

Accurate and Scalable Word Sense Disambiguation in the Biomedical Domain

Bridget T. McInnes

January 19, 2007

What is WSD?

The task of identifying the appropriate sense of a word that has multiple senses.

Example: *bat*

Sense 1: The flying mammal that only comes out at night

Sense 2: Something you hit a baseball with

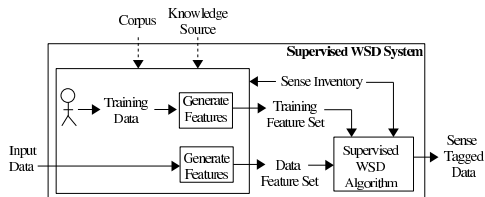
Why is WSD important?

- ▶ Text-to-speech
- ▶ Machine Translation
- ▶ Information Retrieval
- ▶ Concept Mapping ([our research focus](#))

4 Types of WSD systems

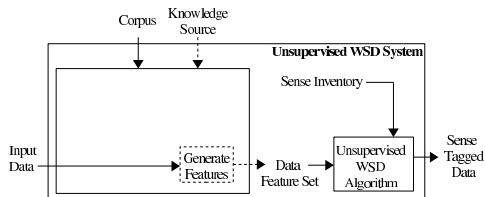
- ▶ Supervised WSD System
- ▶ Unsupervised WSD System
- ▶ Knowledge-based WSD System
- ▶ Bootstrapping WSD System

Supervised WSD System



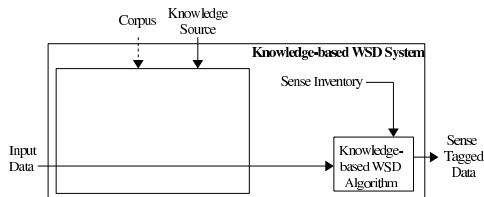
- ▶ **Relys:** on a manually annotated corpus
- ▶ Module generating feature sets
- ▶ Algorithm assigning a sense to each instance of a target word
- ▶ May be augmented by an external corpus and/or knowledge source
- ▶ **Problem:** not scalable due to training data

Unsupervised WSD System



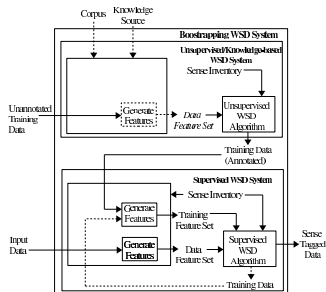
- ▶ **Relys:** on “distributional characteristics of an unannotated corpus”
- ▶ Module generating feature sets
- ▶ Algorithm assigning a sense to each ambiguous word
- ▶ Requires an external corpus
- ▶ May be augmented by an external knowledge source
- ▶ **Problem:** not as accurate as a supervised WSD system

Knowledge-based WSD System



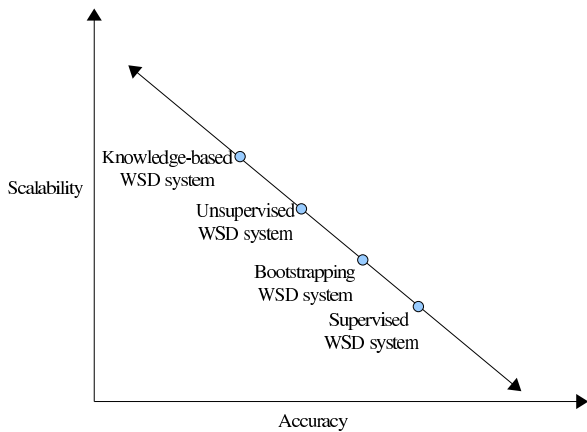
- ▶ **Relys:** on information from an explicit knowledge source
- ▶ Algorithm assigning a sense to each ambiguous word
- ▶ Requires an external knowledge source
- ▶ May be augmented by an external corpus
- ▶ **Problem:** not as accurate as a supervised WSD system

Bootstrapping WSD System

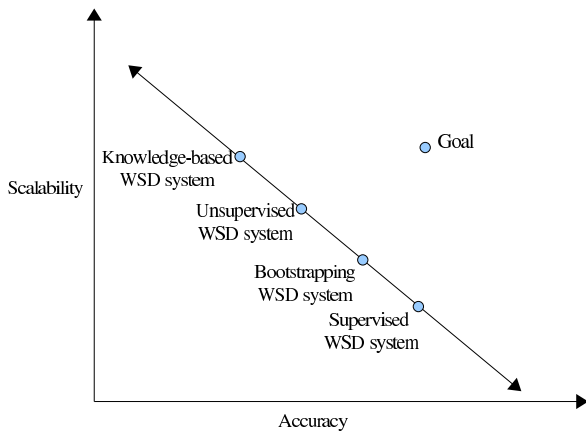


- ▶ **Combines:** a unsupervised or knowledge-based WSD system and a supervised WSD system
- ▶ Training data is (semi-) automatically generated by an unsupervised or knowledge-based system then fed into a supervised system
- ▶ **Problem:** not scalable due to training data

Recap



Thesis Goal



NLM-WSD Data set

- ▶ Created by the National Library of Medicine
- ▶ Contains 50 highly frequent ambiguous Unified Medical Language System (UMLS) concepts from the 1998 MEDLINE abstracts
- ▶ 100 ambiguous instances of each target word
- ▶ Instances were manually disambiguated by 11 evaluators who assigned the target word to a UMLS Concept or assigned the sense as “None” if none of the UMLS concepts described the sense
- ▶ 15 out of the 50 terms whose majority sense is less than 65%
 - ▶ Those terms are used in our preliminary work.

Related Work

- ▶ Supervised WSD systems: 14 papers
- ▶ Unsupervised WSD systems: 10 papers
- ▶ Knowledge-based WSD systems: 34 papers
- ▶ Bootstrapping WSD systems: 2 papers

- ▶ **Totaling: 60 papers**

Leroy and Rindflesch (2004)

- ▶ Algorithm: Naive Bayes
- ▶ Features:
 - ▶ Headword
 - ▶ POS
 - ▶ Semantic Types
 - ▶ Semantic Relations
- ▶ Data set: NLM-WSD

Joshi, Pedersen and Maclin (2005)

- ▶ Algorithm: SVM
- ▶ Features:
 - ▶ Unigrams
 - ▶ abstract level
 - ▶ sentence level
 - ▶ Bigrams
 - ▶ abstract level
 - ▶ sentence level
- ▶ Data set: NLM-WSD

Preliminary Work Research Questions

- Q1: Would using the *UMLS Concepts* of the surrounding words be an improvement over using the unigrams or semantic types of the surrounding words?
- Q2: Would *Journal Descriptors* contain enough information to disambiguate words in biomedical text using a supervised WSD system?
- Q3: Would changing the *algorithm* change the accuracy of the results if the feature set remained the same?

Defining UMLS Concepts

- ▶ Defined as the “meaning” of a word in the UMLS
- ▶ A given word may have more than one concept
 - ▶ e.g. *mole*
 - ▶ The mammal
 - ▶ Benign melanocytic nevus
- ▶ Multiple words may have the same concept
 - ▶ e.g. *exocytosis* and *secretion*
 - ▶ Cellular release of material membrane-limited vesicles by fusion of the vesicles with the cell membrane

Q1: Comparison using UMLS Concepts

Leroy and Rindflesch (2004)

- ▶ Algorithm: Naive Bayes
- ▶ Best performing feature set:
 - ▶ LeroyR04
 - ▶ Head word of the target word
 - ▶ POS of the target word
 - ▶ Semantic types of the surrounding words

Joshi, Pedersen and Maclin (2005)

- ▶ Algorithm: SVM
- ▶ Best two performing features sets:
 - ▶ Unigrams at the sentence level
 - ▶ Unigrams at the abstract level

Leroy and Rindflesch (2004) Comparative Results

Table: UMLS Concept Results using Naive Bayes

	baseline	LeroyR04	Concept	LeroyR04 + Concept
<i>overall average</i>	55	65.60	70.13	73.27
versus baseline	–	+10%	+15% ($p \leq .05$)	+18% ($p \leq .05$)
versus LeroyR04	–	–	+5%	+8%
versus Concept	–	–	–	+3%

Joshi, Pedersen and Maclin (2005) Comparative Results

Table: UMLS Concept Results using SVMs

	baseline	s-unigram	a-unigrams	Concept
<i>overall average</i>	55	73.47	76.07	68.27
versus baseline	–	+18% ($p \leq .001$)	+21% ($p \leq .001$)	+13% ($p \leq .001$)
versus a-unigram	–	–	+3%	-5%
versus s-unigram	–	–	–	-8%

Q1: UMLS Concept Results

- ▶ Introducing a new feature UMLS Concept
 - ▶ Significantly increases the baseline results by 15%.
 - ▶ Increases results reported by Leroy and Rindflesch (2004) but not significantly
 - ▶ Lower than the results reported by Joshi, Pedersen and Maclin (2005) but not significantly

Defining Journal Descriptors

- ▶ Manually assigned tag to the journals in the National Library of Medicines' List of Serials Indexed for Online Users.
- ▶ Currently 123 Journal Descriptors
 - ▶ *Cardiology*
 - ▶ *Surgery*
- ▶ A journal may have more than one descriptor
- ▶ All citations inherit the Journal Descriptors of their respective journals

Q2: Journal Descriptors Results

Table: Journal Descriptor Results using Naive Bayes

	baseline	JD	JD-subset
<i>overall average</i>	55	51.40	86.40

► Conclusions

- Has potential but can not say how much
- Due to the limited number of journals in our data set that have a Journal Descriptors.

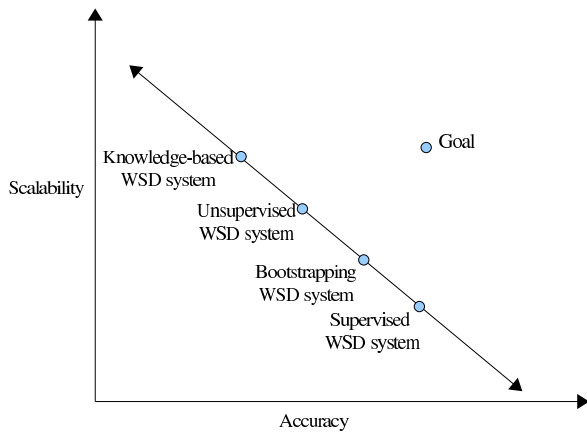
Q3: Algorithm Results

Table: Overall Average for Naive Bayes and SVM

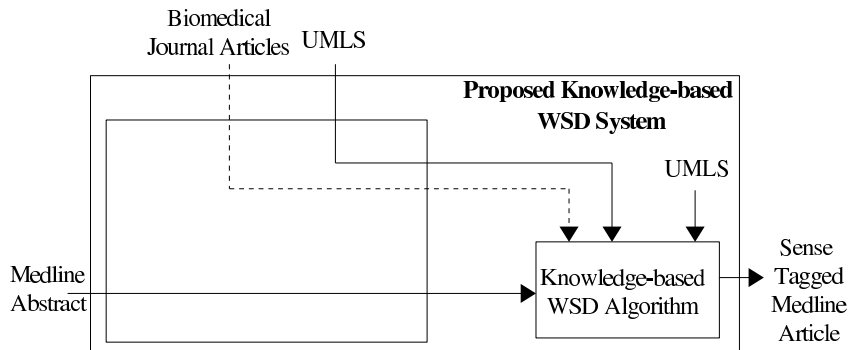
feature set	Naive Bayes	SVM
bag-of-words	69.93	66.47
head	58.47	56.80
POS	56.67	56.57
head+POS	58.20	57.40
ST	61.93	59.47
Concept	70.13	68.27
ST+Concept	72.53	68.40

- ▶ Comparing the Naive Bayes and SVM algorithms
 - ▶ Naive Bayes obtained a higher accuracy than SVM
 - ▶ The results are not significant

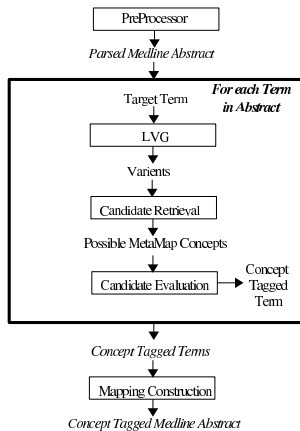
Proposed Work



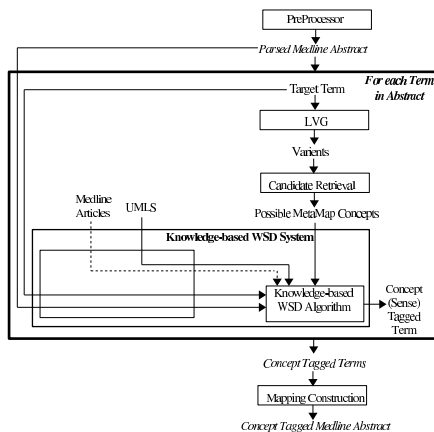
Proposed System



MetaMap



WSD-enhanced MetaMap



Evaluation Plan of Proposed Work

- ▶ Run our proposed system on the NLM-WSD data set
 - ▶ Sense Inventory: NLM-WSD data set
 - ▶ Compare with
 - ▶ MetaMap
 - ▶ Supervised WSD systems
 - ▶ Knowledge-based WSD systems

- ▶ Sense Inventory: MetaMap Candidate Retrieval module
- ▶ Compare with
 - ▶ MetaMap

Ways Forward: External Knowledge Questions

- Q1: Would UMLS Concepts of surrounding words *frequently co-occurring* with the target word help?
- Q2: Would UMLS Concepts of surrounding words *highly associated* with the target word help?
- Q3: Would UMLS Concepts of the surrounding words *related* to the target word help?
- Q4: Would Semantic Types of surrounding words that are either *highly related*, *highly associated* or *frequently co-occur* with the target word help?
- Q5: How well would Journal Descriptors perform as a *coarse-grained* filter?

Ways Forward: Algorithm

- ▶ first step forward toward testing our above questions
 - ▶ vector-based knowledge-based WSD system

- ▶ second step
 - ▶ investigate other ways of determining the distance between vectors

Contributions

- ▶ Intended Contribution
 - ▶ An accurate and scalable word sense disambiguation system for the biomedical domain

- ▶ Four potential applications, if successful
 - ▶ Text-to-speech
 - ▶ Machine Translation
 - ▶ Information Retrieval
 - ▶ Concept Mapping