Do not do processing, when you can look up: Towards a Discrimination Net for WSD

**Diptesh Kanojia, Pushpak Bhattacharyya, Raj Dabre, Siddhartha Gunti
& Manish Shrivastava**
IIT Bombay
dipteshkanojia@gmail.com, pb@cse.iitb.ac.in, prajdabre@gmail.com,
siddhartha.gunti191@gmail.com, mani.shrivastava@gmail.com

Abstract

The task of Word Sense Disambiguation (WSD) incorporates in its definition the role of 'context'. We present our work on the development of a tool which allows for automatic acquisition and ranking of 'context clues' for WSD. These clue words are extracted from the contexts of words appearing in a large monolingual corpus. These mined collection of contextual clues form a *discrimination net* in the sense that for targeted WSD, navigation of the net leads to the correct sense of a word given its context. Utilizing this resource we intend to develop efficient and light weight WSD based on look up and navigation of memory-resident knowledge base, thereby avoiding heavy computation which often prevents incorporation of any serious WSD in MT and search. The need for large quantities of sense marked data too can be reduced.

1 Introduction

Word Sense Disambiguation (WSD) is formally defined as the task of computationally identifying senses of a word in a context. Chatterjee et al. (2011) showed that contextual evidence is the predominant parameter for human (and hence machine) sense disambiguation process.

Joshi et al. (2013) had conducted experiments on eye tracking for sense disambiguation in which they studied the cognitive aspects of human sense disambiguation. They demonstrated that annotators do not focus on sentential structure but look for specific words that help identify the domain of the word and narrow down the number of senses.

Kanojia et al. (2012) had developed a basic WordNet navigation and clue selection tool, "Sense Discrimination Tool", which we have studied and improved upon. We realized that this tool can be improved to include many useful functionalities, the most important being automated clue word acquisition using word context (see section 2) and clue ranking based on the relative importance of a clue word. Thus, to utilize context efficiently we have developed a tool which can help mark clues for each word sense along with providing weights indicating their importance. It can also automatically generate clue word suggestions from large monolingual corpus; leading to the development of a new resource for context based WSD. This tool will later evolve into a memory resident knowledge base whose look up and navigation can perform high quality, light weight WSD. This would avoid the need for sense marked data which it is expensive to create. Such a static WSD system will essentially amount to look up and navigation to discriminate amongst word senses, thereby avoiding expensive computation.

2 Clue Marker Tool[1]

"Sense Discrimination Tool" developed by Kanojia et al. (2012) provided simple functionality of allowing lexicographers to traverse WordNet senses and annotate them with clues which were added manually during this process.

The Clue Marker Tool which we present here has embedded within it a number of functionalities which transcend beyond mere marking words with clues. It is language independent and we plan to expand it to many other languages later. For now we describe our work on Hindi. Refer to snapshots attached for each subsection. The tool allows for the following actions:

---

[1] http://www.cfilt.iitb.ac.in/~diptesh/admin/login.php

2.1 Centralized User Management

In order to track what work was done by which lexicographer we created a registration/login mechanism. This ensures that no one can tamper with the data and also determines how much work was done by a particular person. After the first registration the request is sent to the admin who can regulate the tool usage by the person.

2.2 Phonetic Typing and Devanagari Keyboard

We integrated the Google Transliterate API into our tool which simplifies the task of data entry. For people who find the phonetic typing difficult we have also incorporated a visual Devanagari keyboard.

2.3 WordNet Synsets Navigation

Wordnets have emerged as crucial resources for Natural Language Processing (NLP). They are lexical structures composed of synsets and semantic relations (Fellbaum, 1998). Our tool allows one to navigate through the complete Hindi Word-Net (Dipak Narayan et al., 2002). One can proceed in a sequential manner by viewing previous or next synsets. If one wishes to view any arbitrary synset they can just type its 'id' in a search box and get redirected to it. One can also search for a word and the tool will display all the synsets that contain that word and the user can select any one.

2.4 Add Clues

Synset words, Gloss and Example are possible clue sources. We have provided a mechanism so that if a user selects any text on the page, it can be added to the clues box with a "add"/"add to clues" button. After the lexicographer is sure, she can "submit" the clues to make sure they are finally added to the database. Adding clues only from synset words, gloss or example can be quite restrictive and thus we incorporated a corpus search mechanism known as the concordancer search.

2.5 Concordancer Search

The concordancer is a tool in which, given a corpus and any word to be searched, it returns a set of sentences which contain the word. We provided mechanisms to control the number of sentences to be displayed for lexicographer's convenience. Any word from the sentences returned by the concordancer search results can also be added to the clue word list by the "add to clues" button. The corpus we used, initially, consisted of around 0.22 million sentences from tourism, health and BBC news corpus[2]. We then considered incorporating 0.45 million lines of Wikipedia corpus and 0.97 million lines of crawled news data. Thus we collated a total of approximately 1.4 million lines of monolingual corpus for Hindi.

2.6 Generate Clues automatically

Even with the above concordancer, the lexicographers still have to go through a large number of sentences to decide on the clue words. The primary feature of this tool is being able to generate clues automatically from concordancer sentences. To alleviate this problem we developed a mechanism to automatically generate candidate clue words. The lexicographer can click on the "search for possible clues" button to get a set of words which the tool proposes to be prominent clues. The procedure to generate the clue words is given below:

1. Select N sentences (N=10 for the results reported here) from the concordancer search results by using the first word of the synset as a search term.
2. Run the Hindi part of speech CRF tagger[3] on these sentences.
3. Select the nouns and verbs from the tagged words.
4. Remove stop words, noise and duplicates.

We select nouns and verbs because the lexicographers determined that they are the best candidates for clues. These are, however, not ordered by relative importance, which was the objective of developing the tool. We thus made investigations on the association between the clue words and the synset words leading to some interesting results and insights which are given in the next section.

For each word in the list returned, we calculated a score and sorted the list based on this score. The result is a reordered list of clues presented to the lexicographers who reject the wrong ones. Since the best clues are at the top the lexicographers found their task much simpler than before.

[2] www.cfilt.iitb.ac.in/wsd/annotated_corpus/

[3] http://www.cfilt.iitb.ac.in/tools/POS_tagger.zip

S. No.	Word	Clues
1.	अपराध (aparādha) (crime)	**अपराधी(aparādhi - criminal), दण्ड(daṇḍa - penalty), सजा(sajā - punishment), हत्या(hatyā - killing),साधुजी(sādhuji - sage), चौंका(cauṅkā - surprised)**, बंगले(bangle - bungalow), लौटा(lautā - return),**घटनाक्रम(ghatnākrama - development)**, सोकर(sokar - slept)
2.	पुष्पित (puṣpita) (flowering)	**आनंद(ānanda - joy), वनस्पति(vanaspati - flora), स्पर्श(sparśa - touch)**, स्थिरता(sthiratā - stability), सखी(sakhī - girlfriend), **सम्पर्क(samparka - contact)**, शांति(śānti – silence, peace), **पवन(pavana - wind), समन्वित (samanvita - incorporated)**
3.	अनाथ (anātha) (orphan)	**अनाथों (anātho - orphans), अनाथालय(anāthālaya - orphanage), मां-बाप(maa-baap - parents)**, बताती(batāti - inform), मारती(mārti – to hit), चलाना(calānā – to operate), मैनेजर(mainējara - manager), **असहाय(asahāya - helpless), खोकर(khokar - lose)**
4.	अपमान (apamāna) (insult, affront)	**जनक(janak - originator), सहन(sahan – to endure), मरना(marnā – to die), समझ(samajh - understanding), कहे(kahe - said)**, भूखों(bhukho - hungry), परीक्षित(parikshita - tested), सूचनाओं(sucanao - information), **मुँह(muṇh - mouth)**

Table 1: Clues after PMI ranking

3 Clue Words Ranking

We considered a set of 80 synsets and studied them to form an idea of the basis of ranking the clue words. We used Hindi Synsets for our study. For each synset:

1. Generate the set of possible/candidate clue words by corpus searching, POS tagging and filtering as described in section 2.6.
2. For each clue word generate scores
3. Sort list of scored clues in descending order and consider top 10 clues.

Scoring techniques which include the co-occurrence factor between two words seemed intuitive since they would rate the clues statistically. We studied some prominent scoring mechanisms such as contingency table measure and PMI given by Terra et al. (2003) amongst which PMI fared better.

3.1 Pointwise Mutual Information

PMI, a concept from information theory, is indicative of the degree of association between two words, in this case: the current synset member and the potential clue word. The formulae used are:

$$\text{PMI(target, clue word)} = \log_2 \frac{\text{p(target, clue word)}}{\text{p(target)*p(clue word)}}$$
$$\ldots (3.1)$$

$$\text{p(x,y)} = \frac{\#(\text{number of sentences containing x and y})}{\#(\text{number of sentences})}$$
$$\ldots (3.2)$$

$$\text{p(x)} = \frac{\#(\text{number of sentences containing x})}{\#(\text{number of sentences})}$$
$$\ldots (3.3)$$

For words that are independent, then PMI is 0.

3.2 Results with PMI

We present in Table 1 above, four synsets for which there were strong clues after PMI based ranking. The clues in bold are relevant ones. Over the complete set of 80 words studied, an average of 5 relevant clue words occurred in the top 10 after PMI ranking. This situation freed the lexicographers from looking for clue words manually, by reading sentences from the concordancer search.

4 Synset reinforced clue ranking

In PMI based ranking, we would only consider the first word of a synset to retrieve clues which led the tool to produce the same set of clues for all synsets which had this word as the first word. We solved this problem by reinforcing the clues using other members of the given synset. We also use a different metric for clue word selection and ranking.

This modified clue acquisition mechanism, instead of using just the first word of the synset, uses the first three words of the synset. Using more members of the same synset helps in highlighting those clues which are more important for a given synset.

As before, we retrieve the sets of candidate clue words for each of the 3 synset words and then perform further processing. Instead of just top 10 clues we now consider as many as possible to ensure coverage. We find clue word overlaps between the three different sets of clues obtained. Those candidate clues which are present in more than one set are obviously good indicators of sense and are given a higher ranking. This added metric counters polysemy, even when first synset

S. No.	Word senses	Top overlapped clues
1.	जन्मा (janma) (born)	काल(*kaal* - time),मृत्यु(*mrityu* - death),रूप(*roop* - form, shape),आज(*aaj* - to-day),दुनिया(*duniya* - world),युग(*yuga* - era)
	जन्मा (janma) (originate)	प्रयोगशाला(*prayogshalaa* - laboratory), कारण(*kaaran* - reason), अनुसंधान(*anusandhaan* - research), अध्ययन(*adhyyan* - study), भाषा(*bhashaa* - language), तर्क(*tark*a - argument)
2.	आदिवासी (aadivaasi) (tribe)	अभाव(*abhaav* - scarcity),कारण(*kaaran* - reason),प्रदेश(*Pradesh* - territory), शिक्षा(*shiksha* - education), जनजाति(*janjaati* - tribe, folk), भाषांतरण(*bhashaantaran* - translation), विवाद(*vivaada* - debate), अवस्थापन(*avasthaapan* - habitation, abode)
	आदिवासी (aadivaasi) (domicile)	जनसंख्या(*janasankhya* - population),राज्य(*rajya* - state), सीमाओं(*seemaon* - bor-ders), संस्कृति(*sanskriti* - culture), आकलनों(*aakalanon* - estimations)
3.	यूरोपीय,यूरोपी (yuropiya, yuropi) (related to Europe)	संघ(*sangha* - union), रूप(*roop* - form), देशों(*deshon* - countries) शक्ति(*shakti* - power) विश्व(*vishwa* - world)
	यूरोपी,यूरोपीय (yuropi , yuropiya) (European citizen)	भाषा(*bhasha* - language), लोगों(*logon* - people), परिवार(*parivaar* - family)
4.	जल्दी (jaldi) (rapidity)	काम(*kaam* - work), कारण(*kaaran* - reason), लोग(*log* - people), अभिनय(*abhinaya* - acting), विषय(*vishaya* - topic), नुकसान(*nuksaan* - loss)
	जल्दी, सवेरे (jaldi, savere) (early morning)	स्नान(*snaana* - bath), सुबह(*subaha* - morning), दिन(*din* - day), दूध(*doodh* - milk), देर(*der* - delay), व्रत(*vrata* - fast, fasting)

Table 2: Overlapped clues

word is same for different senses, since having clues which are generated from members of the given synset would help greatly in disambiguating using the overlapping clues. Such clue overlaps would be able to help us distinguish between fine grained word senses and eliminate the unrelated sense, thus improving our accuracy.

4.1 Results with reinforcement

Table 2 presents such cases where clue overlaps are able to distinguish specifically between the different senses for the same word.

5 Error Analysis

For every wrong clue generated we studied the sentences from the concordancer which lead to its coming up. We believe that these wrong clues appear due to the following reasons:

5.1.1 Chance co-occurrence

Consider for अनाथ (anātha) (orphan) the clue word मैनेजर (manager). Here अनाथ mostly occurred with अनाथालय (orphanage) (a strong clue) which has an association with मैनेजर; but मैनेजर can occur with any organization like banks, companies and so on. Similarly, Proper nouns can also

occur by chance without giving any information about the senses.

5.1.2 Lack of Context

Retrieval of relevant clue words is greatly affected by the sentences that are chosen to get the context. Currently, we are using 10 sentences from the concordancer output to get a list of potential clues. Using more sentences can help in some cases by providing more relevant clues. We have refrained from increasing this number to avoid runtime computation time. We expect to reduce pre-processing time to enable us to include more sentences.

5.1.3 Absence of word in corpus

The tool cannot provide any clues if the word is not present in the monolingual corpus. This can happen for two reasons: if the word is rare or if the word is not matched by the concordancer due to corpus tokenization errors. We realized that 1.4 million domain specific sentences can be restrictive. We are currently in the process of collecting more, clean and good quality, corpus from the web.

6 Discrimination Net

The tool is expected to produce a structured net (Figure 1) with the synset words (green) connected to the clues (yellow), as neighbors, with weighted edges given by the scoring mechanism, which for now is PMI. Using wordnet sematic relations, relevant clues can be brought closer to the sense that they indicate. This structured net will be further augmented by inclusion of semantic relations from WordNet to result in a *Discrimination Net*. To disambiguate a word using this net, we will calculate a score for all the senses of the word and select the sense with highest score based on its clues.

6.1 Scoring mechanism and sample

The score for a particular possible sense will be progressively calculated by traversing from clue words of the given synset in the net, while moving towards the sense word. We are in the process of developing a more efficient scoring mechanism than PMI which will help us in assigning relevant weightage to edges in the discrimination net and improve the potential clue score.
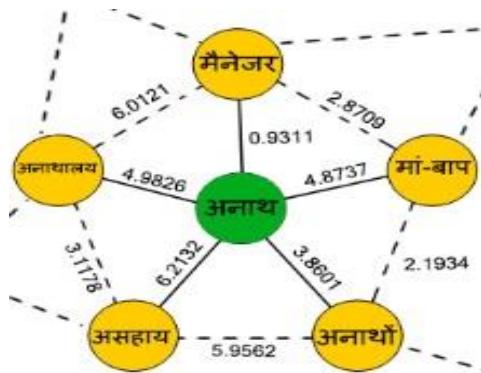
Figure 1: Discrimination Net Sample

7 Conclusions and Future work

We have described the Clue Marker Tool for word senses which allows lexicographers to select relevant clues from a set of ranked candidate clues to disambiguate the sense of the word under consideration. This tool, in addition to being a wordnet browser, is also a corpus browser by way of concordancer based searching. In order to generate high quality clues, we applied PMI based clue ranking and observed its efficacy. The tool is language independent, since by adding synsets of another language to the database and the POS tagger, the clue gathering process can be adapted for the new language. In future we plan to study better measures for clue ranking based on established statistical methods, along with augmenting the corpus to get improvements in generated clues. Finally, we plan to devise efficient and light weight WSD methods that will use the discrimination net, hopefully, bringing about a newer understanding of WSD.

References

Arindam Chatterjee, Salil Joshi, Pushpak Bhattacharyya, Diptesh Kanojia and Akhlesh Meena. 2011. A Study of Human Sense annotation process: Man v/s Machine. *Global WordNet Conference, Matsue, Japan.*

Diptesh Kanojia, Arindam Chatterjee, Salil Joshi and Pushpak Bhattacharyya. 2012. Discrimination Net for Hindi. *COLING, Mumbai, India.*

Salil Joshi, Diptesh Kanojia and Pushpak Bhattacharyya. 2013. More than meets the eye: Study of Human Cognition in Sense Annotation. *NAACL HLT 2013, Atlanta, USA.*

Egidio Terra and C.L.A. Clarke. 2003. Frequency Estimates for Statistical Word Similarity Measures. *NAACL HLT 2003, Edmonton, Canada.*

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. *Cambridge, MA: MIT Press.*

Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande and P. Bhattacharyya. 2002. An Experience in Building the Indo WordNet- a WordNet for Hindi, *International Conference on Global WordNet (GWC 02), Mysore, India.*

Snapshot 1: Clue Marker tool user management

Snapshot 2: Clue Marker tool home / Data entry

Concordancer Hindi Corpus Search

(Text **Transliterates in Hindi** as you type)

Enter the word or phrase: दिल्ली   [Search] Show: 20 results ▼

[Click here for Devanagari Keyboard]

Total **25043** occurences found...

[First] [.....] [Previous] o [Next] [.....] [Last] [Add to clues]

1: मानव धर्म कार्यालय , नयी दिल्ली के श्री दीनानाथ भार्गव दिनेश कृत , भगवद्गीता का हिन्दी पद्यानुवाद , श्री हरिगीता , कहलाता है . उत्तर भारत में लोग गीता के साथ साथ हरिगीता का पाठ भी करते हैं क्योंकि यह लोकभाषा में होने जल्दी समझ आता है और पद्य रूप में होने से आसानी से याद हो जाता है . ।

2: भारत आने के बाद मुसलमानों ने ज़बान-ए-हिन्दी , हिन्दी जुबान अथवा हिन्दी का प्रयोग दिल्ली-आगरा के चारों ओर बोली जाने वाली भाषा के अर्थ में किया ।

3: यह खड़ीबोली पर आधारित है , जो दिल्ली और उसके आस-पास के क्षेत्रों में बोली जाती थी ।

4: हिन्दी और उर्दू दोनों को मिलाकर , बिहार , झारखंड , मध्य प्रदेश , उत्तरांचल , हिमाचल प्रदेश , छत्तीसगढ़ , राजस्थान , हरियाणा , और दिल्ली ।

5: १२वीं शताब्दी के प्रारंभ में , भारत पर भारतीय उपमहाद्वीप का इस्लामिक इतिहास - इस्लामी आक्रमणों के पश्चात , उत्तरी व केन्द्रीय भारत का अधिकांश भाग दिल्ली सल्तनत के शासनाधीन हो गया; और बाद में , अधिकांश उपमहाद्वीप मुगल वंश के अधीन ।

6: राज्यों की चुनी हुई स्वतंत्र सरकारें हैं , जबकि केन्द्रशासित प्रदेशों पर केन्द्र द्वारा नियुक्त प्रबंधन शासन करता है , हालाँकि पॉण्डिचेरी और दिल्ली की लोकतांत्रिक सरकार भी हैं ।

7: भारत के मुख्य शहर हैं - दिल्ली , मुम्बई , कोलकाता , चेन्नई , बंगलोर (बेंगलुरु) ।

8: दो केन्द्र शासित प्रदेशों दिल्ली और पाण्डिचेरी (जो बाद में पुदुचेरी कहा गया) को विधानसभा सदस्यों का अधिकार दिया गया और अब वे छोटे राज्यों के रूप में गिने जाते हैं ।

9: इसके पड़ोसी राज्य हैं उत्तराखंड , हिमाचल प्रदेश , हरियाणा , दिल्ली , राजस्थान , मध्य प्रदेश , छत्तीसगढ़ , झारखंड , बिहार ।

10: यह राज्य उत्तर में नेपाल , तिब्बत दक्षिण में मध्य प्रदेश , पश्चिम में हरियाणा , दिल्ली , राजस्थान तथा पूर्व में बिहार से घिरा है ।

11: लगभग 650 वर्षों तक अधिकांश भारत की तरह उत्तर प्रदेश पर भी किसी न किसी मुस्लिम वंश का शासन रहा , जिनका केन्द्र दिल्ली या उसके आसपास था ।

12: 1526 ई. में बाबर ने दिल्ली के सुल्तान इब्राहीम लोदी को हराया और सर्वाधिक सफल मुस्लिम वंश , मुगल वंश की नींव रखी ।

13: शाहजहाँ ने आगरा व दिल्ली में भी वास्तुशिल्प की दृष्टि से कई महत्त्वपूर्ण इमारतें बनवाईं थीं ।

14: 18वीं शताब्दी में मुगलों के पतन के साथ ही इस मिश्रित संस्कृति का केन्द्र दिल्ली से लखनऊ चला गया , जो अवध के नवाब के अन्तर्गत था और जहाँ साम्प्रदायिक सद्भाव के माहौल में कला , साहित्य , संगीत और काव्य का उत्कर्ष हुआ ।

15: उत्तर प्रदेश के मुख्य नगर वायुमार्ग द्वारा दिल्ली व भारत के अन्य शहरों से जुड़े हुए हैं ।

16: 1312 इसवी में गोआ पहली बार दिल्ली सल्तनत के अधीन हुआ लेकिन उन्हें विजयनगर के शासक हरिहर प्रथम द्वार वहाँ से खदेड़ दिया गया ।

17: अगले सौ सालों तक विजयनगर के शासकों ने यहाँ शासन किया और 1469 में गुलबर्गा के बहामी सुल्तान द्वारा फिर से दिल्ली सल्तनत का हिस्सा बनाया गया ।

18: दिल्ली , जयपुर , शिमला , कुल्लू , कसौली , मनाली , अमृतसर , जालंधर , लुधियाना , हरिद्वार , देहरादून आदि शहरों से यहां के लिए नियमित बस सेवाएं हैं ।

19: इस प्रदेश का दूसरा प्रमुख शहर टाटानगर (जमशेदपुर) दिल्ली कोलकाता मुख्य रेलमार्ग पर बसा हुआ है जो राँची से 120 किलोमीटर दक्षिण में बसा है ।

20: राज्य का एकमात्र राष्ट्रीय हवाई अड्डा राँची का बिरसा मुंडा हवाई-अड्डा है जो देश के प्रमुख शहरों; मुंबई , दिल्ली , कोलकाता और पटना से जुड़ा है ।

Snapshot 3: Clue marker tool concordancer pane

Automated Clue Search Mechanism

139 Possible Clue words:

अज, अजावय:, अर्थ, अवतरण, अवतार, अवधारणा, अविकारी, आएगा, आत्मा, आधार, इच्छा, ईश्वर, उत्तराधिकार, उद्देश्य, उपासना, कथा, कदाचन, कहकर, कहलाता, कामदेव, काया, कारण, कार्य, काल, कोख, क्षेत्रय, खोने, गम, गुणगान, गुणों, चल, चित्रगुप्त, चीज, चेतना, छगल, छाग, जन्म, जन्मता, जयंती, जरथुश्त्र, जानता, जायेहं, जीव, ज्ञान, ताज, दादा, दिखायी, देता, देने, देवता, देश-काल, धर्म, धारण, ध्यान, नंबर, नाम, नामों, निराकार, निर्विकार, न्यायकारी, पति, परमतत्त्व, परमात्म, परमेश्वर, परिवर्तन, पल, पवित्र, पाने, पारसी, पिता, पुंज, पुनर्जन्म, पुरुष, पूषा, प्रतिनिध, प्रत्यक, प्रत्यक्ठों, प्राणियों, प्रिंस, प्रेत, बंधन, बकरे, बच्चा, बताया, बतायें, बसा, बात, बोध, ब्रह्मा, भाषा, भेद, मनाई, मरता, मरते, मानना, मानने, माना, माया, मारा, मारे, मार्गदर्शक, मोहजालों, रूप, लीला, लेता, लेते, लोग, वर्णन, वर्तमान, वस्तु, वासुदेव-अनेक, विशेषण, विश्वपंच, वेदों, शंकर, शरीर, शक्रिया, शिव, संकल्प, संदर्भ, संस्थापक, संहारक, सत्य, सनातन, सर्वभूतानां, सर्वाधार, सर्वेश्वर, सशरीर, साहित्य, सिद्ध, सूनत, सूची, सृष्टिकर्ता, स्वरूप, हरि, हिरण्यगर्भ, होकर, होना, होनेवाला, ,

[Add to Clues]

[Go to top]

[Go to top]

Snapshot 4: Clue marker tool automated clue search

Snapshot 5: Search query box for synset word search