Word Sense Disambiguation using Semantic Categories, Domain Information and Knowledge Sources

Presenter: Siva Reddy
Guide: Prof. Rajeev Sangal
Panel Members: Dr. SP Kishore, Dr. Suryakanth

Language Technologies Research Centre
IIIT Hyderabad
India

Outline

Outline

Computers versus Humans

- *Polysemy* - most words have many possible meanings.
- Ambiguity is rarely a problem for humans in their day to day communication, except in extreme cases.
- A computer program has no basis for knowing which one is appropriate, even if it is obvious to a human.

Ambiguity for a Computer

1. The fisherman jumped off the **bank** and into the water.
2. The **bank** down the street was robbed!
3. Back in the day, we had an entire **bank** of computers devoted to this problem.
4. The plane took a **bank** to the left, and then headed off towards the mountains.

Word Sense Disambiguation (WSD)

Example

- The roads are full of snow and the **coach** is delayed.
- Any **coach** can instruct you to hit the ball waist high.

Definition

The task of assigning the most appropriate sense for a word in a given context. Senses of a word are defined by a dictionary.

Word Sense Disambiguation (WSD)

Example

- The roads are full of snow and the **coach** is delayed.
- Any **coach** can instruct you to hit the ball waist high.

Definition

The task of assigning the most appropriate sense for a word in a given context. Senses of a word are defined by a dictionary.

Outline

Motivation

Machine Translation

- Translate **bat** from English to Telugu
    - Is it *Gabbilam* (Animal) or
    - Is it *chekka mukka* (Instrument)

Information Retrieval

- Find all Web Pages about **cricket**
    - The *sport* or the *insect*?

Question Answering

- When is **NTR's** marriage going to happen?
    - Is it *Junior* or *Senior* NTR.

Knowledge Acquisition

- Add to KB: Banda Karthika Reddy is the first lady mayor of **Hyderabad**.
    - *India* or *Pakistan*?

Previous Work

- Oldest problems in computational linguistics.
  - Dates back to early 1950s.
- Many Methods
  - Unsupervised
  - Semi-Supervised
  - Supervised
  - Knowledge Based (Current direction of research)
- State-of-art
  - Results of SENSEVAL-1, 2 and SEMEVAL-2007, 2010
  - Unsupervised, Knowledge Based perform below baseline.
  - Baseline - Most frequent sense of the manually tagged corpora.
    - In WordNet, senses are ranked according to the frequency of Semcor.

Outline

Domain Specific WSD

Domain Specific WSD

- Even baseline fails.
- Frequent sense may not be the frequent sense in the domain.
- Much harder problem.

Example

Bank

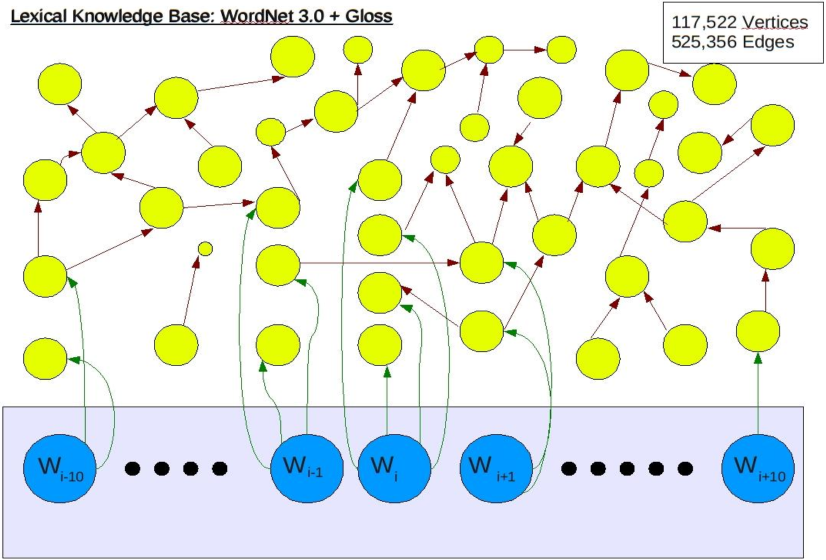
1. Sloping land beside a body of river.
2. financial institution

In the domain of *banking*, sense 2 is more frequent than sense 1.

Knowledge Based Methods

Graph Based systems

- PageRank Algorithm
  - Uses the hierarchial organization of WordNet
  - WordNet is converted to a graph each synset as a node (synset node) and the relationships in WordNet (hypernymy, hyponymy etc.) as edges between synset nodes.
- Personalized PageRank (PPR)
  - Uses contextual information.
  - State-of-art.


Lexical Knowledge Base: WordNet 3.0 + Gloss
117,522 Vertices
525,356 Edges
W_{i-10}
W_{i-1}
W_i
W_{i+1}
W_{i+10}

Outline

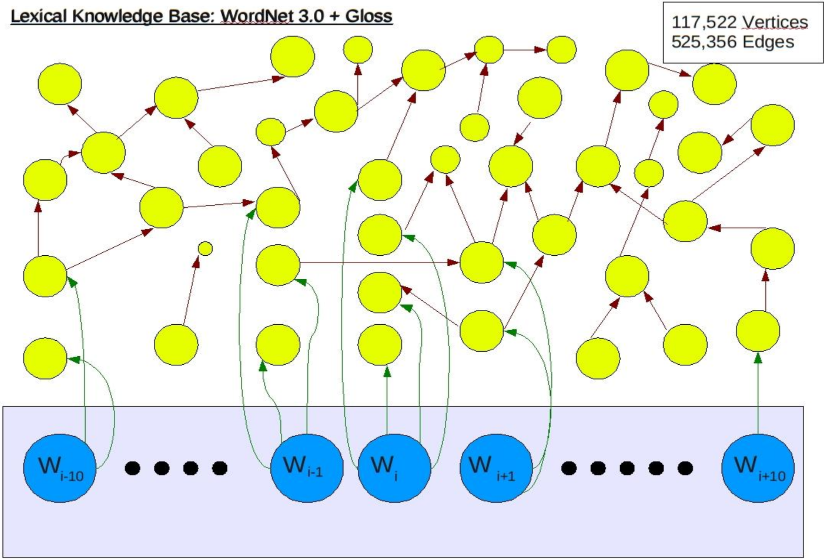
Domain Personalization

Normal personalization does not work for Domain Specific WSD.

- A need for Domain Specific Personalization.

Domain Specific Personalization

- Harness domain knowledge from the domain text corpora.
- Integrate this domain knowledge in a personalized page rank based WSD algorithm


Lexical Knowledge Base: WordNet 3.0 + Gloss
117,522 Vertices
525,356 Edges
W_{i-10}
W_{i-1}
W_i
W_{i+1}
W_{i+10}

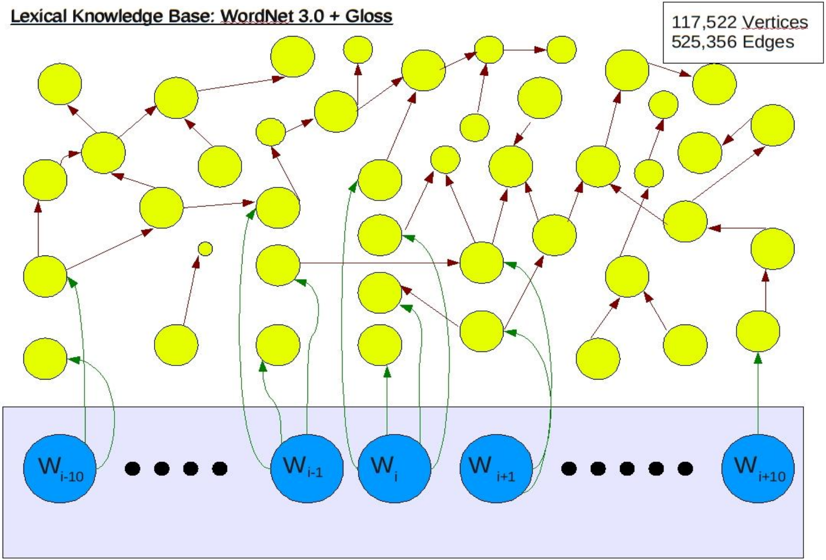
Domain knowledge

Keyword Ranking Scores (KRS)

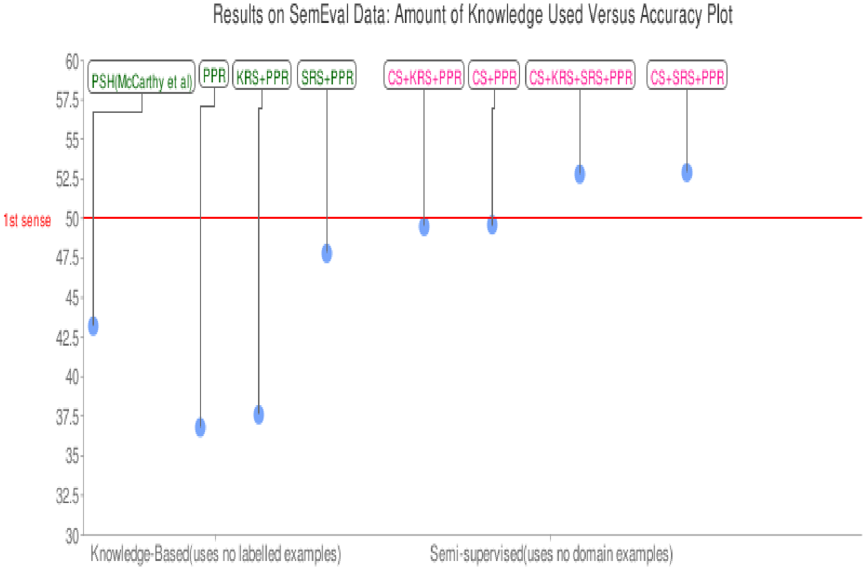
- Words salient in the domaim are better disambiguators.
- Compare frequency lists of Domain Corpora to General corpora.

Sense Ranking Scores (SRS)

- Senses in the given domain are ranked.
- We use McCarthy et al. (2004) to find sense distributions from raw text.
  - For every word, build its thesaurus entries from the raw text
    - Words behaving syntactically similar.
  - Use thesaurus entries to rank senses of a word.


Lexical Knowledge Base: WordNet 3.0 + Gloss
117,522 Vertices
525,356 Edges
W_{i-10}
W_{i-1}
W_i
W_{i+1}
W_{i+10}

Outline


Results on SemEval Data: Amount of Knowledge Used Versus Accuracy Plot
60
57.5
55
52.5
1st sense
50
47.5
45
42.5
40
37.5
35
32.5
30
PSH(McCarthy et al)
PPR
KRS+PPR
SRS+PPR
CS+KRS+PPR
CS+PPR
CS+KRS+SRS+PPR
CS+SRS+PPR
Knowledge-Based(uses no labelled examples)
Semi-supervised(uses no domain examples)

Summary

- **11%** improvement is observered on using sense ranking scores.
  - Big achievement
  - Placed 3^{rd} and 4^{th} in ACL SemEval 2010.
- Key word ranking did not yeild significant improvements
- Adding sense distributions from Semcor improved the overall performance
- Our unsupervised methods yet to beat the $1^{s}t$ sense baseline

Knowledge Sources for WSD

Information from many knowledge sources found to be useful.

- Part-of-speech
- Dictionary Knowledge
  - Sense Relatedness measures
- Domain
- Collocations or Context
- Structural information
  - Selectional preferences
  - Grammatical structures
- Discourse

Limitations of the existing methods

Most methods use information from one or two knowledge sources

Knowledge based methods like graph-based (Agirre and Soroa, 2009; Sinha and Milhalcea, 2007):

- Crucially rely on lexical knowledge base(LKB). Perform disambiguation using a graph based algorithm such as PageRank
- Edges represent binary relations like *is-a, type-of*
- Lack the ability to encode ternary and N-ary relations.

Limitations of the existing methods

Supervised methods (Yarowsky and Florian, 2002; Lee and Ng, 2002; Stevenson and Wilks, 2001)

- Mainly discriminative or aggregative models. Discriminative models aim to identify the most informative feature and aggregative models make their decisions by combining all features - feature vector models.
- Disambiguate word by word and not collectively and hence do not capture all the relationships (e.g sense relatedness) among all the words.
- Lack the ability to directly represent constraints like one sense per discourse.

Limitations of the existing methods

- Limitation due to the underlying method or framework

  Goal: *A framework that allows interaction of evidences from various knowledge sources to arrive at a global optimal solution*

Outline

Distributed Constraint Optimization Problem (DCOP)

Definition

A group of agents distributedly choose values for a set of variables such that the cost of a set of constraints over the variables is either minimized or maximized.

Distributed Constraint Optimization Problem

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

- A = $\{a_1, a_2, \ldots a_n\}$ is a set of n agents,
- V = $\{x_1, x_2, \ldots x_n\}$ is a set of n variables, each one associated to an agent,
- D = $\{D_1, D_2, \ldots D_n\}$ is a set of finite and discrete domains each one associated to the corresponding variable,

DCOP: Constraints and Objective funtion

- C = $\{f_k : D_i \times D_j \times \ldots D_m \longrightarrow \Re\}$ is a set of constraints described by various utility functions f_k. The utility function f_k is defined over a subset of variables V. The domain of f_k represent the constraints C_{f_k} and $f_k(c)$ represents the utility associated with the constraint c, where $c \in C_{f_k}$.
- F = $\sum_k z_k \cdot f_k$ is the objective function to be maximised where z_k is the weight of the corresponding utility function f_k

Outline

WSD as a DCOP

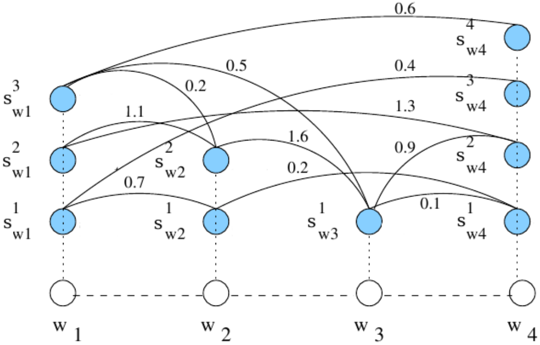
Given a sequence of words W= $\{w_1, w_2, \ldots w_n\}$ with senses of the word w_i, $D_{w_i} = \{s^1_{w_i}, s^2_{w_i} \ldots\}$, WSD is modelled as DCOP as follows.

WSD as a DCOP

- **Agents:** Each word w_i is modelled as an agent.
- **Variables:** Sense of a word is the variable. Each agent w_i is associated with the variable s_{w_i}.
- **Domains:** The set of senses D_{w_i}, is the domain of the variable s_{w_i}.
- **Utility function:** Information from each knowledge source is modelled as an utility function f_k.
- **Objective function:** An objective function F is defined over all the knowledge sources.
  - $F = \sum_k z_k \cdot f_k$
  - Words are assigned sense values such that F is maximized.

Semantic Relatedness as an utility function

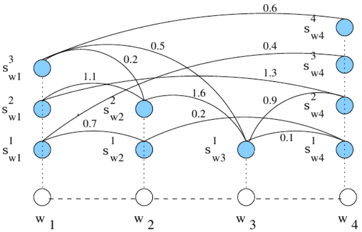
Words in a context must relate to each other to bring out a meaning.
Choose the senses which are maximally related.



[1]

[1] Image courtesy: (Sinha and Mihalcea, 2007)

Sense Relatedness as an utility function



For every pair of words w_i and w_j

- A utility function $f_{(w_i,w_j)} : D_{w_i} \times D_{w_j} \longrightarrow \Re$ is defined.
- $f_{(w_i,w_j)}$ returns sense relatedness (utility) between senses based on sense taxonomy and gloss overlaps defined in the WordNet.

Objective function over sense relatedness

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

Since we used only one type of knowledge source z_k=1

- $F = \sum_{(w_i,w_j)} f_{(w_i,w_j)}$
- DCOP assigns sense values to each word such that F is maximized i.e. a solution for which sense relatedness is maximized.

DCOP experimental details:

- Distributed Pseudotree Optimization Procedure (DPOP) algorithm to solve the DCOP.
- Open source toolkit FRODO[2].

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

Outline

Evaluation and Results

We performed disambiguation on SENSEVAL-2 and SENSEVAL-3 All words WSD data.

Senseval-3 All Words data set					
	noun	verb	adj	adv	all
P_dcop	**62.31**	43.48	**57.14**	**100**	54.68
R_dcop	**60.97**	**42.81**	**55.17**	**100**	**53.51**
F_dcop	**61.63**	**43.14**	**56.14**	**100**	**54.09**
P_Sinha07	61.22	**45.18**	54.79	**100**	**54.86**
R_Sinha07	60.45	40.57	54.14	**100**	52.40
F_Sinha07	60.83	42.75	54.46	**100**	53.60
MFS	69.3	53.6	63.7	92.9	62.3

Discourse

Discourse can be modelled as a n-ary utility function. For instance, *one sense per discourse*

- $w_i, w_j, \ldots w_m$ are the occurrences of a same word
- $f_{dis} : D_i \times D_j \times \ldots D_m \rightarrow \Re$
- f_{dis} returns higher utility when $s_{w_i} = s_{w_j} = \ldots s_{w_m}$

Objective function: $F = z_{dis} \cdot f_{dis} + \sum_k z_k \cdot f_k$

Summary

- We modelled WSD as a Distributed Constraint optimization problem.
- We also showed how to model sense relatedness.
- Results are slightly better than a PageRank based approach.
- We also describe modelling other knowledge sources.
- Future Directions
  - Perform disambiguation using mutliple knowledge sources
  - Determining the weight of each knowledge source.

Outline

Lessons learned from WSD

- Higher inter annotator agreements with coarse grained senses (Palmer, Fellbaum & Dang, 2001).
- WSD should not commit to a single sense. (Ramakrishnan et al., 2004; Erk & McCarthy, 2009).
- Coarse grained distinctions are relatively easier to learn and more useful in practical applications. (Ide, 2001)

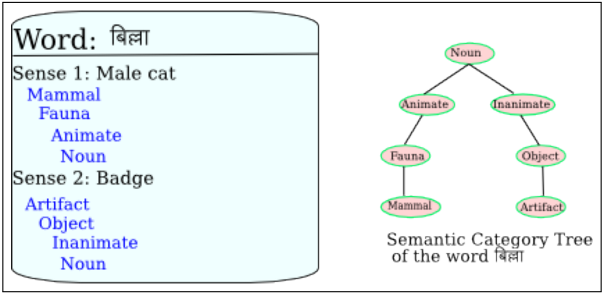
Problem Statement

Given a sequence of words we have to tag each word with it's corresponding semantic category.

1. *kuwwe/Dog* (*Mammal*) `ko` *xeKawe/seeing* (*NaturalEvent*) `hI` **billA/cat** (*Mammal*) *pedZa/tree* (*NaturalObject*) `para/on` `caDZa/climbed` (*VerbOfAction*) `gayA`
2. *saBA/Meeting* (*Event*) `meM/in` *Aye/came* (*VerbOfAction*) `saBI/all` *svayaMsevaka/volunteers* (*Group*) **billA/badge** (*Artifact*) `lagAye/wear` (*VerbOfState*) `hue` `We`

Where do the Semantic categories come from ?

- We use **Ontological Categories** of Hindi Wordnet.
- Hindi wordnet has **25,000** synsets and **168** Onto cats.
- Each sense of a word is mapped to some place in ontological hierarchy.



Applications of SCL

- Two features namely **Animate-Inanimate, Human-NonHuman** raised the parser accuracy. (Bharti et al., 2008)
- Machine Translation. (Vickrey et al., 2005;Carpuat & Wu, 2007)
- Information Extraction. (Resnik, 2006)
- Semantic Web

Outline

Lesk Algorithm

- Assign the sense whose gloss maximally overlaps with the gloss of neighbouring words.
- **Intuition:** Words appearing together in a sentence must inevitably be related in some way.

Example

The rate of interest at my bank is . . .
The glosses of *interest* and *bank* share *money* and *mortgage* and the glosses of *interest* and *rate* share *charge*.

Lesk Algorithm

Several variations of Lesk algorithm

- Simplified Lesk Algorithm
- Adapted Lesk Algorithm

*Underlying premise: Choose the sense which is **maximally related** to the words in context.*

Foundation to our work.

Lesk Algorithm

Several variations of Lesk algorithm

- Simplified Lesk Algorithm
- Adapted Lesk Algorithm

*Underlying premise: Choose the sense which is **maximally related** to the words in context.*

Foundation to our work.

Outline

Central Idea

Choose the semantic category which maximizes the semantic relatedness with its neighbours.

Our approach

- Lesk Algorithm combined with Semantic Relatedness Measures.
- Algorithm:
  - Each sense (semantic category) of the target word is taken
  - Its semantic relatedness is measured w.r.t the context word senses using relatedness measures.
  - Pick up the sense with highest score.
- We used 6 Semantic Relatedness measures and compare their results.

Leacock and Chodorow (lch), 1998

$$Rel_{lch} = -log\frac{length}{2D}$$

where length is the length of the shortest path between two concepts using node-counting, and D is the maximum depth of the taxonomy.

Wu and Palmer (wup), 1994

$$Rel_{wup} = \frac{2 * depth(LCS)}{depth(concept_1) + depth(concept_2)}$$

where LCS is the least common subsumer (LCS).

Lin, 1998

$$Rel_{lin} = \frac{2 * IC(LCS)}{IC(concept_1) + IC(concept_2)}$$

- IC is defined as:

$$IC(c) = -logP(c)$$

- P (c) is the probability of encountering an instance of concept 'c'
- Since we do not have tagged corpora, we assume "*each category of a word is equally likely*" and calculate P(c)
- We used a Hindi web corpora of size 324 MB to collect these statistics.

Jiang and Conrath, 1997

$$Rel_{jnc} = \frac{1}{IC(concept_1) + IC(concept_2) - 2 * IC(LCS)}$$

Lesk as Semantic Relatedness Measure

- Gloss overlap as semantic relatedness measure.
- Gloss of semantic category w.r.t a word
  - Gloss of all synsets of the word linked to it.

$$Rel_{lesk} = Overlap(gloss_{concept1}, gloss_{concept2})$$

Adapted Lesk as Semantic Relatedness Measure

$$Rel_{adpLesk} =$$

$$Overlap(extendedGloss_{concept1}, extendedGloss_{concept2})$$

Outline

Evaluation Data

- The data is developed by Indian language machine translation consortium(ILMT)
- Articles from news and tourism domain.
- 7200 manual annotated sentences covering 133 semantic categories
- The average semantic category ambiguity of a word is 2.18 and synset ambiguity is 2.57.

Results: Semantic Category Labelling

Model	*Precision*	*Recall*	*F-Measure*
Baseline	*84.76*	*84.76*	*84.76*
lch	*74.87*	*61.30*	*67.40*
wup	*75.33*	*61.68*	*67.82*
lin	*74.11*	*60.17*	*66.41*
jcn	*71.93*	*51.43*	*59.97*
lesk	*74.75*	*72.73*	*73.72*
adpLesk	*76.05*	*74.09*	***75.05***

Table: Evaluation of Semantic Category Labeling of Nouns

Results: Synset Disambiguation

Model	***Precision***	***Recall***	***F-Measure***
Baseline	*78.23*	*78.23*	*78.23*
lch	*67.14*	*54.98*	*60.45*
wup	*67.45*	*55.23*	*60.73*
lin	*65.05*	*52.81*	*58.29*
jcn	*62.52*	*44.70*	*52.12*
lesk	*65.27*	*63.51*	***64.37***
adpLesk	*63.36*	*61.73*	*62.53*

Table: Evaluation of Synset Assignment of Nouns

Summary

- Semantic Category labelling is easier than Synset labelling
- Baseline is tough to beat. (State of art unsupervised)
- **Adapted Lesk** performs better for semantic category labeling
- **Lesk** performs better for synset assignment.
- Why??

  Possible reason: When committing an error, adapted lesk assigns the synset closest to the actual one.

Summary

- Other relatedness measures perform worse. Why?
- Are these measures really useful for WSD?
- WordNet provide similarity kind of measures rather than relatedness measure.
  - **Orange:n** (sense:Fruit) is related with **Eat:v** (sense:Action of taking in).
  - We need cross pos links (like V-N links) apart from is-a, part-of relations.
- Why is performance of dictionary based methods lower than baseline?

Dream synset for WSD

Dream synset with the following properties

- Cross POS linkages
- Not coarse grained or fine grained sense. Just the right one.
- Adapts its rank according to the domain or the corpora at hand.
- It has lots of information in it
  - Collocations (words, frequency information, morph and syntactic patterns)
  - More corpus instances (Examples)
  - Topic Signatures (Domain Capture)

Related Work

- Lesk algorithm for Hindi WSD (Sinha et al., 2004)
- Lesk approach combined with similarity measures (Patwardhan et al., 2004)
- Graph based approaches with similarity measures to calculate edge weights. (Sinha and Mihalcea, 2007, Agirre et al. 2009]
- Bayesian model (Probabilistic approach) (Yarowsky, 1992)

Preliminaries

Example

I ate an orange.
Monkey is eating a banana.
She eats an apple everyday.

- **First order collocational features** of a word *W* describe the context of the word *w*.
- Feature templates used are *<sw>* , *<sw, pos(sw)>*, *<sw, pos(sw), pos(w)>* . **FOCF(eat)**= `{I, orange, monkey, banana, she, apple}`, given context size =2.

Preliminaries

Example

I ate an orange.
Monkey is eating a banana.
She eats an apple everyday.

- **First order collocational features** of a word *W* describe the context of the word *w*.
- Feature templates used are *<sw>* , *<sw, pos(sw)>*, *<sw, pos(sw), pos(w)>* . **FOCF(eat)**= `{I, orange, monkey, banana, she, apple}`, given context size =2.

Preliminaries

Example

I ate an orange.
Monkey is eating a banana.
She eats an apple everyday.

- **Second order collocates**: A word *x* is said to be second order collocate of *y* w.r.t feature *f*, iff the feature *f* is a first order collocational feature of *x* and *y*. `{orange, banana, apple}` are second order collocates of each other w.r.t feature $<$ **eat** $>$.

Outline

Our Unsupervised Approach

- At the core, all the unsupervised approaches exploit the redundancy in the data.
- Dekang Lin (Lin, 1997) uses the intuition that

  Two different words are likely to have similar meanings if they occur in identical local contexts.

- Our assumption is

  Two different words are likely to have similar semantic category if they have identical first order collocational features i.e. if they are second order collocates to each other.

Our Unsupervised Approach

- **Training Phase**
  - **Step 1:** Collect second order collocates w.r.t all the features present in the training corpus.
  - **Step 2:** Build training models from second order collocation sets. Aim of this step is to calculate the likelihood of a category *cat* given a feature.
- **Disambiguation Phase**
  - **Step 3:** Using the above training models for Semantic Category labeling.

- We propose two methods that differ in the way the probabilites are calculated.
- **Step 1** is common to both the methods.
- **Flat Semantic Category labeling:** assumes a flat list of categories
- **Hierarchical Semantic Category labeling:** exploits the hierarchical organization of these categories.

Step 1

- First order collocational features of all the words are collected using feature templates. The best templates are found to be
  1. (sw_k)
  2. (sw_k , posOf(sw_k) , posOf(sw_0))
- For every feature f_j in F, **S**econd **O**rder **C**ollocate sets w.r.t f_j, SOC_{f_j}, are calculated i.e. all the words which have feature f_j as first order collocational feature are collected.

FSCL: Training

- Aim of this step is to calculate the expectation of occurence of category *cat* with feature f_j. To calculate this, we use the following equations.

$$
\begin{aligned}
Pr(cat|f_j) &= \frac{Count(cat, SOC_{f_j})}{\sum_{cat} Count(cat, SOC_{f_j})} \\
AE(cat|f_j) &= \frac{Pr(cat|f_j)}{Pr(f_j)}
\end{aligned}
\tag{1}
$$

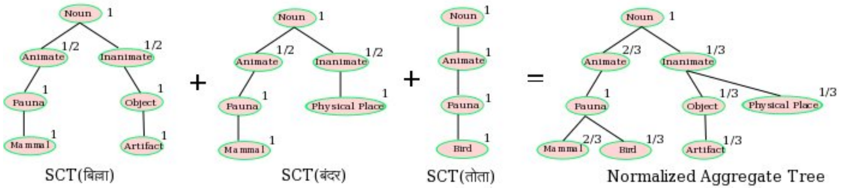
FSCL: Disambiguation Step

- This is the disambiguation phase where an utterance of a word w_i with leaf semantic categories $SC_{w_i} = \{c_1, c_2, \ldots\}$ is assigned a category according to the following equation.

$$\operatorname*{argMax}_{c_k \in SC_{w_i}} \sum_{j=1}^{F} AE(c_k | f_j)$$

- where $f_1, f_2, \ldots, f_F$ are the first order collocational features of w_i.

Hierarchical Semantic category labeler



- The example shows the second order collocates w.r.t feature **cadZa**.
- billa, wowa, bandar are the candidates.
- HSCL doesn't reflect take into account the global position of the category.
- Instead the scoring is based on number of siblings or the number of children it's parent has.

Advantages of HSCL:

- HSCL disambiguates level by level. Number of categories to be disambiguated in the top level are less compared to the number of leaves of the semantic category tree.
- No need of semantic similarity/relatedness measures.
- The nodes at top levels are shared by large number of words. This makes the learning effective for these nodes and hence the method takes better decesions at top levels.
- This can handle unseen category instances because the disambiguation proceeds in top down manner.
- This method can stop at a level which has high confidence score.

Outline

FSCL Results

Model	P	R
Baseline	85.6	85.6
FSCL trained on raw text	75.6	75.6
FSCL with k=2 trained on raw text	84.7	53.9
FSCL with k=3 trained on raw text	87.8	50.0
FSCL with k=2 trained on pos tagged text	83.2	63.4

Table: Accuracies of FSCL and Baseline for **nouns** (P: precision and R:recall)

HSCL Results

	Baseline		HSCL $(k = 5)$	
Level	P	R	P	R
1	96.9	96.9	99.4	94.0
2	91.5	91.5	96.4	63.8
3	89.8	89.8	95.4	52.0
4	87.7	87.7	94.4	46.4
5	76.8	76.8	83.1	64.4

Table: Level wise accuracies of HSCL for **nouns**

Summary

- Two Unsupervised Methods for Semantic Categoey labeling
    - You just need raw text
- A novel top-down approach.
- HSCL more reliable than FSCL.

Contribution of this thesis

- State-of-art method for Domain Specific WSD.
  - Reveals the importance of domain information.
- Multi Agent Framework for modelling various knowledge sources.
  - Powerful enough to model any knowledge source.
  - First of its kind in WSD.
- Introducing the problem of Semantic Category labeling (SCL)
  - Very relevant in the context of Indian Languages.
- Evaluation of various semantic relatedness measures for SCL
- Unsupervised method for SCL
  - A novel top-down approach contrast to the existing bottom-up approaches.