# An Extended Analysis of a Method of All-words Sense Disambiguation

Varada Kolhatkar
kolha002@d.umn.edu
Advisor: Dr. Ted Pedersen

Thesis defense, July 28, 2009

# Word Sense Disambiguation

Words in a natural language often have multiple senses.

*Sir William Walton was a British composer and* **conductor**.

*conductor* $\longrightarrow$ *the person who leads a musical group*
*conductor* $\longrightarrow$ *a substance that readily conducts electricity and heat*

- Humans are fairly adept in solving ambiguity by drawing on context and their knowledge of the world.
- Useful in various applications if software could distinguish between different senses of a word.
  - Examples: Machine Translation, Information Retrieval, Question Answering etc.
- **Word Sense Disambiguation** (WSD) is the process of selecting the correct sense of a word in given context.

# All-words Sense Disambiguation

- **All-words Sense Disambiguation (all-words)** is the process of disambiguating all words in a text.
- Why all-words?
  - Helps understand the overall meaning of a sentence.
  - Can be used more generally in the translation, searching or summarization of a text.
- Complex problem : Mapping between words and senses is many-to-many.
- Current state-of-the-art accuracy remains a long way off far from natural human abilities.

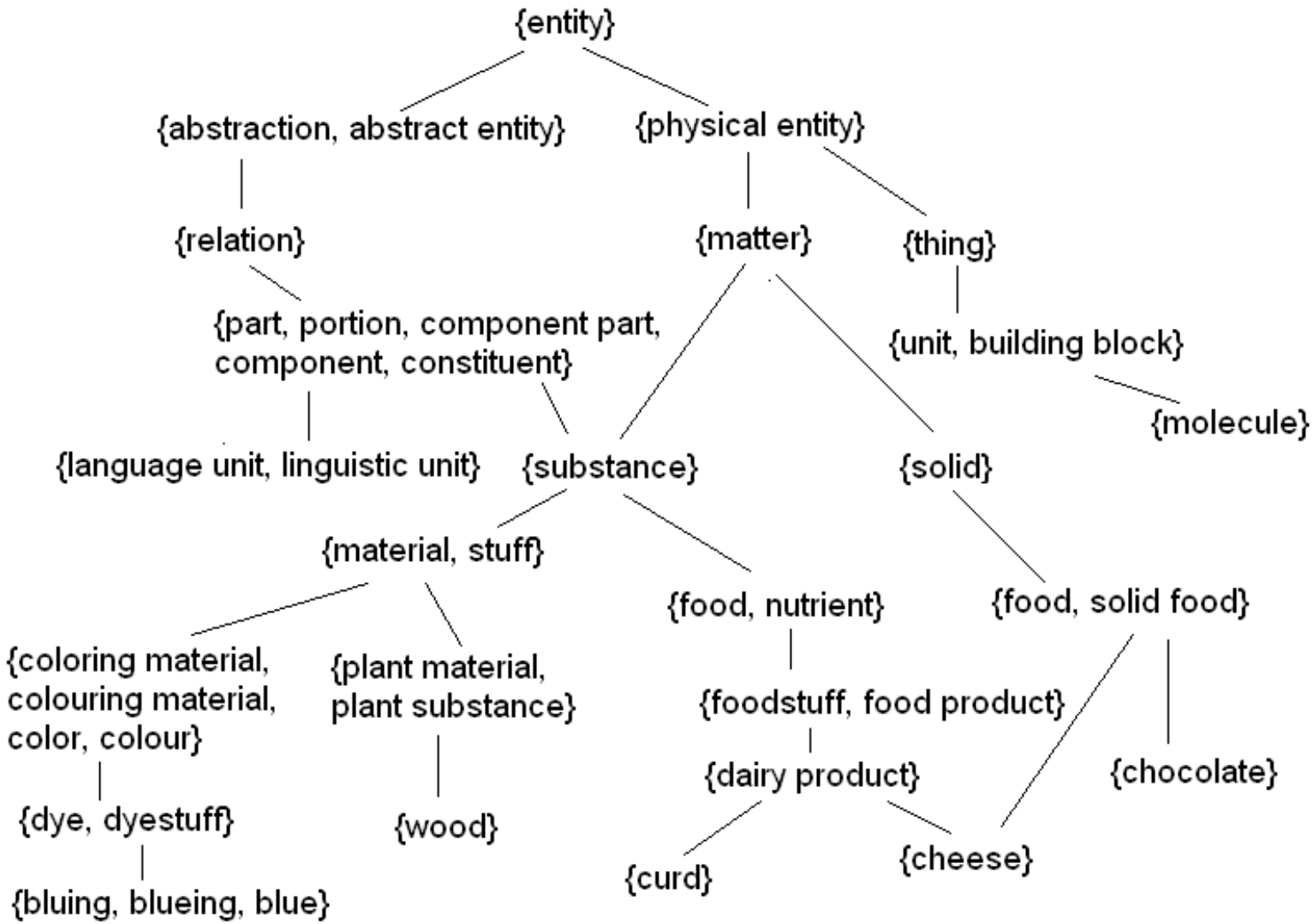
# Contributions of the Thesis

- The thesis starts by formalizing the algorithm of Michelizzi, 2005 for all-words.
- The time complexity is also examined.
- **The thesis presents our analysis of some of the components that might be contributing to the level of error currently plaguing all-words sense disambiguation.**
- Enhanced the method of Michelizzi in significant ways (via version 0.19 which is freely available on the Web).

# WordNet Overview

- Lexical database based on psycholinguistic principles.
- Contains only open class words.
  nouns (n), verbs (v), adjectives (a), adverbs (r)
- Concepts are organized in a semantic network.
- Nodes represent cognitively synonymous concepts called **synsets**.
  e.g. {conductor, music director, director} is a synset of concept 'the person who leads a musical group'.
- Edges represent relations between concepts.
- Separate network for each part of speech.
- The networks of nouns and verbs may be viewed as hierarchies.
- squash#n#1 means the first sense of the noun squash.

# WordNet *is-a* Hierarchy



# WordNet Overview

- Other relations include has-a, antonyms, pertaining-to, derived-from etc.
  - *ship* has-a *deck*.
  - *rich* is an antonym of *poor*.
  - *dental* pertains to *tooth*.
- About 155,287 words organized in over 115,000 synsets
- A total of 207,000 word-sense pairs.
- Contains about 117,700 nouns, 11,500 verbs, 21,400 adjectives and 4400 adverbs.
- Structure is well suited for the tasks where interpretation of a word based on its lexical semantics is required
- A very useful resource for the research in WSD.

# The Measures of Similarity and Relatedness

- Between which pair is the stronger relation?

  rose and flower or
  rose and calculator

- A variety of similarity and relatedness measures that exploit structure of WordNet

- Similarity Vs. Relatedness
  - *rich* and *poor* are related (antonyms), however they are not similar.
  - Similarity is limited to *is-a* hierarchies
  - Relatedness is more general and considers all relations.

# Path Based Measures

- Counting number of edges between two synsets.
- The greater the path-length, less similar the synsets are.
- Unfortunately not well suited where each node has different interpretation.
- Measures that use path-lengths that incorporate a variety of correcting factors
- Depth of the taxanomy [Leacock and Chodorow, 1994 (lch)]
- Depth of the least common subsumer [Wu and Palmer, 1994 (wup)]
- Relatedness measure, lexical chains [Hirst and St-Onge, 1998 (hso)]

# Information Content Based Measures

- Information Content ($IC$): Measure specificity of a concept.
  - general concept *entity*: low $IC$, specific concept *cheese*: high $IC$.
- $IC$ of the least common subsumer [Resnik, 1995 (res)].
- $IC$ to find semantic distance between concepts [Jiang and Conrath, 1997 (jcn)].
- concepts sharing a lot of specific information are more similar [Lin, 1997 (lin)]

# Gloss Based Measures

- Relatedness measures
- Can be applied to all parts-of-speech
- Extended gloss overlap (lesk) [Banerjee and Pedersen, 02]
  - Combines advantages of [Lesk, 86] gloss overlap with the structure of concept hierarchy
  - doesn't differentiate between a single word and a phrasal overlap
- Context vectors measure (vector) [Patwardhan and Pedersen, 03]
  - Gloss vector
    - Context vector formed by considering WordNet gloss as the context
  - relatedness is measured by measuring the cosine of the angle between the normalized gloss vectors.

# Overview

- Unsupervised knowledge based algorithm for all-words.
- Originally developed by Michelizzi, 2005.
- Assigns a WordNet sense to each content word in a sentence that is most related or similar to the surrounding words.
- Processes one sentence per line and one line per sentence.
- Formats
  - (raw) Sir William Walton was a British composer and conductor
  - (tagged) Sir_William_Walton#NNP was#VBD a#DT British#JJ composer#NN and#CC conductor#NN
  - (wntagged) Sir_William_Walton#n was#v a British#a composer#n and conductor#n
- If the format is raw, converts text into lower case and removes punctuation.
- Follow the steps below sequentially.

# Compoundify

*Input: the* **white house** *is the official residence of the president of the u.s.*
*Output: the* **white_house** *is the official residence of the president of the u.s.*

- Compounds are multi-word terms found in WordNet.
- Non-compositional meaning.
- If the format is raw, crucial to identify compounds for correct disambiguation.
- No combination of senses of *white* and *house* would imply *"the residence and office of the President of the United States."*
- 40% strings in WordNet are compounds.
- WordNet::Tools Perl module for compound identification.
- Greedy search to find the longest compound.

# Stop Words Removal

*Input:* *sir_william_walton was* **a** *british composer and conductor*
*Output:* *sir_william_walton was british composer and conductor*

- Stop words (prepositions, determiners etc.) are not included in WordNet.
- Only disambiguate content words i.e. words found in WordNet.
- Stop words are automatically excluded.
- Some commonly used stop words have unusual senses in WordNet.
  e.g. an : Associate in Nursing, AN - (an associate degree in nursing)
- Stoplist (a list of stop words) to eliminate commonly used stop words that have unusual senses in WordNet.
- Each content word occurring in the stoplist is not further considered for disambiguation.

# Lemmatization

*Input:* *sir_william_walton* **was** *british composer and conductor*
*Output:* *sir_william_walton* **be** *british composer and conductor*

- Obtaining the base forms of a word
  **be** is the lemma of **was**
- If *was* isn't lemmatized to *be*, WN-SRAW might consider
  Washington, WA – (a state in northwestern United States on the Pacific)
- Uses a simple lemmatization provided by WordNet::QueryData Perl module.
- Given a word or word#pos, it provides a list of all alternate forms (alternate spellings, conjugations, plural/singular forms, etc
- WordNet::QueryData also provides an interface to WordNet.

# Disambiguation

*Input:* *sir_william_walton be british composer conductor*
*Output:* *sir_william_walton#n#1 be#v#1 british#a#1 composer#n#1 conductor#n#1*

- Input of the algorithm is a sentence
  - which contains only content words (words found in WordNet),
  - in which compounds are identified,
  - in which stop words are eliminated,
  - which is lemmatized
- Each word is disambiguated separately
- Word being disambiguated is the **target** and the surrounding words form **window context**

# Window Size = $n$

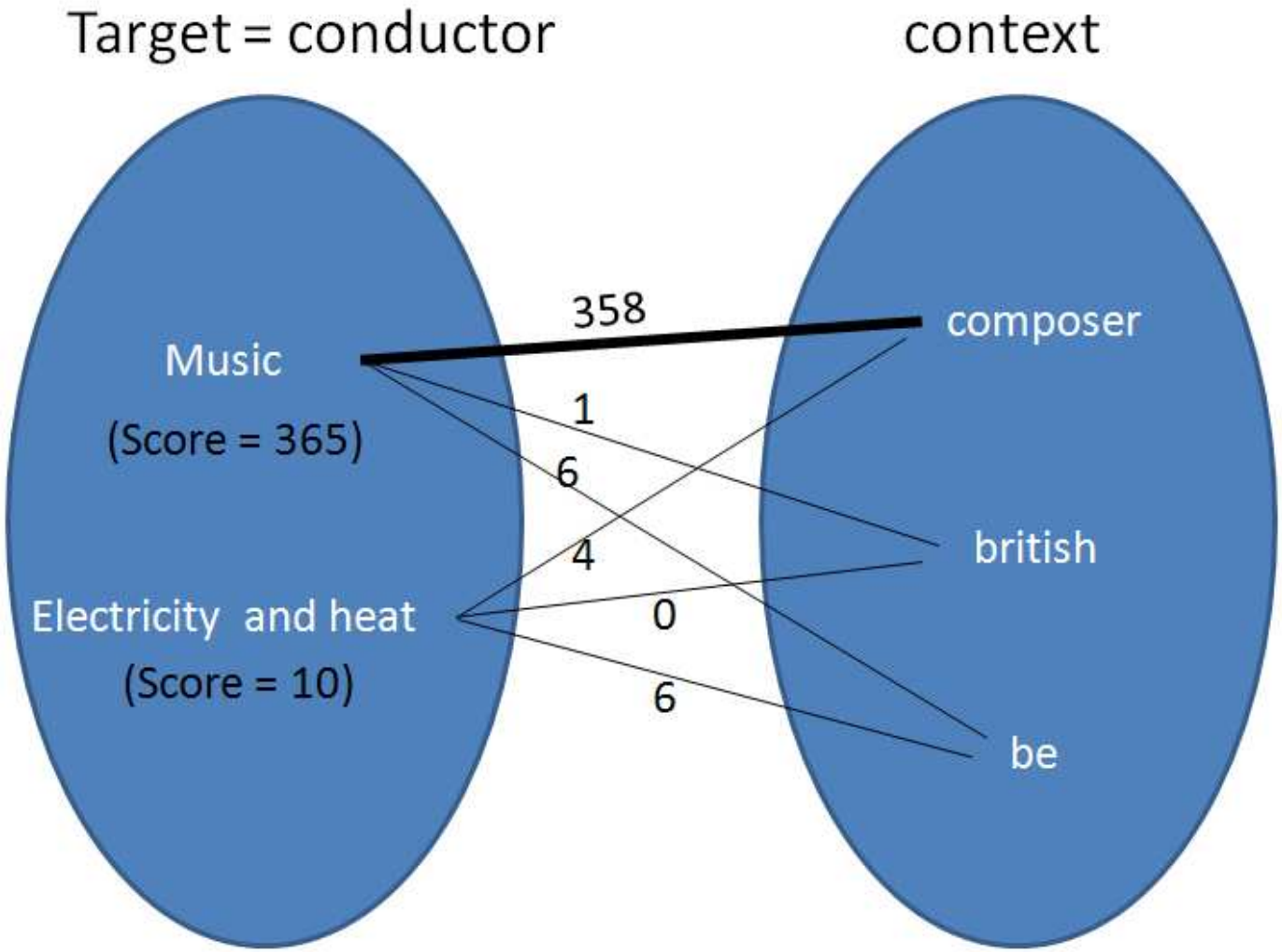
*sir_william_walton be british composer conductor*

- balance context according to the size $n$
- ceil$((n-1)/2)$ words on the left of the target
- floor $((n-1)/2)$ words on the right of the target

  *ceil(x) = smallest integer not less than x as a real number*
  *floor(x) = largest integer not greater than x as a real number*

- window = 3, target = british
  **context window = {be, composer}**
- window = 4, target = british
  **context window = {sir_william_walton, be, composer}**
- window = 7, target = conductor
  **context window = {composer, british, be}**

# Disambiguation$_1$




Higher weight $\rightarrow$ closely related, Lower weight $\rightarrow$ not really related

# Disambiguation$_2$

- Unfortunately context words also have multiple senses

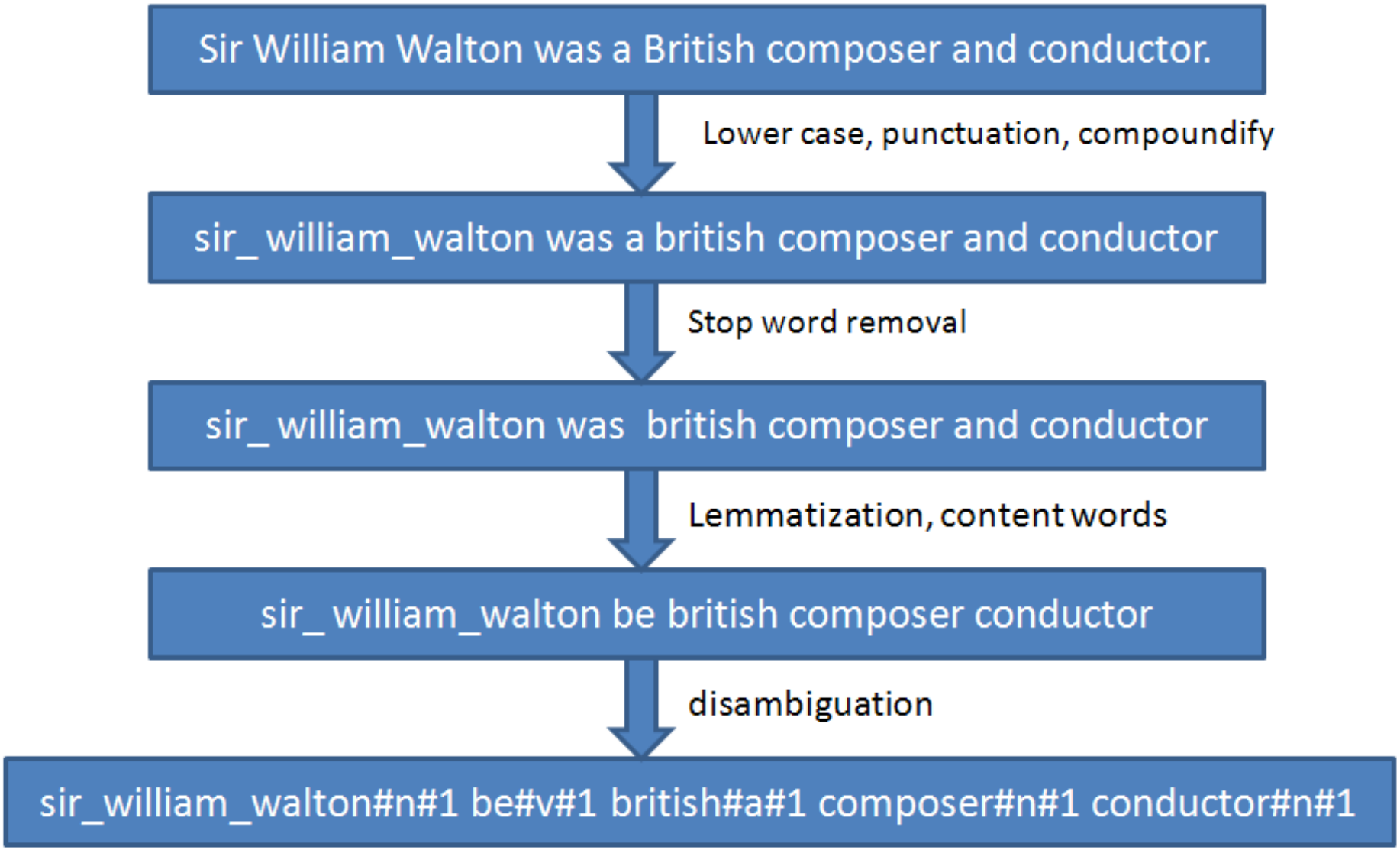
$$W(Music, british) = \max_{1 \leq \ell \leq \#senses(british)} (relatedness(music, british_\ell))$$

$british_\ell$ denotes $\ell^{th}$ sense of *british*.

- Each sense is assigned a score by summing the weights associated with its incoming edges

- The sense with the highest score is the winning sense.

# Algorithm Summary



# Definitions

**A monoseme**

A word or a phrase with a single meaning.

1. (12) Tuesday, Tues – (the third day of the week; the second working day)

**Polysemy**

Having or being characterized by multiple meanings.

I went **walking**. I went for **a walk**. I **walk the dog**. I took a **graduation walk**.

# Manually Sense-tagged Corpora

- Human annotators assign sense tags to each content word in a text using WordNet.
- Manually sense-tagging all words is a time consuming, expensive and error prone process.
- Involves learning senses of a number of words.
- Relatively less sense-tagged corpora available.

# SemCor

- Widely-used freely available manually sense-tagged corpus.
- Created at Princeton University.
- Comprises of $\approx$ 234,000 semantically annotated words.
  (80% Brown Corpus, 20% a novel, "The Red Badge of Courage")
- All open class words are manually tagged with WordNet 1.6 senses.
- WN-SRAW uses SemCor 3.0 that is compatible with WordNet 3.0.
  (translated by Rada Mihalcea)
- SemCor 3.0
  - 185,273 manually sense-tagged open class words.

# SensEval

- Competitions held in order to evaluate various WSD systems.
- Four competitions held so far.
  (SensEval-1 in 1998, SensEval-2 in 2001, SensEval-3 2004 and SensEval-4 in 2007)
- Includes a number of different tasks.
- Different types of data sets created for all-words task.

| corpus | nouns | verbs | adjectives | adverbs |
|---|---|---|---|---|
| SemCor | 87,002 (47%) | 47,570 (26%) | 31,754 (17%) | 18,947 (10%) |
| SensEval-2 | 1,057 (47%) | 509 (23%) | 417 (18%) | 277 (12%) |
| SensEval-3 | 884 (46%) | 719 (37%) | 322 (17%) | 12 (0.6%) |

# SensEval-2 and SensEval-3

- SensEval-2
  - Small subset of the Penn Treebank corpus.
  - Three Wall Street Journal articles.
  - 2,260 open class words found in WordNet.
- SensEval-3
  - Small subset of the Penn Treebank corpus.
  - Three articles
    (2 from Wall Street Journal, 1 is a work of fiction from Brown corpus).
  - 1,937 open class words found in WordNet.
- SensEval-1 didn't have an all-words task.
- SensEval-4 data would be an interesting data to work on.

# General Methodology and Evaluation Measures

General Methodology

- Performance is evaluated using manually sense-tagged corpora.
- Extract the key (the gold standard).
- Extract the part of speech tagged text from the corpus.
- Disambiguate the extracted text using WN-SRAW.
- Score the answers of WN-SRAW against the key.

Evaluation Measures

- Results are reported using precision (p) , recall (r) and F-score (F)
- $p = \frac{\#instances\ \ assigned\ \ correct\ \ senses}{\#attempted\ \ instances}$
- $r = \frac{\#instances\ \ assigned\ \ correct\ \ senses}{\#total\ \ instances\ \ in\ \ the\ \ corpus}$
- $F = \frac{2.p.r}{p+r}$

# Baselines

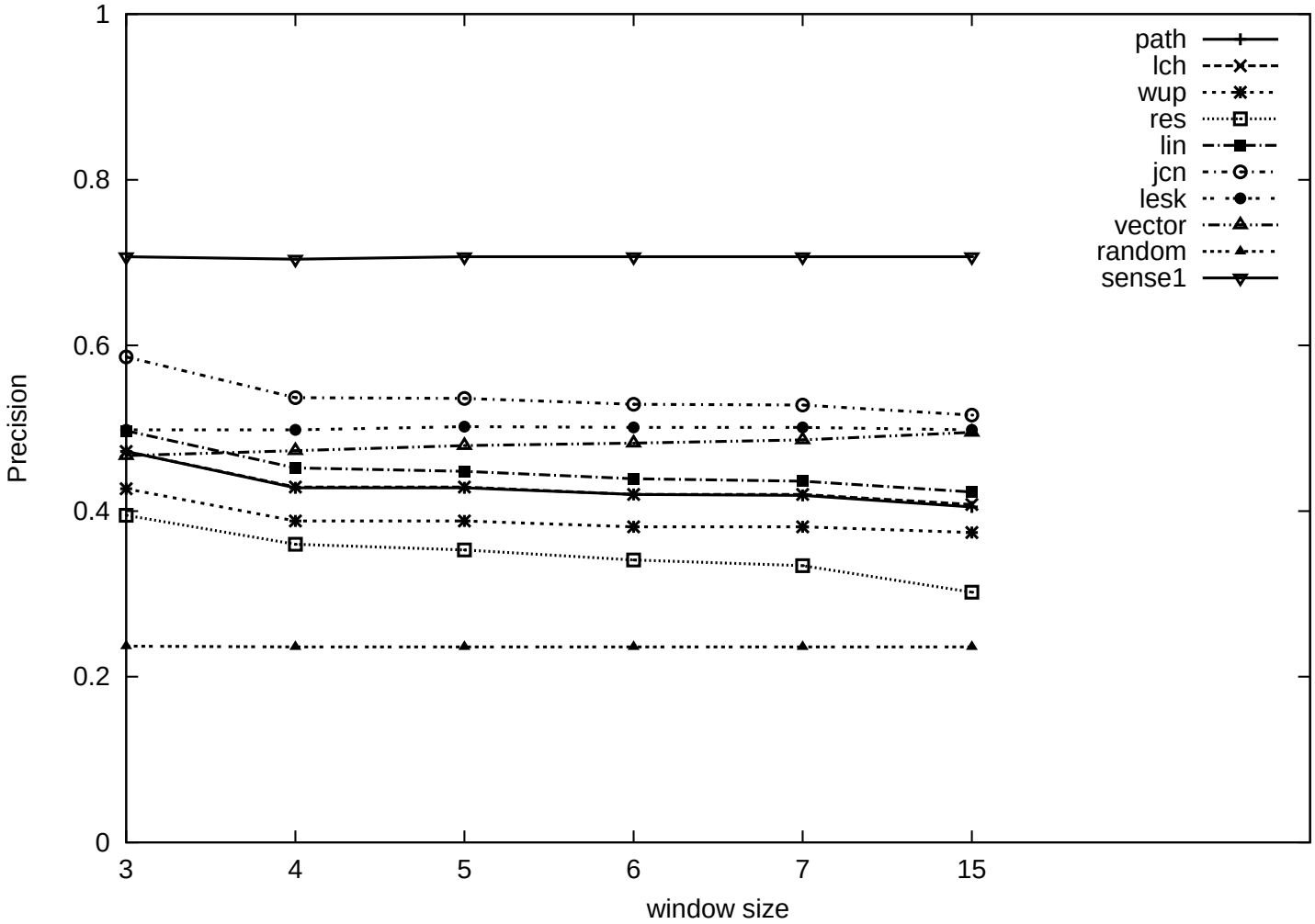
Random Scheme (lower bound)

- assignment of a random sense to each instance.
- done after lemmatization, leaving a relatively few senses from which to choose a random sense.

Sense1 Scheme (upper bound)

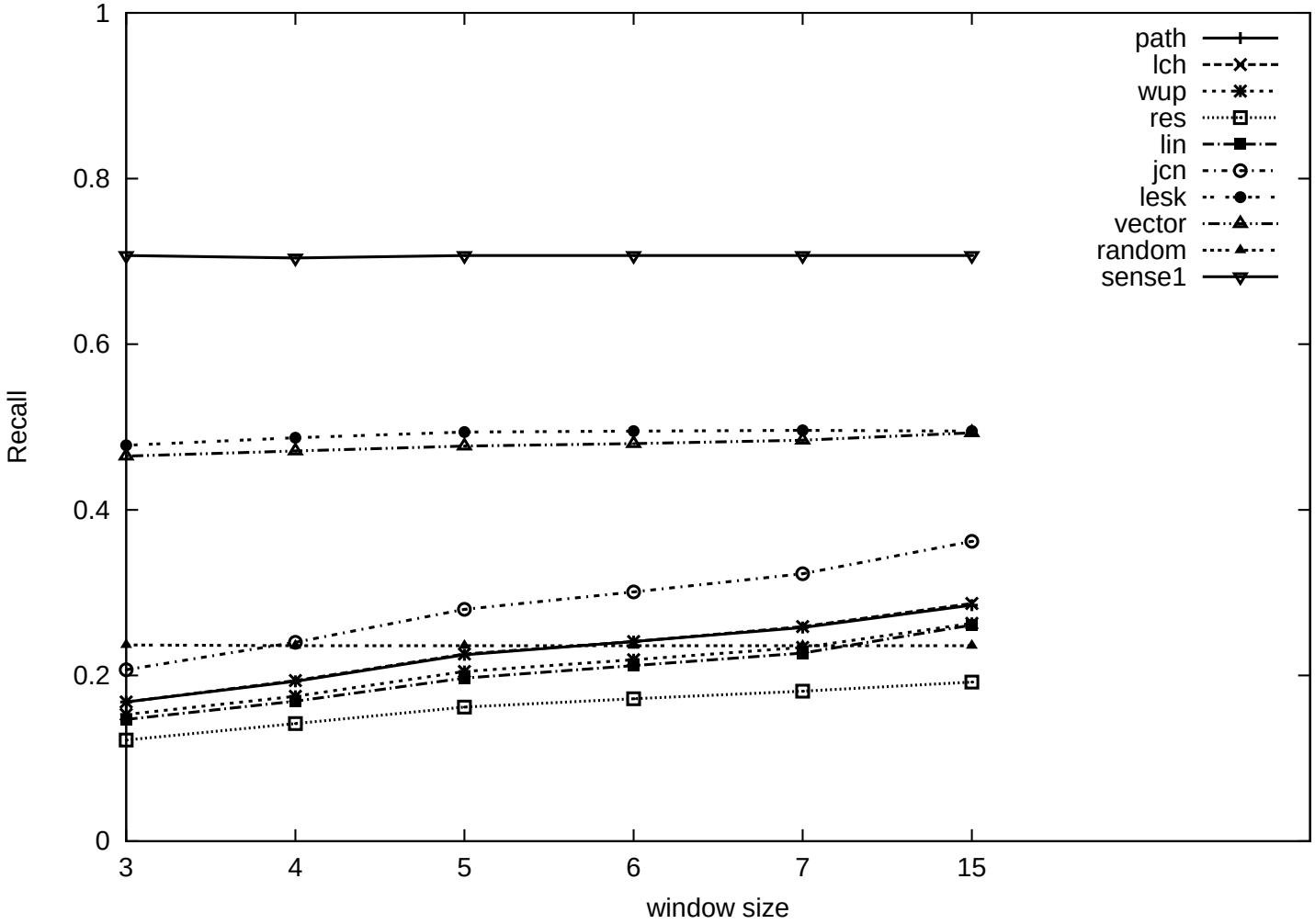
- WordNet senses are arranged according to their frequencies in SemCor.
- Assignment of sense1 in WordNet to all instances.
- Like a supervised system which uses information about distribution of senses.
- Works great for the available sense-tagged corpora.
- Won't generalize well for the text in a different domain.

# Expanding Context Window$_1$




- Precision decreases with increased window size.

# Expanding Context Window$_2$




- Recall, F-score increase with increased window size.

# Polysemy and Difficulty

| Polysemy | P | R | F | # instances |
|---|---|---|---|---|
| 1 | 1.000 | 1.000 | 1.000 | 28,673 (19.67 %) |
| 2 | 0.677 | 0.666 | 0.672 | 23,417 (16.06 %) |
| 3 | 0.680 | 0.673 | 0.677 | 25,525 (17.51 %) |
| 4 | 0.515 | 0.513 | 0.514 | 18,776 (12.88 %) |
| 5 | 0.473 | 0.470 | 0.471 | 13,210 (9.06 %) |
| 6 | 0.412 | 0.410 | 0.411 | 9,944 (6.82 %) |
| 7 | 0.381 | 0.379 | 0.380 | 9,056 (6.21 %) |
| 8 | 0.363 | 0.362 | 0.363 | 5,123 (3.51 %) |
| 9 | 0.329 | 0.328 | 0.328 | 4,726 (3.24 %) |
| 10 | 0.302 | 0.301 | 0.302 | 5,465 (3.75 %) |
| 11 | 0.351 | 0.347 | 0.349 | 5,437 (3.73 %) |
| 12 | 0.296 | 0.296 | 0.296 | 2,355 (1.62 %) |

- Spearman's rank correlation rho between Polysemy and **F = -0.820**.
- Polysemy is directly proportional to the difficulty of disambiguation.

# Results of Frequently Occurring Types

| word type | P | R | F | # Instances | Polysemy |
|---|---|---|---|---|---|
| be#v | 0.624 | 0.621 | 0.623 | 8,400 (4.5%) | 13 |
| person#n | 1.000 | 0.987 | 0.993 | 6,696 (3.6%) | 3 |
| not#r | 1.000 | 0.984 | 0.992 | 1,703 (0.91%) | 1 |
| group#n | 0.981 | 0.981 | 0.981 | 1,329 (0.71%) | 3 |
| have#v | **0.124** | **0.123** | **0.124** | 1,126 (0.61%) | 19 |
| say#v | **0.215** | **0.210** | **0.212** | 1,005 (0.54%) | 11 |
| location#n | 0.955 | 0.952 | 0.952 | 993(0.0.53%) | 4 |
| make#v | **0.085** | **0.085** | **0.085** | 757 (0.41%) | 49 |
| man#n | 0.674 | 0.672 | 0.673 | 576(0.31%) | 11 |
| see#v | **0.053** | **0.053** | **0.053** | 549 (0.29%) | 24 |
| know#v | **0.280** | **0.268** | **0.274** | 512 (0.28%) | 11 |
| time#n | **0.103** | **0.103** | **0.103** | 511 (0.%28) | 10 |

**Some frequently occurring types consistently perform poorly.**

# Tagged and Raw Format Experiments (lesk, window=5)

| | | brill tagged | raw |
|---|---|---|---|
| Nouns | P | 0.535 | 0.504 |
| | R | 0.525 | 0.501 |
| | F | 0.530 | 0.503 |
| Verbs | P | 0.389 | 0.313 |
| | R | 0.380 | 0.310 |
| | F | 0.384 | 0.311 |
| Adjectives | P | 0.541 | 0.422 |
| | R | 0.487 | 0.420 |
| | F | 0.513 | 0.421 |
| Adverbs | P | 0.436 | 0.283 |
| | R | 0.418 | 0.279 |
| | F | 0.427 | 0.281 |
| All | P | 0.484 | 0.419 |
| | R | 0.469 | 0.416 |
| | F | 0.476 | 0.417 |

**Knowing part-of-speech tag is helpful.**

# Best Performing Measures for Polysemous Instances

| POS | SemCor | SensEval-2 | SensEval-3 |
|---|---|---|---|
| Nouns | $jcn_{15}(0.574)$ | $vector_{15}(0.547)$ | $lesk_{15}(0.481)$ |
| Verbs | $vector_{15}(0.410)$ | $vector_{15}(0.342)$ | $vector_{15}(0.387)$ |
| Adj. | $lesk_{7}(0.582)$ | $lesk_{6}(0.597)$ | $lesk_{6}(0.494)$ |
| Adv. | $lesk_{7}(0.469)$ | $lesk_{7}(0.509)$ | - |

# Related Work

- Miller, et al., 1996
  - proposes benchmarks (random, sense1)
- Mihalcea and Faruque, 2004
  - minimally supervised system
  - uses parsing (syntax) and co-occurrences
- Navigli and Lapata, 2007
  - graph-based unsupervised algorithm, uses WordNet
  - correct sense is identified by using a variety of measures that analyze the connectivity of graph structures.
- Preiss, et al., 2009
  - starts with sense1 and tries to refine it
  - supervised, uses ranking algorithm and a Wikipedia similarity measure.

# Conclusion

The experimental results provide evidence in favor of the following hypotheses

- The degree of difficulty in disambiguating a word is proportional to the number of senses of that word (polysemy).
- A significant percentage of word sense disambiguation error is caused by just a few highly frequent word types.
- Part-of-speech tagged text will be disambiguated more accurately than raw text.

Other Observations

- Expanding the context window around a polysemous target word improves recall significantly, but lowers precision suggesting that expanding the context may add significant noise.

# Future Work

- Flexible context selection
  - varying the context selection according to the situation.
  - Avoiding context words that might add noise.
- Incorporating the notion of syntax.
- Performance in terms of time
  - storing similarity scores.
  - parallelize the disambiguation.
- Combination method that combines best performing measures based on part-of-speech.

# WordNet::SenseRelate::AllWords

- **Project**
  `http://search.cpan.org/dist/WordNet-SenseRelate-AllWords/`
- **Web Interface**
  `http://talisker.d.umn.edu/allwords/allwords.html`
- **Data**
  `http://www.cse.unt.edu/~rada/downloads.html`
- **Publication**
  `http://www.d.umn.edu/~tpederse/Pubs/pedersenk09-demo-final.pdf`

THANKS!