. 2011. It takes two to tango: A bilingual unsupervised approach for estimating sense distributions using expectation maximization. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 695–704, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

K. Yoong Lee, Hwee T. Ng, and Tee K. Chia. 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 137–140.

Els Lefever and Veronique Hoste. 2010. Semeval-2010 task 3: cross-lingual word sense disambiguation. In Katrin Erk and Carlo Strapparava, editors, *SemEval 2010 : 5th International workshop on Semantic Evaluation : proceedings of the workshop*, pages 15–20. ACL.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In *COLING*.

Rajat Mohanty, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra, and Aditya Sharma. 2008. Synset based multilingual dictionary: Insights, applications and challenges. In *Global Wordnet Conference*.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47, Morristown, NJ, USA. ACL.

T. Pedersen, S. Banerjee, and S. Patwardhan. 2005. Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. Research Report UMSI 2005/25, University of Minnesota Supercomputing Institute, March.

Lucia Specia, Maria Das Graças, Volpe Nunes, and Mark Stevenson. 2005. Exploiting parallel texts to produce a multilingual sense tagged corpus for word sense disambiguation. In *In Proceedings of RANLP-05, Borovets*, pages 525–531.