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.

Purpose of the work

- To experiment with automatic extraction of clues which will be useful in context based WSD.
- To study the relevance of using association based method such as PMI for ranking extracted clues.
- To develop a basic framework of the Discrimination Net comprising of the clues.

Roadmap

- Motivation
- Introduction
- Clue Marker Tool
- Automatic clue generation
- Clue ranking
- Synset reinforced clue ranking
- Results
- Error Analysis
- Discrimination Net
- Conclusions & Future work

Motivation

- Current WSD methods do not make use of the context effectively.
- Heavy weight memory utilization due to large size probabilistic models
- Intensive processing because of inference over graphical models.
- Need of a light weight (memory resident) high accuracy WSD mechanism.

Word Sense Disambiguation

- WSD entails computationally finding the sense of a word in a given ***context***.
- *Current WSD models* based on probabilistic methods which require heavy processing.
- Our work is based on the development and use of a tool to mine collection of context clues to form a *Discrimination Net*.

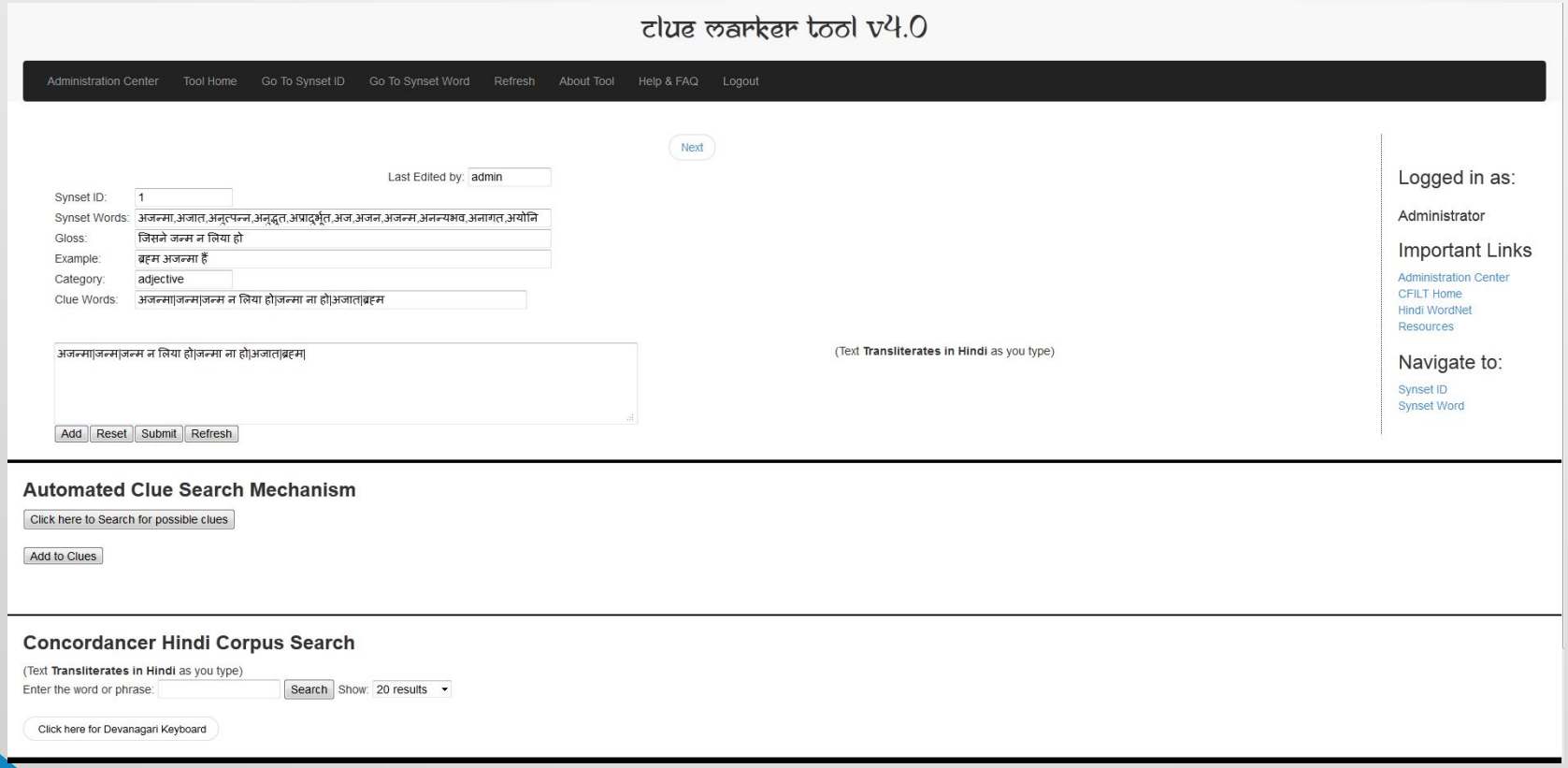
Previous work

- Chatterjee et al. (2011) showed that contextual evidence is predominant parameter for human (and hence machine) sense disambiguation process.
- Joshi et. al. (2013) showed that annotators do not focus on sentential structure but look for specific words, thus helping identify the domain.
- Kanojia et al. (2012) developed a basic tool for Wordnet navigation and manual clue selection by annotators.
- Using insights from these works, the concept of Discrimination Net was born, which started out from the Clue Marker Tool.

Clue Marker Tool

- Clue marker tool (earlier Sense discrimination tool) developed by Kanojia et. al. provided with simple functionality of allowing lexicographers to tag clues to senses.
- Clues were to be added from gloss, example sentences present in the Wordnet database.
- We improvised on it by embedding concordancer and automatic clue mining from concordancer.
- Previous features:
  - Centralized user management system
  - Phonetic typing and Devanagri keyboard.
  - Wordnet navigation
  - Manual clue addition to database.

Clue Marker Tool: Screenshot



Clue Marker Tool: Navigation & Clue Addition

Previous | First Page | Next

Last Edited by: jaya

Synset ID:	4
Synset Words:	पवित्र स्थान,चैत्य स्थान,पूण्य भूमि,पूण्य-स्थल,पूण्य स्थल
Gloss:	वह स्थान जो पवित्र माना जाता हो
Example:	हिंदूओं के लिए काशी एक पवित्र स्थान है
Category:	noun
Clue Words:	पवित्र|शुद्ध|स्थान|पवित्र स्थान|चैत्य स्थान|पूण्य-स्थल|हिंदूओं|काशी|आर्यसमाज|मन्दिर|यज्ञ

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

पवित्र|शुद्ध|स्थान|पवित्र स्थान|चैत्य स्थान|पुण्य-स्थल|हिंदुओं|काशी|आर्यसमाज|मन्दिर|यज्ञ|

Add | Reset | Submit | Refresh

Clue Marker Tool: Automatic Clue Generation

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, क्षेत्रग्य, खोने, गम, गुणगान, गुणों, चल, चित्रगुप्त, चीज, चेतना, छगल, छाग, जन्म, जन्मता, जयंती, ज़रथुश्त्र, 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, प्रत्यक्ष, प्रत्यक्षों, प्राणियों, प्रिंस, प्रेत, बंधन, बकरे, बच्चा, बताया, बताये, बसा, बात, बोध, ब्रम्हा, भाषा, भेड, 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, शवक्रिया, शिव, संकल्प, संदर्भ, संस्थापक, संहारक, सत्य, सनातन, सर्वभूतानां, सर्वाधार, सर्वेश्वर, सशरीर, साहित्य, सिद्ध, सूक्त, सूची, सृष्टिकर्ता, स्वरूप, हरि, हिरण्यगर्भ, होकर, होना, होनेवाला

Add to Clues

Clue Marker Tool: Concordancer

Concordancer Hindi Corpus Search

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

Enter the word or phrase: देवता Search Show: 20 results

Click here for Devanagari Keyboard

Total **1359** occurences found...

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

1: पूजा की रीति इस तरह है : पहले कोई भी देवता चुनें , जिसकी पूजा करनी है ।
2: इनमें एक अज्ञात मातृदेवी की मूर्तियाँ , शिव पशुपति जैसे देवता की मुद्राएँ , लिंग , पीपल की पूजा , इत्यादि प्रमुख हैं ।
3: हिन्दू धर्मग्रन्थ उपनिषदों के अनुसार ब्रह्म ही परम तत्व है (इसे त्रिमूर्ति के देवता ब्रह्मा से भ्रमित न करें) ।
4: ये देवता कौन हैं , इस बारे में तीन मत हो सकते हैं : ।
5: जैसे , कृष्ण को परमेश्वर माना जाता है जिनके अधीन बाकी सभी देवी-देवता हैं , और साथ ही साथ , सभी देवी-देवताओं को कृष्ण का ही रूप माना जाता है ।
6: जो भी सोच हो , ये देवता रंग-बिरंगी हिन्दू संस्कृति के अभिन्न अंग हैं ।
7: बाद के हिन्दू धर्म में नये देवी देवता आये (कई अवतार के रूप में)-- गणेश , राम , कृष्ण , हनुमान , कार्तिकेय , सूर्य-चन्द्र और ग्रह , और देवियाँ (जिनको माता की उपाधि दी जाती है) जैसे-- दुर्गा , पार्वती , लक्ष्मी , शीतला , सीता , राधा , सन्तोषी , काली , इत्यादि ।
8: ये सभी देवता पुराणों मे उल्लिखित हैं , और उनकी कुल संख्या 33 करोड़ बतायी जाती है ।
9: यह माना जाता है कि गाय में सम्पूर्ण ३३ करोड देवि देवता वास करते हैं ।
10: हाँ , यह भी हुआ कि वैदिक ऋषि जहाँ यह पूछ कर शांत हो जाते थे कि यह सृष्टि किसने बनाई है? और कौन देवता है जिसकी हम उपासना करें? वहाँ उपनिषदों के ऋषियों ने सृष्टि बनाने वाले के संबंध में कुछ सिद्धांतों का निश्चय कर दिया और उस सत का भी पता पा लिया जो पूजा और उपासना का वस्तुत: अधिकार है ।
11: प्रकृति के प्रत्येक रूप में एक नियंत्रक देवता की कल्पना करते-करते वैदिक आर्य बहुदेववादी हो गए थे ।
12: अन्य उपनिषद् तत्तद् देवता विषयक होने के कारण तांत्रिक माने जाते हैं ।
13: इनमें एक अज्ञात मातृदेवी की मूर्तियाँ , शिव पशुपति जैसे देवता की मुद्राएँ , लिंग , पीपल की पूजा , इत्यादि प्रमुख हैं ।
14: हिन्दू धर्मग्रन्थ उपनिषदों के अनुसार ब्रह्म ही परम तत्व है (इसे त्रिमूर्ति के देवता ब्रह्मा से भ्रमित न करें) ।
15: ये देवता कौन हैं , इस बारे में तीन मत हो सकते हैं : ।
16: अद्वैत वेदान्त , भगवद गीता , वेद , उपनिषद् , आदि के मुताबिक सभी देवी-देवता एक ही परमेश्वर के विभिन्न रूप हैं (ईश्वर स्वयं ही ब्रह्म का रूप है) ।
17: मीमांसा के अनुसार सभी देवी-देवता स्वतन्त्र सत्ता रखते हैं , और उनके उपर कोई एक ईश्वर नहीं है ।
18: जैसे , कृष्ण को परमेश्वर माना जाता है जिनके अधीन बाकी सभी देवी-देवता हैं , और साथ ही साथ , सभी देवी-देवताओं को कृष्ण का ही रूप माना जाता है ।
19: जो भी सोच हो , ये देवता रंग-बिरंगी हिन्दू संस्कृति के अभिन्न अंग हैं ।
20: बाद के हिन्दू धर्म में नये देवी देवता आये (कई अवतार के रूप में)-- गणेश , राम , कृष्ण , हनुमान , कार्तिकेय , सूर्य-चन्द्र और ग्रह , और देवियाँ (जिनको माता की उपाधि दी जाती है) जैसे-- दुर्गा , पार्वती , लक्ष्मी , शीतला , सीता , राधा , सन्तोषी , काली , इत्यादि ।

Automatic clue extraction

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 on these sentences.
3. Select the nouns and verbs from the tagged words.
    1. Previous works (Chatterjee et al., Joshi et al.) showed that Nouns and Verbs are best indicators of a word sense
    2. This point of view is supported by in house linguists.
4. Remove stop words, noise and duplicates.

Clue word ranking

- Automatically generated clue words may not all be good for future usage in disambiguation.
- Using association based measures (PMI in our case) one can rank extracted clues in order of their importance for disambiguation.
- Algorithm
  - Generate the set of possible/candidate clue words by corpus searching, POS tagging and filtering .
  - For each clue word generate scores. (Formula in next slide)
  - Sort list of scored clues in descending order and consider top 10 clues.

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.

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

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

$$p(x) = \frac{\#(\textit{number of sentences containing x})}{\#(\textit{number of sentences})}$$

Results of PMI ranking

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**(muñh - mouth)

Synset reinforced clue ranking

- In PMI based ranking, Only first word of synset used to retrieve clues.
  - Same set of clues for all synsets with same first synset words were produced.
- Remedied this by considering additional members of each synset (three in our case) and all possible clues (instead of top 10).
- For each synset found intersection of set of clues.
- The common clues are the stronger indicators of word sense.

Results of reinforced clue ranking

S. No.	Word senses	Top overlapped clues
1.	जन्मा (janma) (born)	काल(kaal - time), मृत्यु(mrityu - death), रूप(roop – form, shape), आज(aaj - today), 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(tarka - 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 - borders), 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)

Error Analysis

- Studied sentences from concordancer which led to wrong clues.
- Three main sources of poor clues:
  - **Chance co-occurrence:**
    - अनाथ (anātha) (orphan) has clue word मैनेजर (manager), but is a poor clue because it co-occurs with अनाथालय (orphanage).
  - **Lack of Context:**
    - Utilization of only 10 sentences limits number of clues.
  - **Absence of word in corpus:**
    - Prevents any clues from being generated. Limits us to example and gloss.

Discrimination Net

- Our Discrimination Net is expected to produce a

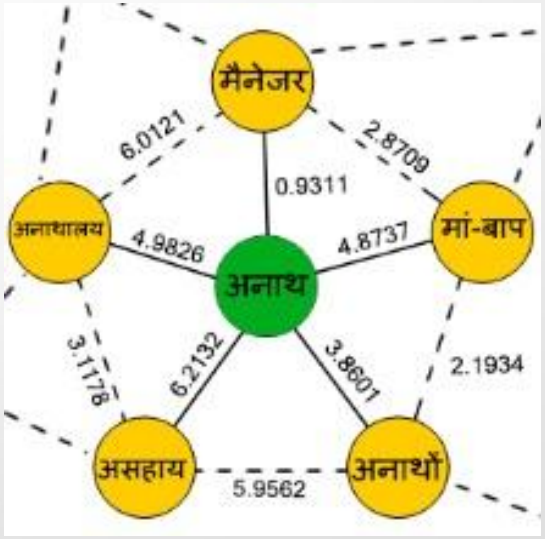
Structured net with synset word (green) connected to Clues (yellow) as neighbors.

- The Weighted edges give the scoring which,

for now, is PMI.

- This structured net will be further augmented

by inclusion of semantic relations from WordNet.



Conclusions

- Clue marker tool is useful in extracting clues which will assist in better utilization of context in word sense disambiguation.
- Automatic clue extraction reduces cognitive load on lexicographers.
- PMI is a reasonable clue ranking mechanism.
- Clue overlap gives stronger and more indicative clues.

Future Work

- Extracted clues in the form of a graph can be memory resident and will be useful in light weight WSD module.
- Accuracy for such a model to be compared with other WSD methods.
- Finally, A framework is to be developed which can Discriminate between fine grained WSD senses using clue words in context.

Resource

- CFILT Resources available at:
  - www.cfilt.iitb.ac.in
- Publications & References:
  - http://www.cse.iitb.ac.in/~pb/pubs-yearwise.html
- Clue Marker Tool:
  - www.cfilt.iitb.ac.in/~diptesh

Thank you! ☺️

Refrences

- Pushpak Bhattacharyya, Arindam Chatterjee, Salil Joshi, Diptesh Kanojia and Akhlesh Meena. 2011. A Study of Human Sense annotation process: Man v/s Machine. *Global WordNet Conference, Matsue, Japan.*
- Pushpak Bhattacharyya, Arindam Chatterjee, Salil Joshi and Diptesh Kanojia. 2012. Discrimination Net for Hindi. *COLING, Mumbai, India*.
- Pushpak Bhattacharyya, Salil Joshi and Diptesh Kanojia. 2013. More than meets the eye: Study of Human Cognition in Sense Annotation. *NAACL HLT 2013, Atlanta, USA.*
- Charles Clarke and Egidio Terra. 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.*
- Pushpak Bhattacharyya, Debasri Chakrabarty, Dipak Narayan and Prabhakar Pande. 2002. An Experience in Building the Indo WordNet- a WordNet for Hindi, *International Conference on Global WordNet (GWC 02), Mysore, India*.