WordNet and Word Sense Disambiguation

Sudha Bhingardive
Department of CSE,
IIT Bombay

Supervisor
Prof. Pushpak Bhattacharyya
WordNet
Outline

• What is WordNet?
• WordNet Synset
• Principles used for Synset Creation
• WordNet Lexico-Semantic Relations
• Important WordNets: English, Hindi, IndoWordNet, BabelNet
• Applications
Outline

• What is WordNet?
• WordNet Synset
• Principles used for Synset Creation
• WordNet Lexico-Semantic Relations
• Important WordNets: English, Hindi, IndoWordNet, BabelNet
• Applications
What is WordNet?

Dictionary = Words + meanings

WordNet = Words + meanings + Semantic and Lexical Relations
What is WordNet? contd..

- A lexical knowledge database for a language
- Consists of synsets and lexico-semantic relations
- Categorizes synsets into four main parts-of-speech categories: nouns, adjectives, adverbs and verbs
- **Monolingual WordNet**
  - English
  - Hindi
  - Sanskrit
- **Multilingual WordNet**
  - IndoWordNet
  - EuroWordNet
  - BabelNet
Outline

• What is WordNet?
• WordNet Synset
• Principles used for Synset Creation
• WordNet Lexico-Semantic Relations
• Important WordNets: English, Hindi, IndoWordNet, BabelNet
• Applications
WordNet Synset

Each synset consist of:
• Sense ID
• Parts-of-speech category
• Synset Members (Synonyms words)
• Gloss or Concept Definition
• Example Sentence

Synset of a boy:
(10305010) (n) male child, boy (a youthful male person) "the baby was a boy"; "she made the boy brush his teeth every night"; "most soldiers are only boys in uniform"
Outline

• What is WordNet?
• WordNet Synset
• Principles used for Synset Creation
• WordNet Lexico-Semantic Relations
• Important WordNets: English, Hindi, IndoWordNet, BabelNet
• Applications
Principles used for Synset Creation

• Minimality
  – The minimal set of words to make the concept unique

• Coverage
  – The maximal set of words ordered by frequency in the corpus to include all possible words standing for the sense.

• Replaceability
  – The example sentence should be such that the most frequent words in the synset can replace one another in the sentence without altering the sense.

Sysnet of bank:
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"
Outline

• What is WordNet?
• WordNet Synset
• Principles used for Synset Creation
• WordNet Lexico-Semantic Relations
• Important WordNets: English, Hindi, IndoWordNet, BabelNet
• Applications
WordNet Lexico–Semantic Relations

- Synonymy
- Antonymy
- Gradation
- Hypernymy / Hyponymy
- Meronymy / Holonymy
- Entailment
- Attribute
- Nominalization
- Ability Link
- Capability Link
- Function Link
Lexical Relations

• Relation between words

• **Synonymy**: relationship between words in a synset.
  – \{plant, flora\}, ‘plant’ and ‘flora’ are related through synonymy relation.

• **Antonymy**: relationship between words having an opposite meaning.
  – ‘day’ and ‘night’ are antonyms of each other.

• **Gradation**:
  – ‘morning’, ‘afternoon’, ‘evening’ are related through gradation relation
Semantic Relations

• Relation between synsets

• **Hypernymy / Hyponymy**: is-a-kind-of relation
  – ‘fruit’ is a hypernym of ‘mango’ and ‘mango’ is a hyponym of ‘fruit’.

• **Meronymy / Holonymy**: part-whole relation
  – ‘hand’ is a meronym of ‘body’ and ‘body’ is a holonym of ‘hand’
Semantic Relations contd..

- **Entailment:**
  - ‘snore’ entails ‘sleep’

- **Attribute:** relationship between noun and adjective synsets
  - ‘hot’ is a value of or attribute of ‘temperature’

- **Nominalization:** relationship between noun and verb synsets
  - ‘service’ nominalizes the verb ‘serve’
Semantic Relations contd..

- **Ability Link**: specifies the inherited features of a nominal concept
  - ‘animal’ and ‘walk’, ‘fish’ and ‘swim’

- **Capability Link**: relationship specifies the acquired features of a nominal concept
  - ‘person’ and ‘swim’

- **Function Link**: relationship specifies the function of a nominal concept
  - ‘vehicle’ and ‘move’ and ‘teacher’ and ‘teach’
WordNet Lexico–Semantic Relations

- boy
  - boyhood
  - schoolboy
  - scout
- male, male person
- girl
- Hypernymy
- Antonymy
- Derivationally related form
- Hyponymy
Outline

• What is WordNet?
• WordNet Synset
• Principles used for Synset Creation
• WordNet Lexico-Semantic Relations
• Important WordNets: English, Hindi, IndoWordNet, BabelNet
• Applications
Some important wordnets

- **English WordNet** (Fellbaum, 1998):
  - First semantic net created at Princeton University
- **Hindi WordNet** (Narayan et. al, 2002)
  - First Indian language Wordnet which is created from English WordNet using expansion approach at IIT Bombay
- **IndoWordnet** (Bhattacharyya, 2010)
  - A Multilingual Wordnet for 17 Indian Languages
- **BabelNet** (Navigli, 2010)
  - A very large, wide coverage multilingual semantic network
  - 271 languages, 14 million synsets, and about 745 million word senses
  - Obtained by automatic integration of Wikipedia (encyclopedic) and WordNet (lexicographic)
English WordNet Interface

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: boy

Display Options: (Select option to change)  Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) male child, boy (a youthful male person) "the baby was a boy"; "she made the boy brush his teeth every night"; "most soldiers are only boys in uniform"
- S: (n) boy (a friendly informal reference to a grown man) "he likes to play golf with the boys"
- S: (n) son, boy (a male human offspring) "their son became a famous judge"; "his boy is taller than he is"

http://wordnetweb.princeton.edu/perl/webwn
WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: boy

Display Options: (Select option to change) - Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) male child, boy (a youthful male person) "the baby was a boy"; "she made the boy brush his teeth every night"; "most soldiers are only boys in uniform"
  - direct hyponym / full hyponym
  - direct hypernym / inherited hypernym / sister term
    - S: (n) male, male person (a person who belongs to the sex that cannot have babies)
  - antonym
    - W: (n) female child [Opposed to: male child] (a youthful female person) "the baby was a girl"; "the girls were just learning to ride a tricycle"
    - W: (n) girl [Opposed to: boy] (a youthful female person) "the baby was a girl"; "the girls were just learning to ride a tricycle"
Hindi WordNet Interface

http://www.cfilt.iitb.ac.in/wordnet/webhwn/
Hindi WordNet Interface contd..

http://www.cfilt.iitb.ac.in/wordnet/webhwn/
Hindi WordNet Structure

Synonyms: आम, रसाल, आम, अंब, अम्ब, प्रियांबु, केशवायुध, कामायुध, कामशर्मा

Gloss: एक फल जो खाया या चूसा जाता है

Example Sentence: तोता पेड़ पर बैठकर आम खा रहा है

Synset

Is-a-kind-of relation (Hyponym)

Is-a-kind-of (Hyponymy)

Is-a-part-of (Meronymy)

Is-a-part-of (Holonymy)
Hindi WordNet Mobile App

The Hindi WordNet is a system for bringing together different lexical (Word based) and semantic (Meaning or Sense based) relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Hindi WordNet is inspired by the famous English WordNet.

You can browse the Hindi WordNet through this interface. It will require Internet to work.

For Querying in Hindi, you can either:
1. Type your query with your mobile keyboard e.g. Type 'साम', press Enter to Transliterate to 'साम' in Hindi, and then click on Search.

The WordNet is a lexical database that is a combination of several components, including:

- **Concepts**: The basic units of information in the WordNet are concepts, which are organized into hierarchical, taxonomic structures.
- **Synsets**: Each concept is represented by a synset, which is a set of synonyms or closely related words.
- **Hypernyms**: More general concepts are represented by hypernyms, which are above synsets in the hierarchy.
- **Hyponyms**: More specific concepts are represented by hyponyms, which are below synsets in the hierarchy.
- **Synset relations**: Relations between synsets include synonymy, hypernymy, hyponymy, meronymy, and holonymy.

The WordNet is a tool for understanding the relationships between words and their meanings, which can be very useful for tasks such as natural language processing and machine translation.

For more information, visit the official WordNet website: https://wordnet.princeton.edu/wordnet
IndoWordNet

http://www.cfilt.iitb.ac.in/indowordnet/
IndoWordNet contd..

IndoWordNet
Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Manipuri, Malayalam, Marathi, Nepali, Odia, Punjabi, Sanskrit, Tamil, Telugu, Urdu

Dravidian Wordnet
Kannada, Malayalam, Tamil, Telugu

North-East Wordnet
Assamese, Bodo, Manipuri, Nepali

Other Wordnets
Hindi, Marathi, Sanskrit

Indradhanush Wordnet
Bengali, Gujarati, Kashmiri, Konkani, Odia, Punjabi, Urdu
Institutes involved in creating IndoWordNet

- Indian Institute of Technology, Bombay – Hindi, Marathi, Sanskrit
- Goa University, Goa – Konkani
- Gauhati University, Guwahati – Assamese, Bodo
- University of Hyderabad, Hyderabad – Odia
- Jawaharlal Nehru University, New Delhi – Urdu
- Dharmsinh Desai University, Nadiad – Gujarati
- University of Kashmir, Srinagar – Kashmiri
- Punjabi University, Patiala – Punjabi
- Thapar University, Patiala – Punjabi
- Manipur University, Imphal – Manipuri
- Assam University, Silchar – Nepali
- Amrita Vishwa Vidyapeetham, Coimbatore – Malayalam
- University of Mysore, Mysore – Kannada
- Tamil University, Tanjavur – Tamil
- Dravidian University, Kuppam – Telugu
IndoWordNet linked Synset

(4265) (n)

ছেলে, বালক

কম বয়সের পুরুষ,
বিশেষত অবিবাহিত

"ময়দানে ছেলেরা
ক্রিকেট খেলছে"  

Bengali
WordNet

(4265) (n)

लड़का, बालक, बाल, बच्चा, छोकड़ा, छोरा, छोकरा

कम उम्र का पुरुष,
विशेषकर अविवाहित

"मैदान में लड़के क्रिकेट
खेल रहे हैं।"

Hindi
WordNet

(4265) (n)

मुलगा, पोरगा, पोर, पोरगे

साधारणतः सोला
वर्षाखालील पुरुष

"तो मुलगा खुपच हुशार
आहे।"

Marathi
WordNet
## IndoWordNet Synset Statistics

<table>
<thead>
<tr>
<th>Language</th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>Adverb</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>29664</td>
<td>3626</td>
<td>6313</td>
<td>534</td>
<td>40137</td>
</tr>
<tr>
<td>Assamese</td>
<td>9065</td>
<td>1676</td>
<td>3805</td>
<td>412</td>
<td>14958</td>
</tr>
<tr>
<td>Bengali</td>
<td>27281</td>
<td>2804</td>
<td>5815</td>
<td>445</td>
<td>36346</td>
</tr>
<tr>
<td>Bodo</td>
<td>8788</td>
<td>2296</td>
<td>4287</td>
<td>414</td>
<td>15785</td>
</tr>
<tr>
<td>Gujarati</td>
<td>26503</td>
<td>2805</td>
<td>5828</td>
<td>445</td>
<td>35599</td>
</tr>
<tr>
<td>Kannada</td>
<td>12765</td>
<td>3119</td>
<td>5988</td>
<td>170</td>
<td>22042</td>
</tr>
<tr>
<td>Kashmiri</td>
<td>21041</td>
<td>2660</td>
<td>5365</td>
<td>400</td>
<td>29469</td>
</tr>
<tr>
<td>Konkani</td>
<td>23144</td>
<td>3000</td>
<td>5744</td>
<td>482</td>
<td>32370</td>
</tr>
<tr>
<td>Malayalam</td>
<td>20071</td>
<td>3311</td>
<td>6257</td>
<td>501</td>
<td>30140</td>
</tr>
<tr>
<td>Manipuri</td>
<td>10156</td>
<td>2021</td>
<td>3806</td>
<td>332</td>
<td>16351</td>
</tr>
<tr>
<td>Marathi</td>
<td>23271</td>
<td>3146</td>
<td>5269</td>
<td>539</td>
<td>32226</td>
</tr>
<tr>
<td>Nepali</td>
<td>6748</td>
<td>1477</td>
<td>3227</td>
<td>261</td>
<td>11713</td>
</tr>
<tr>
<td>Odiya</td>
<td>27216</td>
<td>2418</td>
<td>5273</td>
<td>377</td>
<td>35284</td>
</tr>
<tr>
<td>Punjabi</td>
<td>23255</td>
<td>2836</td>
<td>5830</td>
<td>443</td>
<td>32364</td>
</tr>
<tr>
<td>Sanskrit</td>
<td>31476</td>
<td>1247</td>
<td>4004</td>
<td>265</td>
<td>36997</td>
</tr>
<tr>
<td>Tamil</td>
<td>16312</td>
<td>2803</td>
<td>5827</td>
<td>477</td>
<td>25419</td>
</tr>
<tr>
<td>Telugu</td>
<td>12078</td>
<td>2795</td>
<td>5776</td>
<td>442</td>
<td>21091</td>
</tr>
<tr>
<td>Urdu</td>
<td>22990</td>
<td>2801</td>
<td>5786</td>
<td>443</td>
<td>34280</td>
</tr>
</tbody>
</table>
IndoWordNet Visualizer Interface

IndoWordNet Visualizer

<table>
<thead>
<tr>
<th>Sense ID</th>
<th>PoS</th>
<th>Meaning</th>
<th>Example</th>
<th>Synset</th>
</tr>
</thead>
<tbody>
<tr>
<td>3373</td>
<td>NOUN</td>
<td>नर संतान</td>
<td>“कृष्ण दुसुधेव के पुत्र थे। पुत्र कुमार हो सकता है लेकिन भावा कुमारता नहीं हो सकती”</td>
<td>पुत्र, बेटा, लड़का, लड़की, सूत, बच्चा, सूत, नन्दन, नन्दन, पूरा, पति, तनुजा, आलम, आलमाल, तनुज, लड़का, कुमार, गिरीरीछा, गिरेश, किशोर, हुय, बच्चा, नौकर, मोठा, तनुजा, लड़का, पति, दुभादम, तनुजा, तोइ, तनुजा, भरकंड, फरसंद, फरसंद, आलम, आलमाल, आलम, आलम-संबंध, आलम-संबंध, आलम-संबंध, आलम-संबंध, आलम-संबंध, आलम-संबंध, आलम-संबंध, आलम-संबंध, आलम-संबंध, आलम-संबंध, आलम-संबंध</td>
</tr>
<tr>
<td>5896</td>
<td>NOUN</td>
<td>यह छोटी अपस्सा का पुरुष जो नींदक का काम करे</td>
<td>“दुकानदार ने लड़के से बार-बार बातें की”</td>
<td>लड़का, छोकड़ा, छोकरा</td>
</tr>
<tr>
<td>4265</td>
<td>NOUN</td>
<td>क्रम ऊपर का मुरुग विद्येशक अविष्कृत</td>
<td>“गैरान में लड़के क्रिकेट होता रहता है”</td>
<td>लड़का, बालक, बच्चा, छोकड़ा, छोरा, छोकरा, टौंडा, डुंडा, पूजुक्त, टिशिया, डुंड, बटुक, दहर</td>
</tr>
</tbody>
</table>

http://www.cfilt.iitb.ac.in/Drawgraph/input.html
IndoWordNet Visualizer contd..

http://www.cfilt.iitb.ac.in/Drawgraph/input.html
IndoWordNet Visualizer contd..

http://www.cfilt.iitb.ac.in/Drawgraph/input.html
BabelNet Interface

boy

Noun

boy, male child
A youthful male person
ID: 00012559n | Concept

boy
A friendly informal reference to a grown man
ID: 00012570n | Concept

boy, son
A male human offspring
ID: 00012571n | Concept

http://babelnet.org/
BabelNet Synset

http://babelnet.org/
Wordnets in the World

- The Global WordNet Organization gives access of wordnets in the world
- Albanian, Arabic, Spanish, Catalan, Basque, Italian, Bulgarian, Czech, Greek, Romanian, Serbian, Turkish, Chinese, Danish, Dutch, Estonian, French, German, Hungarian, Icelandic, Portuguese, Irish, Japanese, Korean, Kurdish, Latin, Macedonian, Norwegian, Persian, Polish, Russian, Swedish
WordNet API’s and similarity tools

• English:
  – Java API: extJWNL, JAWS, JWNL
  – Python API: NLTK
  – WordNet:::Similarity tool

• Hindi:
  – Java API: JHWNL
  – Python API
  – IndoWordNet:::Similarity tool
Outline

• What is WordNet?
• WordNet Synset
• Principles used for Synset Creation
• Lexico-Semantic Relations
• Important WordNets: English, Hindi, IndoWordNet, BabelNet, EuroWordNet
• Applications
WordNet Applications

- Machine Translation
- Word Sense Disambiguation
- Sentiment Analysis
- Information Retrieval
- MultiWord Expression Detection
- Document structuring and categorization
- Cognitive NLP
Word Sense Disambiguation
Outline

• Introduction
  – Ambiguity
  – WSD Definition
  – Position of WSD in NLP layers
• Motivation
• WSD block diagram
• Lexical Resources needed
  – Sense Repository
  – Sense Annotated Corpus
• WSD approaches
  – Knowledge based
  – Corpus based (Supervised, Unsupervised)
• Applications
Outline

• Introduction
  – Ambiguity
  – WSD Definition
  – Position of WSD in NLP layers
• Motivation
• WSD block diagram
• Lexical Resources needed
  – Sense Repository
  – Sense Annotated Corpus
• WSD approaches
  – Knowledge based
  – Corpus Based (Supervised, Unsupervised)
• Applications
Ambiguity

• A word, phrase or sentence is ambiguous if it has more than one meaning

• Structural ambiguity: due to the sentence structure
  – *A boy saw a man with a telescope* (English)
  – राम ने दौड़ते हुए शेर को देखा (Hindi)

• Lexical ambiguity: due to polysemous words
  – *She put her glasses on the table* (English)
  – पड़ोसी ने हमारे घर में आग लगायी (Hindi)
WSD Definition

• Word Sense Disambiguation (WSD) is the problem of computationally determining the ‘sense’ or ‘meaning’ of a word in a particular context.
WSD Example

यहाँ फोटो खींचना मना है

to capture  to pull  to extract

वह कुएँ से रस्सी खींचती है

to capture  to pull  to extract
Why WSD is difficult?

• Sometimes human even fails to disambiguate

‘उसका हाथ मशीन के नीचे आ गया’

1. हाथ, बाजू, हस्त, बाँ, बाह, बांज धान कर, - कन्धे से पंजे तक का वह अंग जिससे चीजें पकड़ते और काम करते हैं। गांधीजी के हाथ बहुत लंबे थे। / भीम की भुजाओं में बहुत बल था।

2. हाथ, कर, पंजा, पाणि - कहा भाग "उसका हाथ मशीन के नीचे आ गया।"

3. हाथ, हस्त, कर, पाणि - कोहनी से पंजे के सिरे तक का भाग "दुर्घटना में उसका दाहिना हाथ टूट गया।"

4. हाथ, हस्त - चौबीस अंगुल की एक नाप या "इस वस्त्र की लंबाई दो हाथ है।"

5. हाथ - ताश के खेल में एक दौर में गिरने वाला था। उसके बाद खेल से बाहर हो जाए, "मेरे सात हाथ बन चुके हैं।"
Why WSD is difficult? contd..

• From practical point of view, it is essential to make sense distinction according to the needs of the application

• **Coarse grained senses** – Information Retrieval, Information Extraction, Document Categorization, Machine Translation

• **Fine grained senses** – Language Learning, Machine Translation of distant languages like Chinese-English
Why WSD is difficult? contd..

- Generally verbs are more polysemous as compared to other parts-of-speech

<table>
<thead>
<tr>
<th>Verb</th>
<th>#Senses</th>
<th>Verb</th>
<th>#Senses</th>
<th>Verb</th>
<th>#Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>निकलना</td>
<td>31</td>
<td>निकालना</td>
<td>29</td>
<td>लगना</td>
<td>26</td>
</tr>
<tr>
<td>लगाना</td>
<td>23</td>
<td>चढ़ना</td>
<td>22</td>
<td>उतरना</td>
<td>21</td>
</tr>
<tr>
<td>पड़ना</td>
<td>19</td>
<td>आना</td>
<td>19</td>
<td>छोड़ना</td>
<td>19</td>
</tr>
<tr>
<td>चढ़ना</td>
<td>18</td>
<td>चलना</td>
<td>17</td>
<td>उठना</td>
<td>17</td>
</tr>
<tr>
<td>मिलना</td>
<td>17</td>
<td>देखना</td>
<td>16</td>
<td>बोलना</td>
<td>16</td>
</tr>
<tr>
<td>टूटना</td>
<td>16</td>
<td>खुलना</td>
<td>15</td>
<td>उड़ना</td>
<td>15</td>
</tr>
<tr>
<td>उठना</td>
<td>14</td>
<td>खोलना</td>
<td>14</td>
<td>छूटना</td>
<td>14</td>
</tr>
<tr>
<td>बनना</td>
<td>14</td>
<td>लेना</td>
<td>13</td>
<td>रहना</td>
<td>13</td>
</tr>
<tr>
<td>जमना</td>
<td>12</td>
<td>बांधना</td>
<td>12</td>
<td>बैठना</td>
<td>12</td>
</tr>
<tr>
<td>खाना</td>
<td>12</td>
<td>काटना</td>
<td>12</td>
<td>बांधना</td>
<td>12</td>
</tr>
</tbody>
</table>
Position of WSD in NLP layers

- Morphology
- POS tagging
- Parsing
- Semantics
- Discourse
- Pragmatics

Level of Complexity

WSD
Outline

• Introduction
  – Ambiguity
  – WSD Definition
  – Position of WSD in NLP layers

• Motivation

• WSD block diagram

• Lexical Resources needed
  – Sense Repository
  – Sense Annotated Corpus

• WSD approaches
  – Knowledge based
  – Corpus Based (Supervised, Unsupervised)

• Applications
Motivation

- TE
- MT
- SA
- NER
- SP
- CLIR
Outline

• Introduction
  – Ambiguity
  – WSD Definition
  – Position of WSD in NLP layers
• Motivation
• WSD block diagram
• Lexical Resources needed
  – Sense Repository
  – Sense Annotated Corpus
• WSD approaches
  – Knowledge based
  – Corpus Based (Supervised, Unsupervised)
• Applications
Block diagram of WSD

Training corpora:
- Sense tagged or
- Parallel or
- Comparable or
- Untagged

Test corpora:
- Untagged text

Knowledge resources:
- Wordnet, Thesauri,
- Ontologies

WSD System

Sense tagged test corpora
Outline

• Introduction
  – Ambiguity
  – WSD Definition
  – Position of WSD in NLP layers
• Motivation
• WSD block diagram
• Lexical Resources needed
  – Sense Repository
  – Sense Annotated Corpus
• WSD approaches
  – Knowledge based
  – Corpus Based (Supervised, Unsupervised)
• Applications
Lexical Resources for WSD

• Sense Repository
  – Dictionary
  – Thesaurus
  – Wordnet

• Sense Annotated Corpus
WordNet

• Lexical knowledge base
• Consists of synsets and semantic relations
• For example: Senses of ‘boy’ from WordNet
  – (10305010) S: (n) **male child, boy** (a youthful male person) "the baby was a boy"; "she made the boy brush his teeth every night"; "most soldiers are only boys in uniform"

  – (09890332) S: (n) **boy** (a friendly informal reference to a grown man) "he likes to play golf with the boys"

  – (10643436) S: (n) **son, boy** (a male human offspring) "their son became a famous judge"; "his boy is taller than he is"
WordNet: Lexico-Semantic relations

- Boy
  - male, male person
  - boyhood
    - Derivationally related form
  - Schoolboy
    - hyponymy
  - Scout
    - hyponymy
  - Girl
    - antonymy

- hypernymy
- hyponymy
http://www.cfilt.iitb.ac.in/indowordnet/
लड़का, बालक, बाल, बच्चा, छोकड़ा, छोरा, छोकरा
कम उम्र का पुरुष,
विशेषकर अविवाहित
"मैदान में लड़के क्रिकेट खेल रहे हैं।"

मुलगा, पोरगा, पोर, पोरगे
साधारणतः सोळा वर्षाखालील पुरुष व्यक्ती
"तो मुलगा खूपच हुशार आहे।"
Sense Annotated Corpus

- Corpus annotated with sense tags from wordnet
  - English corpus:
    - SemCor Corpus, OntoNotes, DSO, Senseval, SemLink
  - Indian language corpus:
    - CFILT corpus (Hindi and Marathi Health-Tourism)
  - Japanese corpus
    - Jsemcor corpus
  - Dutch corpus:
    - DutchSemCor
  - Spanish corpus:
    - SpsemCor
Sense Annotated Corpus contd..

CFILT corpus: (Hindi-health domain)

• व्यायाम_5939 शरीर_1961 को स्वस्थ_1831 और तन्दुरुस्त_1831 रखने_ में सहायता_3623 करता_ है

• दैनिक_6246 व्यायाम_5939 सबसे उत्कृष्ट_2360 लाभ_2751 प्रदान_1694 करते_ हैं

• स्वास्थ्य_8407 शारीरिक_9166 , मानसिक_2151 और सामाजिक_3540 सुख_3538 की एक_187 अवस्था_652 है

• इसमें केवल_4509 बीमारी_1423 की अनुपस्थिति_6745 से भी अधिक_2403 शामिल_10810 है
Outline

• Introduction
  – Ambiguity
  – WSD Definition
  – Position of WSD in NLP layers
• Motivation
• WSD block diagram
• Lexical Resources needed
  – Sense Repository
  – Sense Annotated Corpus
• WSD approaches
  – Knowledge based
  – Corpus Based (Supervised, Unsupervised)
• Applications
WSD approaches

• Knowledge-based WSD:
  – uses an explicit lexicon (machine readable dictionary (MRD), thesaurus) or ontology (e.g. WordNet).

• Corpus-based WSD: (Supervised & Unsupervised)
  – the relevant information about word senses is gathered from training on a large corpus.

• Hybrid approach:
  – combining aspects of both of the aforementioned methodologies
# Knowledge-based WSD

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSD using Selectional Restrictions</td>
<td>44% on Brown Corpus</td>
</tr>
<tr>
<td>Lesk’s algorithm</td>
<td>50-60% on short samples of “Pride and Prejudice” and some “news stories”.</td>
</tr>
<tr>
<td>WSD using conceptual density</td>
<td>54% on Brown corpus.</td>
</tr>
<tr>
<td>WSD using Random Walk Algorithms</td>
<td>54% accuracy on SEMCOR corpus which has a baseline accuracy of 37%.</td>
</tr>
<tr>
<td>Walker’s algorithm</td>
<td>50% when tested on 10 highly polysemous English words.</td>
</tr>
</tbody>
</table>
Simple Lesk Algorithm

• Example: *pine cone*

  pine 1 kinds of evergreen tree with needle-shaped leaves
       2 waste away through sorrow or illness
cone 1 solid body which narrows to a point
       2 something of this shape whether solid or hollow
       3 fruit of certain evergreen trees

• Dictionary definitions of pine1 and cone3 literally overlap: “evergreen” + “tree”

• So “pine cone” must be pine1 + cone3
Simplified Lesk Algorithm

The **bank** can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

given the following two WordNet senses:

<table>
<thead>
<tr>
<th>bank$^1$</th>
<th>Gloss:</th>
<th>a financial institution that accepts deposits and channels the money into lending activities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Examples:</td>
<td>“he cashed a check at the bank”, “that bank holds the mortgage on my home”</td>
</tr>
<tr>
<td>bank$^2$</td>
<td>Gloss:</td>
<td>sloping land (especially the slope beside a body of water)</td>
</tr>
<tr>
<td></td>
<td>Examples:</td>
<td>“they pulled the canoe up on the bank”, “he sat on the bank of the river and watched the currents”</td>
</tr>
</tbody>
</table>

• Count words in the context (sentence) which are also in the Gloss or Example for 1 and 2;
• Choose the word-sense with most “overlap”
Corpus Based approaches

• A corpus-based approach extracts information on word senses from a large annotated data collection.
• Distributional information about an ambiguous word refers to the frequency distribution of its senses
• collocational or co-occurrence information
• part-of-speech
• ...

Corpus Based approaches

• There are two possible approaches to corpus-based WSD systems:

  – **Supervised approaches**
    • use annotated training data
    • basically amount to a classification task

  – **Unsupervised algorithms**
    • applied to raw text material
    • annotated data is only needed for evaluation
    • correspond to a clustering task rather than a classification.

  – **Bootstrapping**
    • looks like supervised approaches
    • it needs only a few seeds instead of a large number of training examples
## Supervised Approaches

<table>
<thead>
<tr>
<th>Approach</th>
<th>Average Precision</th>
<th>Average Recall</th>
<th>Corpus</th>
<th>Average Baseline Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>64.13%</td>
<td>Not reported</td>
<td>Senseval3 – All Words Task</td>
<td>60.90%</td>
</tr>
<tr>
<td>Decision Lists</td>
<td>96%</td>
<td>Not applicable</td>
<td>Tested on a set of 12 highly polysemous English words</td>
<td>63.9%</td>
</tr>
<tr>
<td>Exemplar Based disambiguation (k-NN)</td>
<td>68.6%</td>
<td>Not reported</td>
<td>WSJ6 containing 191 content words</td>
<td>63.7%</td>
</tr>
<tr>
<td>SVM</td>
<td>72.4%</td>
<td>72.4%</td>
<td>Senseval 3 – Lexical sample task (Used for disambiguation of 57 words)</td>
<td>55.2%</td>
</tr>
<tr>
<td>Perceptron trained HMM</td>
<td>67.60</td>
<td>73.74%</td>
<td>Senseval3 – All Words Task</td>
<td>60.90%</td>
</tr>
</tbody>
</table>
Unsupervised approaches

• Supervised WSD performs well but needs sense tagged corpora

• Obtaining sense tagged corpora is costly in terms of time and money

• A high degree of language dependence and makes it difficult to apply them to a variety of languages

• Despite of the less accuracy, unsupervised approaches are chosen for their resource consciousness and robustness
Classification of Unsupervised WSD Methods

Strictly Unsupervised Methods

Distributional Methods

Type Based
- Hyperlex
- Latent Semantic Indexing (LSA)
- Hyper Space Analogue to Language (HAL)
- Clustering by Committee (CBC)

Token Based

Translational Equivalence Methods

- Brown et al
- Context Group Discrimination
- McQuitty’s Similarity Analysis
<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Average Recall</th>
<th>Corpus</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin’s Algorithm</td>
<td>68.5%.</td>
<td>Not reported</td>
<td>Trained using WSJ corpus containing 25 million words. Tested on 7 SemCor files containing 2832 polysemous nouns.</td>
<td>64.2%</td>
</tr>
<tr>
<td>Hyperlex</td>
<td>97%</td>
<td>82%</td>
<td>Tested on a set of 10 highly polysemous French words</td>
<td>73%</td>
</tr>
<tr>
<td>WSD using Roget’s Thesaurus categories</td>
<td>92% (average degree of polysemy was 3)</td>
<td>Not reported</td>
<td>Tested on a set of 12 highly polysemous English words</td>
<td>Not reported</td>
</tr>
<tr>
<td>WSD using parallel corpora</td>
<td>SM: 62.4%</td>
<td>SM: 61.6%</td>
<td>Trained using a English Spanish parallel corpus Tested using Senseval 2 – All Words task (only nouns were considered)</td>
<td>Not reported</td>
</tr>
<tr>
<td></td>
<td>CM: 67.2%</td>
<td>CM: 65.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hyperlex (Veronis, 2004)

- Target word WSD developed for Information Retrieval applications
- Instead of using “dictionary defined senses” extract the “senses from the corpus” itself
- Works only for nouns and adjectives
- Co-occurrence graph is constructed for words which co-occur with the target word
- Words which are syntactically correlated are connected with edges
- Weight of an edge is determined by following formula:

\[ w_{AB} = 1 - \max(P(A|B), P(B|A)) \]
Example of co-occurrence graph

Co-occurrence graph for the word वीज (electricity/lightening)
Root Hubs Detection

Co-occurrence graph for the word वीज (electricity/lightening)

- Root hubs are identified as the most connected nodes of each strongly connected component
Target Word Added

Co-occurrence graph for the word वीज (electricity/lightening)

- Target word is added to the graph and connected to root hubs using edges of zero weight
Minimum Spanning Tree found

Co-occurrence graph for the word वीज (electricity/lightening)

Then score vector for each word is computed as follows:

\[
S_i = \begin{cases} 
\frac{1}{1 + d(h_i, v)} & \text{if } v \in \text{component } i \\
0 & \text{otherwise}
\end{cases}
\]

Where, \( d(h_i, v) \) is the distance between the root hub \( h_i \) and node \( v \)
Hyperlex contd..

• For the given occurrence of a target word, only words from its context take part in the scoring process

• The score vectors of all words are added for the given context

• The component with highest score becomes the winner sense

• Accuracy: 97% for 10 highly polysemous French words
## Comparing WSD approaches

<table>
<thead>
<tr>
<th></th>
<th>Supervised</th>
<th>Semi-Supervised</th>
<th>Unsupervised</th>
<th>Knowledge based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>high</td>
<td>moderate</td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td><strong>Need of tagged corpora</strong></td>
<td>yes</td>
<td>Very few</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td><strong>Need of Knowledge resources</strong></td>
<td>No</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
Outline

- Introduction
  - Ambiguity
  - WSD Definition
  - Position of WSD in NLP layers
- Motivation
- WSD block diagram
- Lexical Resources needed
  - Sense Repository
  - Sense Annotated Corpus
- WSD approaches
  - Knowledge based
  - Corpus Based (Supervised, Unsupervised)
- Applications
WSD Applications

• Machine Translation
  – Translate “bill” from English to Spanish
  – Is it a “pico” or a “cuenta”? 
  – Is it a bird jaw or an invoice?

• Information Retrieval
  – Find all Web Pages about “cricket” 
  – The sport or the insect?

• Question Answering
  – What is George Miller’s position on gun control? 
  – The psychologist or US congressman?
WSD @ IIT Bombay
Unsupervised WSD approaches

• Approach 1:
  – Bilingual WSD using Expectation Maximization (EM) algorithm
    (Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya, Neighbor Help: Bilingual Unsupervised WSD Using Context, Association for Computational Linguistics (ACL) 2013, Sofia, Bulgaria, 4-9 August, 2013 )

• Approach 2:
  – Most Frequent Sense Detection using Word vectors or embeddings
Unsupervised WSD approaches

• Approach 1:
  – Bilingual WSD using Expectation Maximization (EM) algorithm
    (Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya, Neighbor Help: Bilingual Unsupervised WSD Using Context, Association for Computational Linguistics (ACL) 2013, Sofia, Bulgaria, 4-9 August, 2013 )

• Approach 2:
  – Most Frequent Sense Detection using Word vectors or embeddings
Problem Statement

• For a given untagged text of two languages perform word sense disambiguation using unsupervised technique
Overview of the approach

- Extension of Bilingual WSD (Khapra et al., 2011) by adding context
- Two resource scarce languages can help each other without the need of any sense tagged corpora in either languages.
- Approach uses untagged corpora and the aligned wordnets
- Approach relies on the key observation that sense distribution of any language remains same within a domain
- Context-based EM formulation is used for estimating the sense distribution
- An improvement of 17% - 35% in verb accuracy
Mode of Working

Marathi Language

\( S_{1}^{mar}, S_{2}^{mar} \) paan पान

Hindi Language

panna \( S_{2}^{hin} \) पन्ना

\( S_{1}^{mar} \) parna पर्ण

parna \( S_{1}^{hin} \) पर्ण

\( S_{3}^{mar} \) patte पत्ते

patta (\( S_{1}^{hin}, S_{3}^{hin} \)) पत्ता

A bipartite graph of translation correspondences
Formulation

Marathi language

\[ P(S_{1}^{mar} | \text{paan}) = \frac{\#(S_{1}^{mar}, \text{paan})}{\#(S_{1}^{mar}, \text{paan}) + \#(S_{2}^{mar}, \text{paan})} \]

Using Cross-links in Hindi:

\[ P(S_{1}^{mar} | \text{paan}) = \frac{\#(S_{1}^{hin}, \text{patta}) + \#(S_{1}^{hin}, \text{parna})}{\#(S_{1}^{hin}, \text{patta}) + \#(S_{1}^{hin}, \text{parna}) + \#(S_{2}^{hin}, \text{panna})} \]

where,

\[ \#(S_{1}^{hin}, \text{patta}) = P(S_{1}^{hin} | \text{patta}) \times \#(\text{patta}) \]

Marathi language

\[ P(S_{1}^{hin} | \text{patta}) = \frac{\#(S_{1}^{mar}, \text{paan}) + \#(S_{1}^{mar}, \text{parna})}{\#(S_{1}^{mar}, \text{paan}) + \#(S_{1}^{mar}, \text{parna}) + \#(S_{3}^{mar}, \text{patte})} \]
Formulation by Khapra et al., 2011

E- Step:

\[
P(S_{L1} | u) = \frac{\sum_{s_i^{L1}} P(\pi_{L2} (S_{L1}^{L1}) | v).\#(v)}{\sum_{s_i^{L1}} \sum_{y} P(\pi_{L2} (S_{L1}^{L1}) | y).\#(y)}
\]

\(s_i^{L1} \in \text{synsets}_{L1}(u)\)

\(v \in \text{crosslinks}_{L2}(u, S_{L1}^{L1})\)

\(y \in \text{crosslinks}_{L2}(u, S_{i}^{L1})\)

M- Step:

\[
P(S_{L2} | v) = \frac{\sum_{s_i^{L2}} P(\pi_{L1} (S_{L2}^{L2}) | u).\#(u)}{\sum_{s_i^{L2}} \sum_{z} P(\pi_{L1} (S_{L2}^{L2}) | z).\#(z)}
\]

\(s_i^{L2} \in \text{synsets}_{L2}(v)\)

\(u \in \text{crosslinks}_{L1}(v, S_{L2}^{L2})\)

\(z \in \text{crosslinks}_{L1}(v, S_{i}^{L2})\)

Two languages mutually help each other in estimating sense distribution
Adding Context

Basic formulation

\[
P(S_{mar}^{1} \mid paan) = \frac{P(S_{1}^{hin} \mid patta) * \#(patta) + P(S_{1}^{hin} \mid parna) * \#(parna)}{P(S_{1}^{hin} \mid patta) * \#(patta) + P(S_{1}^{hin} \mid parna) * \#(parna) + P(S_{3}^{hin} \mid parna) * \#(parna)}
\]

After adding the context

\[
P(S_{mar}^{1} \mid paan, zaad) = \frac{\#(S_{1}^{hin} \mid patta, ped).\#(patta, ped) + \#(S_{1}^{hin} \mid parna, ped).\#(parna, ped) + \#(S_{3}^{hin} \mid parna, ped).\#(parna, ped)}{\#(S_{1}^{hin} \mid patta, ped).\#(patta, ped) + \#(S_{1}^{hin} \mid parna, ped).\#(parna, ped) + \#(S_{3}^{hin} \mid parna, ped).\#(parna, ped)}
\]
Concurrence counts are unreliable

They can make sense only if we have huge amount of corpora

Semantic relatedness gives a good estimation of co-occurrence count.

Adding Semantic Relatedness
New Formulation

After adding semantic relatedness

**E-step:**

\[
P(S_{L1}^{L1}|u,a) = \frac{\sum_{v,b} P(\pi_{L2}(S_{L1}^{L1})|v,b) \cdot \sigma(v,b)}{\sum_{i} \sum_{x,b} P(\pi_{L2}(S_{i}^{L1})|x,b) \cdot \sigma(x,b)}
\]

where, \(S_{i}^{L1} \in \text{synsets}_{L1}(u)\)

\[a \in \text{context}(u)\]

\[v \in \text{crosslinks}_{L2}(u,S_{L1}^{L1})\]

\[b \in \text{crosslinks}_{L2}(a)\]

\[x \in \text{crosslinks}_{L2}(u,S_{i}^{L1})\]

**M-step:**

\[
P(S_{L2}^{L2}|v,b) = \frac{\sum_{u,a} P(\pi_{L1}(S_{L2}^{L2})|u,a) \cdot \sigma(u,a)}{\sum_{i} \sum_{y,b} P(\pi_{L1}(S_{i}^{L2})|y,a) \cdot \sigma(y,a)}
\]

where, \(S_{i}^{L2} \in \text{synsets}_{L2}(v)\)

\[b \in \text{context}(v)\]

\[u \in \text{crosslinks}_{L1}(v,S_{L2}^{L2})\]

\[a \in \text{crosslinks}_{L1}(b)\]

\[y \in \text{crosslinks}_{L1}(v,S_{i}^{L2})\]
Results on Health domain

Hindi Health Domain

Marathi Health Domain
Results on Tourism domain

Hindi Tourism Domain

Marathi Tourism Domain
Error Analysis contd..

The semantic structure of the sentence can help in such situations.

वे पत्ते खेल रहे हैं
(vaha patte khel rahe the)
(They are playing cards)

वे पेड़ के नीचे पत्ते खेल रहे हैं
(vaha ped ke niche patte khel rahe hai)
(They are playing cards below the tree)
Error Analysis

Function words help in disambiguation, since they define semantic relations between two content words.
Error Analysis contd..

• We have considered single word crosslinks in our approach.
• Sometimes one word has multi-word crosslinks in another language.

अब आता, या वेळी, या वेळेस, ह्या वेळी, ह्या वेळेस
(ab) (aata, ya veli, ya veles, hya veli, hya veles)

(Hindi) (Marathi)

Language properties also play an important role
Error Analysis contd..

• Resource related problems:
  – too fine grained HWN senses

We should consider coarse-grained senses to increase accuracy
Bilingual WSD using Word Embeddings

- Word embeddings are used as an approximation to the co-occurrence counts.
- Verb accuracy improved by 8.5% for Marathi.
- Adjective accuracy improved by 7% for Hindi and 2.5% for Marathi.

<table>
<thead>
<tr>
<th>WSD Algorithm</th>
<th>HIN-HEALTH</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>MAR-HEALTH</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOUN</td>
<td>ADV</td>
<td>ADJ</td>
<td>VERB</td>
<td>Overall</td>
<td>NOUN</td>
<td>ADV</td>
<td>ADJ</td>
<td>VERB</td>
<td>Overall</td>
</tr>
<tr>
<td>Combined</td>
<td>59.32</td>
<td>68.98</td>
<td>63.18</td>
<td>60.02</td>
<td>60.94</td>
<td>62.75</td>
<td>61.19</td>
<td>56.22</td>
<td>60.99</td>
<td>61.30</td>
</tr>
<tr>
<td>EM-C-DistSimi</td>
<td>59.59</td>
<td>69.20</td>
<td>63.87</td>
<td>55.73</td>
<td>61.09</td>
<td>63.09</td>
<td>61.82</td>
<td>55.60</td>
<td>43.69</td>
<td>58.92</td>
</tr>
<tr>
<td>EM-C-WnSimi</td>
<td>59.82</td>
<td>67.80</td>
<td>56.66</td>
<td>60.38</td>
<td>59.63</td>
<td>62.90</td>
<td>62.54</td>
<td>53.63</td>
<td>52.49</td>
<td>59.77</td>
</tr>
<tr>
<td>EM</td>
<td>60.68</td>
<td>67.48</td>
<td>55.54</td>
<td>25.29</td>
<td>58.16</td>
<td>63.88</td>
<td>58.88</td>
<td>55.71</td>
<td>35.60</td>
<td>58.03</td>
</tr>
<tr>
<td>WFS</td>
<td>53.49</td>
<td>73.24</td>
<td>55.16</td>
<td>38.64</td>
<td>54.46</td>
<td>59.35</td>
<td>67.32</td>
<td>38.12</td>
<td>34.91</td>
<td>52.57</td>
</tr>
<tr>
<td>RB</td>
<td>32.52</td>
<td>45.08</td>
<td>35.42</td>
<td>17.93</td>
<td>33.31</td>
<td>33.83</td>
<td>38.76</td>
<td>37.68</td>
<td>18.49</td>
<td>32.45</td>
</tr>
</tbody>
</table>
Unsupervised WSD approaches

• Approach 1:
  – Bilingual WSD using Expectation Maximization (EM) algorithm
    (Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya, Neighbor Help: Bilingual Unsupervised WSD Using Context, Association for Computational Linguistics (ACL) 2013, Sofia, Bulgaria, 4-9 August, 2013 )

• Approach 2:
  – Most Frequent Sense Detection using Word vectors or embeddings
Most Frequent Sense Detection

• **Problem Statement:**
  • For a given word, find the most frequent sense of a word using unsupervised technique

• **Motivation:**
  • The first sense heuristic is often used as a baseline for WSD systems
  • For WSD systems, it is hard to beat this baseline (5 out of 26 supervised approaches beat this baseline)
  • Manually tagging data is costly in terms of time and money
  • It would be useful to have a method of ranking senses directly from untagged data
Related Work

- Most Frequent Sense Detection
  - Using Category Information from Thesaurus
  - Using WordNet Semantic Similarity
  - Using Clustering by Committee
  - Using Syntactic Evidence

References:
- Mohammad and Hirst, 2006
- Buitelaar et al., 2001
- McCarthy et al., 2007
- Pantel and Lin, 2002
- Lapata and Brew., 2004
Our Approach [UMFS–WE]

• A unsupervised approach for MFS detection using word embeddings

• Word embedding of a word is compared with sense embeddings and the sense with highest similarity is considered as the most frequent sense

• Extendable and portable: Domain independent approach and easily portable to multiple languages
Word Embeddings

• Represent each word with low-dimensional real valued vector.

• Increasingly being used in variety of Natural Language Processing tasks

• **word2vec tool** (Mikolov et. al, 2013)
  – One of the most popular word embedding tool
  – Source code provided
Word Embeddings contd..

Continuous bag of words model (CBOW)

Skip-gram model
Word Embeddings contd..

- **word2vec tool** (Mikolov et. al, 2013)

  - It captures many linguistic regularities

    \[
    \text{Vector('king')} - \text{Vector('man')} + \text{Vector('woman')} \Rightarrow \text{Vector('queen')}
    \]
**Word Embeddings contd..**

- Distributionally Similar words of **फल** (fala, fruit)

<table>
<thead>
<tr>
<th>words</th>
<th>cosine similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>फ़ल</td>
<td>0.840545</td>
</tr>
<tr>
<td>केला</td>
<td>0.705185</td>
</tr>
<tr>
<td>ल</td>
<td>0.688565</td>
</tr>
<tr>
<td>सीताफल</td>
<td>0.685993</td>
</tr>
<tr>
<td>पपीता</td>
<td>0.682171</td>
</tr>
<tr>
<td>सौंदर्यवर्धक</td>
<td>0.677420</td>
</tr>
<tr>
<td>कन्दमूल</td>
<td>0.672466</td>
</tr>
<tr>
<td>अननास</td>
<td>0.655930</td>
</tr>
<tr>
<td>भाजियाँ</td>
<td>0.650811</td>
</tr>
<tr>
<td>आडू</td>
<td>0.650100</td>
</tr>
</tbody>
</table>
**Sense Embeddings**

- The **sense-bag** for the sense $S_i$ is created as below,

\[ SB(S_i) = \{x | x - \text{Features}(S_i) \} \]

  - Features($S_i$) - WordNet based features for sense $S_i$

- Sense embeddings are obtained by taking the average of word embeddings of each word in the sense-bag

\[ \text{vec}(S_i) = \frac{\sum_{x \in SB(S_i)} \text{vec}(x)}{N} \]

  - $S_i$ - $i^{th}$ sense of a word $W$
  - $N$ - Number of words present in the sense-bag $SB(S_i)$
MFS Detection

- We treat the MFS identification problem as finding the closest cluster centroid (i.e., sense embedding)
- Cosine similarity is used.
- Most frequent sense is obtained

\[ MFS_w = \arg\max_{S_i} \cos(vec(W), vec(S_i)) \]

- \( vec(W) \) - word embedding of a word \( W \)
- \( S_i \) - \( i^{th} \) sense of word \( W \)
- \( vec(S_i) \) - sense embedding for \( S_i \)
MFS Detection

02232196: cricket (leaping insect; male makes chirping noises by rubbing the forewings together)

00477400: cricket (a game played with a ball and bat by two teams of 11 players; teams take turns trying to score runs)
MFS Detection contd..

- insect
- chirping
- rubbing
- forewings

- noises

- cricket
- played game
- ball
- runs
- team
- bat

S_1

S_2
Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      • Hindi (Newspaper)
      • English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
   3. Comparing WSD results using different sense vector models
      • Retrofitting Sense Vector Model (English)
   4. Experiments on WSD for words which do not exists in SemCor

B. Experiments on selected words (34 polysemous words from SENSEVAL-2 corpus)
   1. Experiments using different word vector models
   2. Comparing results with various sizes of vector dimensions
Experiments

A. Experiments on WSD

1. Experiments on WSD using Skip-Gram model
   - Hindi (Newspaper)
   - English (SENSEVAL-2 and SENSEVAL-3)
A.1] Experiments on WSD using skip-gram model

• Training of word embeddings:
  – Hindi: Bojar (2014) corpus (44 M sentences)
  – English: Pre-trained Google-News word embeddings

• Datasets used for WSD:
  – Hindi: Newspaper dataset
  – English: SENSEVAL-2 and SENSEVAL-3

• Experiments are restricted to only polysemous nouns.
### [A.1] Results on WSD

<table>
<thead>
<tr>
<th>HINDI WSD</th>
<th>Newspaper dataset</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-Score</td>
</tr>
<tr>
<td>UMFS-WE</td>
<td>62.43</td>
<td>61.58</td>
<td>62.00</td>
</tr>
<tr>
<td>WFS</td>
<td>61.73</td>
<td>59.31</td>
<td>60.49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ENGLISH WSD</th>
<th>SENSEVAL-2 dataset</th>
<th>SENSEVAL-3 dataset</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-Score</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>UMFS-WE</td>
<td>52.39</td>
<td>52.27</td>
<td>52.34</td>
<td>43.34</td>
<td>43.22</td>
</tr>
<tr>
<td>WFS</td>
<td>61.72</td>
<td>58.16</td>
<td>59.88</td>
<td>66.57</td>
<td>64.89</td>
</tr>
</tbody>
</table>
[A.1] Results on WSD contd..

- F-Score is also calculated for increasing thresholds on the frequency of nouns appearing in the corpus.

**F-Score values for Our Approach vs WFS on Hindi Newspaper Dataset**

**F-Score values for Our Approach vs SemCor on English SENSEVAL-2**
### [A.1] Results on WSD contd..

- WordNet feature selection for sense embeddings creation

<table>
<thead>
<tr>
<th>Sense Vectors Using WordNet features</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB</td>
<td>51.73</td>
<td>38.13</td>
<td>43.89</td>
</tr>
<tr>
<td>SB+GB</td>
<td>53.31</td>
<td>52.39</td>
<td>52.85</td>
</tr>
<tr>
<td>SB+GB+EB</td>
<td>56.61</td>
<td>55.84</td>
<td>56.22</td>
</tr>
<tr>
<td>SB+GB+EB+PSB</td>
<td>59.53</td>
<td>58.72</td>
<td>59.12</td>
</tr>
<tr>
<td>SB+GB+EB+PGB</td>
<td>60.57</td>
<td>59.75</td>
<td>60.16</td>
</tr>
<tr>
<td>SB+GB+EB+PEB</td>
<td>60.12</td>
<td>59.3</td>
<td>59.71</td>
</tr>
<tr>
<td>SB+GB+EB+PSB+PGB</td>
<td>57.59</td>
<td>56.81</td>
<td>57.19</td>
</tr>
<tr>
<td>SB+GB+EB+PSB+PEB</td>
<td>58.93</td>
<td>58.13</td>
<td>58.52</td>
</tr>
<tr>
<td>SB+GB+EB+PGB+PEB</td>
<td><strong>62.43</strong></td>
<td><strong>61.58</strong></td>
<td><strong>62</strong></td>
</tr>
<tr>
<td>SB+GB+EB+PSB+PGB+PEB</td>
<td>58.56</td>
<td>57.76</td>
<td>58.16</td>
</tr>
</tbody>
</table>

Table: Hindi WSD results using various WordNet features for Sense Embedding creation

- **SB**: Synset Bag
- **GB**: Gloss Bag
- **EB**: Example Bag
- **PSB**: Parent Synset Bag
- **PGB**: Parent Gloss Bag
- **PEB**: Parent Example Bag
Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      • Hindi (Newspaper)
      • English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
[A.2] Experiments on WSD using various Word Vector models

- We compared MFS results on various word vector models as listed below:

<table>
<thead>
<tr>
<th>Word Vector Model</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkipGram-Google-News (Mikolov et. al, 2013)</td>
<td>300</td>
</tr>
<tr>
<td>Senna (Collobert et. al, 2011)</td>
<td>50</td>
</tr>
<tr>
<td>MetaOptimize (Turian et. al, 2010)</td>
<td>50</td>
</tr>
<tr>
<td>RNN (Mikolov et. al, 2011)</td>
<td>640</td>
</tr>
<tr>
<td>Glove (Pennington et. al, 2014)</td>
<td>300</td>
</tr>
<tr>
<td>Global Context (Huang et. al, 2013)</td>
<td>50</td>
</tr>
<tr>
<td>Multilingual (Faruqui et.al, 2014)</td>
<td>512</td>
</tr>
<tr>
<td>SkipGram-BNC (Mikolov et. al, 2013)</td>
<td>300</td>
</tr>
<tr>
<td>SkipGram-Brown (Mikolov et. al, 2013)</td>
<td>300</td>
</tr>
</tbody>
</table>

Table: Word Vector Models
[A.2] Experiments on WSD using various Word Vector models contd..

<table>
<thead>
<tr>
<th>WordVector</th>
<th>Noun</th>
<th>Adj</th>
<th>Adv</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkipGram-Google-News</td>
<td>54.49</td>
<td>50.56</td>
<td>47.66</td>
<td>20.66</td>
</tr>
<tr>
<td>Senna</td>
<td>54.49</td>
<td>40.44</td>
<td>28.97</td>
<td>21.9</td>
</tr>
<tr>
<td>RNN</td>
<td>39.07</td>
<td>28.65</td>
<td>40.18</td>
<td>19.42</td>
</tr>
<tr>
<td>MetaOptimize</td>
<td>33.73</td>
<td>36.51</td>
<td>32.71</td>
<td>19.83</td>
</tr>
<tr>
<td>Glove</td>
<td>54.69</td>
<td>49.43</td>
<td>39.25</td>
<td>18.18</td>
</tr>
<tr>
<td>Global Context</td>
<td>48.30</td>
<td>32.02</td>
<td>31.77</td>
<td>20.66</td>
</tr>
<tr>
<td>SkipGram-BNC</td>
<td>53.03</td>
<td>48.87</td>
<td>39.25</td>
<td>23.14</td>
</tr>
<tr>
<td>SkipGram-Brown</td>
<td>30.29</td>
<td>48.87</td>
<td>27.10</td>
<td>13.29</td>
</tr>
</tbody>
</table>

Table: English WSD results for words with corpus frequency > 2
A. Experiments on WSD

1. Experiments on WSD using Skip-Gram model
   • Hindi (Newspaper)
   • English (SENSEVAL-2 and SENSEVAL-3)

2. Experiments on WSD using different word vector models

3. Comparing WSD results using different sense vector models
   • Retrofitting Sense Vector Model (Jauhar et al, 2015)
## [A.3] Results on WSD

<table>
<thead>
<tr>
<th>WordVector</th>
<th>SenseVector</th>
<th>Noun</th>
<th>Adj</th>
<th>Adv</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkipGram-Google-News</td>
<td>Our model</td>
<td><strong>58.87</strong></td>
<td>53.53</td>
<td><strong>46.34</strong></td>
<td>20.49</td>
</tr>
<tr>
<td>Retrofitting</td>
<td></td>
<td>47.84</td>
<td><strong>57.57</strong></td>
<td>32.92</td>
<td><strong>21.73</strong></td>
</tr>
<tr>
<td>Senna</td>
<td>Our model</td>
<td><strong>61.29</strong></td>
<td>43.43</td>
<td><strong>21.95</strong></td>
<td><strong>24.22</strong></td>
</tr>
<tr>
<td>Retrofitting</td>
<td></td>
<td>6.9</td>
<td><strong>68.68</strong></td>
<td>21.95</td>
<td>1.86</td>
</tr>
<tr>
<td>RNN</td>
<td>Our model</td>
<td><strong>42.2</strong></td>
<td>26.26</td>
<td><strong>40.24</strong></td>
<td><strong>21.11</strong></td>
</tr>
<tr>
<td>Retrofitting</td>
<td></td>
<td>10.48</td>
<td><strong>62.62</strong></td>
<td>21.95</td>
<td>1.24</td>
</tr>
<tr>
<td>MetaOptimize</td>
<td>Our model</td>
<td><strong>37.9</strong></td>
<td>50.5</td>
<td><strong>31.7</strong></td>
<td><strong>18.01</strong></td>
</tr>
<tr>
<td>Retrofitting</td>
<td></td>
<td>10.48</td>
<td><strong>62.62</strong></td>
<td>21.95</td>
<td>1.24</td>
</tr>
<tr>
<td>Glove</td>
<td>Our model</td>
<td><strong>58.33</strong></td>
<td>53.33</td>
<td><strong>39.02</strong></td>
<td><strong>17.39</strong></td>
</tr>
<tr>
<td>Retrofitting</td>
<td></td>
<td>9.94</td>
<td><strong>62.62</strong></td>
<td>21.95</td>
<td>1.24</td>
</tr>
<tr>
<td>Global Context</td>
<td>Our model</td>
<td><strong>53.22</strong></td>
<td>37.37</td>
<td><strong>24.39</strong></td>
<td>19.25</td>
</tr>
<tr>
<td>Retrofitting</td>
<td></td>
<td>12.36</td>
<td><strong>68.68</strong></td>
<td>21.95</td>
<td>1.24</td>
</tr>
<tr>
<td>SkipGram-Brown</td>
<td>Our model</td>
<td>29.31</td>
<td>60.6</td>
<td><strong>23.17</strong></td>
<td>11.42</td>
</tr>
<tr>
<td>Retrofitting</td>
<td></td>
<td>11.49</td>
<td><strong>68.68</strong></td>
<td>21.95</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Table: English WSD results for words with corpus frequency > 2
Experiments

A. Experiments on WSD

1. Experiments on WSD using Skip-Gram model
   • Hindi (Newspaper)
   • English (SENSEVAL-2 and SENSEVAL-3)

2. Experiments on WSD using different word vector models

3. Comparing WSD results using different sense vector models
   • Retrofitting Sense Vector Model (English)

4. Experiments on WSD for words which do not exists in SemCor
### [A.4] English WSD results for SENSEVAL-2 words which do not exist in SemCor

<table>
<thead>
<tr>
<th>Word Vector</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkipGram-Google-News</td>
<td>84.12</td>
</tr>
<tr>
<td>Senna</td>
<td>79.67</td>
</tr>
<tr>
<td>RNN</td>
<td>24.59</td>
</tr>
<tr>
<td>MetaOptimize</td>
<td>22.76</td>
</tr>
<tr>
<td>Glove</td>
<td>79.03</td>
</tr>
<tr>
<td>Global Context</td>
<td>28.09</td>
</tr>
<tr>
<td>Multilingual</td>
<td>35.48</td>
</tr>
<tr>
<td>SkipGram-BNC</td>
<td>68.29</td>
</tr>
<tr>
<td>SkipGram-BNC-Brown</td>
<td>74.79</td>
</tr>
</tbody>
</table>

proliferate, agreeable, bell_ringer, audacious, disco, delete, prestigious, option, peal, impaired, ringer, flatulent, unwashed, cervix, discordant, eloquently, carillon, full-blown, incompetence, stick_on, illiteracy, implicate, galvanize, retard, libel, obsession, altar, polyp, unintelligible, governance, bell_ringing.
Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      • Hindi (Newspaper)
      • English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
   3. Comparing WSD results using different sense vector models
      • Retrofitting Sense Vector Model (English)
   4. Experiments on WSD for words which do not exists in SemCor

B. Experiments on selected words (34 polysemous words from SENSEVAL-2 corpus)
   1. Experiments using different word vector models
[B.1] Experiments on selected words

- 34 polysemous nouns, where each one has at least two senses and which have occurred at least twice in the SENSEVAL-2 dataset are chosen

<table>
<thead>
<tr>
<th>Token</th>
<th>Senses</th>
<th>Token</th>
<th>Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>church</td>
<td>4</td>
<td>individual</td>
<td>2</td>
</tr>
<tr>
<td>field</td>
<td>13</td>
<td>child</td>
<td>4</td>
</tr>
<tr>
<td>bell</td>
<td>10</td>
<td>risk</td>
<td>4</td>
</tr>
<tr>
<td>rope</td>
<td>2</td>
<td>eye</td>
<td>5</td>
</tr>
<tr>
<td>band</td>
<td>12</td>
<td>research</td>
<td>2</td>
</tr>
<tr>
<td>ringer</td>
<td>4</td>
<td>team</td>
<td>2</td>
</tr>
<tr>
<td>tower</td>
<td>3</td>
<td>version</td>
<td>6</td>
</tr>
<tr>
<td>group</td>
<td>3</td>
<td>copy</td>
<td>3</td>
</tr>
<tr>
<td>year</td>
<td>4</td>
<td>loss</td>
<td>8</td>
</tr>
<tr>
<td>vicar</td>
<td>3</td>
<td>colon</td>
<td>5</td>
</tr>
<tr>
<td>sort</td>
<td>4</td>
<td>leader</td>
<td>2</td>
</tr>
<tr>
<td>country</td>
<td>5</td>
<td>discovery</td>
<td>4</td>
</tr>
<tr>
<td>woman</td>
<td>4</td>
<td>education</td>
<td>6</td>
</tr>
<tr>
<td>cancer</td>
<td>5</td>
<td>performance</td>
<td>5</td>
</tr>
<tr>
<td>cell</td>
<td>7</td>
<td>school</td>
<td>7</td>
</tr>
<tr>
<td>type</td>
<td>6</td>
<td>pupil</td>
<td>3</td>
</tr>
<tr>
<td>growth</td>
<td>6</td>
<td>student</td>
<td>2</td>
</tr>
</tbody>
</table>
# [B.1] MFS Results on selected words

<table>
<thead>
<tr>
<th>Word Vectors</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkipGram-BNC</td>
<td>63.63</td>
</tr>
<tr>
<td>SkipGram-Brown</td>
<td>48.38</td>
</tr>
<tr>
<td>SkipGram-Google-News</td>
<td>60.6</td>
</tr>
<tr>
<td>Senna</td>
<td>57.57</td>
</tr>
<tr>
<td><strong>Glove</strong></td>
<td><strong>66.66</strong></td>
</tr>
<tr>
<td>Global Context</td>
<td>51.51</td>
</tr>
<tr>
<td>Metaoptimize</td>
<td>27.27</td>
</tr>
<tr>
<td>RNN</td>
<td>51.51</td>
</tr>
<tr>
<td>Multilingual</td>
<td>63.4</td>
</tr>
</tbody>
</table>

Table: English WSD results for selected words from SENSEVAL-2 dataset
Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      • Hindi (Newspaper)
      • English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
   3. Comparing WSD results using different sense vector models
      • Retrofitting Sense Vector Model (English)
   4. Experiments on WSD for words which do not exists in SemCor

B. Experiments on selected words (34 polysemous words from SENSEVAL-2 corpus)
   1. Experiments using different word vector models
   2. Comparing results with various sizes of vector dimensions
[B.2] Comparing MFS results with various sizes of vector dimensions

<table>
<thead>
<tr>
<th>Word Vectors</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkipGram-BNC-1500</td>
<td>60.61</td>
</tr>
<tr>
<td>SkipGram-BNC-1000</td>
<td>60.61</td>
</tr>
<tr>
<td>SkipGram-BNC-500</td>
<td>66.67</td>
</tr>
<tr>
<td>SkipGram-BNC-400</td>
<td><strong>69.69</strong></td>
</tr>
<tr>
<td>SkipGram-BNC-300</td>
<td>63.64</td>
</tr>
<tr>
<td>SkipGram-BNC-200</td>
<td>60.61</td>
</tr>
<tr>
<td>SkipGram-BNC-100</td>
<td>48.49</td>
</tr>
<tr>
<td>SkipGram-BNC-50</td>
<td>51.52</td>
</tr>
</tbody>
</table>
**MFS for Indian Languages**

- Polyglot\(^1\) word embeddings are used for obtaining MFS.
  - word embeddings are trained using Wikipedia data.

- Currently, system is working for Marathi, Bengali, Gujarati, Sanskrit, Assamese, Bodo, Oriya, Kannada, Tamil, Telugu, Malayalam and Punjabi.

- Due to lack of gold data, we could not evaluate results

- APIs are developed for finding the MFS for a word

\(^1\)https://sites.google.com/site/rmyeid/projects/polyglot
MFS for using BabelNet

- MFS is calculated by using BabelNet as a sense repository.

- BabelNet covers 271 languages and is obtained from the automatic integration of: WordNet, Open Multilingual WordNet, Wikipedia, Omega Wiki, Wiktionary, Wikidata.

- System is working for English, Russian, Italian, French, German, and Spanish.

- Due to lack of gold data, we couldn't evaluate results for these language.
Conclusion

- WSD helps in solving ambiguity
- Bilingual WSD approach showed how two resource deprived languages help each other in WSD
- Unsupervised MFS approach showed that how word embeddings captures the MFS of a word
- Both the approaches are language independent
- They can be used in NLP applications
Publications Contributing to Thesis


Publications Contributing to Thesis contd..


• Book Chapter:
  
  **Word Sense Disambiguation Using IndoWordNet**

Other Publications


• Dhirendra Singh, Sudha Bhingardive, Kevin Patel and Pushpak Bhattacharyya, Detecting Multiword Expression using Word Embeddings and WordNet, 37th International Conference of the Linguistic Society of India (LSI-37), Javaharlal Nehru University, Delhi, 15-17 October, 2015.


References

- Pushpak Bhattacharyya, IndoWordNet, Lexical Resources Engineering Conference (LREC), Malta, May, 2010.


References


References


References


References


Thank You !!!
Contact Details

Sudha Bhingardive,
Dept. of CSE,
IIT Bombay, Mumbai

sudha@cse.iitb.ac.in
bhingardivesudha@gmail.com