Page Rank for Word Sense Disambiguation

B Arjun Shubhangi Ghosh

Department of Computer Science
Indian Institute of Technology Madras

NLP Paper Presentation, 2018

Paper to be discussed

Knowledge-Based WSD on Specific Domains: Performing better than Generic Supervised WSD

Eneko Agirre and **Oier Lopez de Lacalle** and **Aitor Soroa**
Informatika Fakultatea, University of the Basque Country

Outline

Why WSD?

Word Sense Disambiguation has been a central topic of research in NLP for years. WSD is a key step to approach language understanding. WSD has many applications such as:

- Parsing
- Machine Translation
- Information Retrieval
- Question Answering

Supervised WSD

- Performs well with sufficient training data.
- Requires hand-annotated corpus.
  Words need to be tagged with their correct sense.

Supervised WSD

The two supervised algorithms used for comparison with our approach are:

- **k-NN:**
  - Memory-based learning method.
  - For each test instance, *k* most **similar** train instances are found.
  - **Similarity** is measured as cosine of their feature vectors.
  - Maximum for sum of weighted votes of nearest neighbours is used to predict sense.
  - k=5 is used.

Supervised WSD

The two supervised algorithms used for comparison with our approach are:

- **k-NN:**
  - Memory-based learning method.
  - For each test instance, k most **similar** train instances are found.
  - **Similarity** is measured as cosine of their feature vectors.
  - Maximum for sum of weighted votes of nearest neighbours is used to predict sense.
  - k=5 is used.
- **Linear SVM**
- Standard features were used for training such as: Local Collocations, Syntactic Dependencies and Bag-of-Words

Knowledge-Based WSD

- Uses Lexical Knowledge Base such as WordNet.
- Doesn't require hand-annotated corpus with correct word sense.

Word Sense Disambiguation

- **Lexical Sample Exercises:**
  Sense of a particular word has to be disambiguated given correct senses of all context words.

Word Sense Disambiguation

- **Lexical Sample Exercises:**
  Sense of a particular word has to be disambiguated given correct senses of all context words.
- **All-Words Exercises:**
  - Disambiguate sense of all words in a running text.
  - Training data per word is much scarcer.
  - Supervised algorithms for WSD are typically trained on SemCor.

Flaws of Supervised WSD:

Supervised WSD gives only a small improvement over the **Most Frequent Sense(MFS)** baseline for **All-words exercises.**

- **Sparseness:**
  - Relatively small amount of training data available.
  - Sequence of words less likely to appear repeatedly as opposed to one word.

Flaws of Supervised WSD:

Supervised WSD gives only a small improvement over the **Most Frequent Sense(MFS)** baseline for **All-words exercises.**

- **Sparseness:**
  - Relatively small amount of training data available.
  - Sequence of words less likely to appear repeatedly as opposed to one word.
- **Corpus Mismatch:**
  - Supervised WSD applied to a different corpus than the one they were trained on.
  - **Solution:**
    Hand-tag examples from every new domain, but infeasible in practice.

Advantages of Knowledge-Based WSD:

- Doesn't require large amount of hand-tagged data.

Advantages of Knowledge-Based WSD:

- Doesn't require large amount of hand-tagged data.
- Performance is better on domain-specific data since it uses a Lexical Knowledge Base and thus avoids problems of corpus mismatch.

Advantages of Knowledge-Based WSD:

- Doesn't require large amount of hand-tagged data.
- Performance is better on domain-specific data since it uses a Lexical Knowledge Base and thus avoids problems of corpus mismatch.
- Exploits structural properties of the graph underlying a LKB, such as WordNet.

Advantages of Knowledge-Based WSD:

- Doesn't require large amount of hand-tagged data.
- Performance is better on domain-specific data since it uses a Lexical Knowledge Base and thus avoids problems of corpus mismatch.
- Exploits structural properties of the graph underlying a LKB, such as WordNet.
- Can be extended to any language if it has a well-developed WordNet.

Prior Approaches to Knowledge Based WSD:

DISTRIBUTIONAL THESAURUS - **Koeling et al., 2005**

- **Distributional Thesaurus** of related words constructed from untagged corpus.
  All words which share similar context are marked as related words for a particular word. This is done using co-occurence counts.
- Wordnet Pairwise similarity is evaluated with all related words to obtain a **Most Predominant Sense** for a given word.
- This method gives very little improvement over MFS, since like MFS this also uses one sense of the word to tag it over all instances in the corpus.

Prior Approaches to Knowledge Based WSD:

GRAPH CENTRALITY MEASURES:
Sinha and Mihalcea, 2007
Navigli and Lapata, 2007

- Graph Centrality measures are used to identify **Most Predominant Sense** for a given word from a LKB(Lexical Knowledge Base).
- Again improvement over MFS(Most Frequent Sense) is not very high.

Motivation

- Consider the following sentence to be disambiguated:

Motivation

- Consider the following sentence to be disambiguated:
- *The* **bank** *can guarantee* **deposits** *will eventually cover future tuition costs.*

Motivation

- Consider the following sentence to be disambiguated:
- *The* **bank** *can guarantee* **deposits** *will eventually cover future tuition costs.*
- Clearly the words bank and deposits are both ambiguous, as bank may refer to the financial institution sense or the land alongside a river, and deposits may refer to a sum of money or the act of placing something in a specific place.

Motivation

- Consider the following sentence to be disambiguated:
- *The* **bank** *can guarantee* **deposits** *will eventually cover future tuition costs.*
- Clearly the words bank and deposits are both ambiguous, as bank may refer to the financial institution sense or the land alongside a river, and deposits may refer to a sum of money or the act of placing something in a specific place.
- Thus there is an inherent circularity in the WSD problem.

Motivation

- Consider the following sentence to be disambiguated:
- *The* **bank** *can guarantee* **deposits** *will eventually cover future tuition costs.*
- Clearly the words bank and deposits are both ambiguous, as bank may refer to the financial institution sense or the land alongside a river, and deposits may refer to a sum of money or the act of placing something in a specific place.
- Thus there is an inherent circularity in the WSD problem.
- We also note that we can resolve the ambiguity by making use of the presence of words like "costs", which is not ambiguous. Also this sense reinforces a particular sense for both "bank" and "deposits", which in turn reinforce each other.

Motivation

- Consider the following sentence to be disambiguated:
- *The* **bank** *can guarantee* **deposits** *will eventually cover future tuition costs.*
- Clearly the words bank and deposits are both ambiguous, as bank may refer to the financial institution sense or the land alongside a river, and deposits may refer to a sum of money or the act of placing something in a specific place.
- Thus there is an inherent circularity in the WSD problem.
- We also note that we can resolve the ambiguity by making use of the presence of words like "costs", which is not ambiguous. Also this sense reinforces a particular sense for both "bank" and "deposits", which in turn reinforce each other.
- If we model the words and its senses as nodes in a graph, this clearly translates to identifying the importance of the senses in the graph, while making use of semantic relations between the senses. **PageRank** algorithm is known to be useful when such a **circularity** problem arises.

Page Rank

Consider the WordNet graph consisting of a set of vertices and edges. G =(V,E).

- The synsets or concepts are the vertices.
- Relationships between synsets such as causality, entailment, hyponymy, meronymy etc.

Page Rank

- Whenever a link from v_i to v_j exists on a graph, a vote from node i to node j is produced, and hence the rank of node j increases.

Page Rank

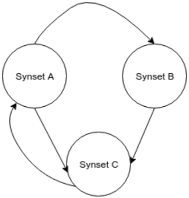
- Whenever a link from v_i to v_j exists on a graph, a vote from node i to node j is produced, and hence the rank of node j increases.
- The strength of the vote from i to j also depends on the rank of node i.

Page Rank

- Whenever a link from v_i to v_j exists on a graph, a vote from node i to node j is produced, and hence the rank of node j increases.
- The strength of the vote from i to j also depends on the rank of node i.
- $$\mathbf{PR} = cM\mathbf{PR} + (1-c)\mathbf{v}$$

PR - The importance scores of the synsets.
v - The initial probability mass over synsets.
c - Damping factor (scalar value between 0 and 1)
M - N x N transition probability matrix, where $M_{ji} = \frac{1}{d_i}$ if a link from i to j exists, and zero otherwise.

Example of Transition Matrix:



$$M = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 1 & 0 \end{bmatrix}$$

Outline

1. Why WSD?
2. Supervised WSD vs. Knowledge-Based WSD
3. Prior Approaches to Knowledge Based WSD
4. Page Rank
   - Static Page Rank
   - Personalized Page Rank
5. Evaluation

Static Page Rank

- Gives equal initial probability mass to all synsets of all words.
- Makes prediction of **one Most Predominant Sense** for all instances of the word in given text.
- In a way, this is analogous to **MFS (Most Frequent Sense)**.

Outline

Personalised Page Rank

- Context words for a given instance are added as extra nodes to the graph with edges to their respective concepts.
- Gives equal initial probability mass to all context word nodes of given instance.
- Initial probability mass of all other nodes are set to 0.
- Thus, for each new context of given instance, a sense is predicted.
- The resulting Personalized PageRank vector can be seen as a measure of the structural relevance of LKB concepts in the presence of the input context.

Personalised Page Rank

Problem:

If one of the target words has two senses which are related to each other by semantic relations, those senses would reinforce each other, and could thus dampen the effect of the other senses in the context.

Personalised Page Rank

Problem:
If one of the target words has two senses which are related to each other by semantic relations, those senses would reinforce each other, and could thus dampen the effect of the other senses in the context.

Solution:
For each target word W_i, the initial probability mass is concentrated in the context words, but not in the target word itself, thus avoiding bias in the initial score of concepts associated to target word W_i.

Variants of Personalised Page Rank

PPRank.maxsense:

Select the sense which is chosen most frequently by Personalized PageRank.

Variants of Personalised Page Rank

PPRank.maxsense:
Select the sense which is chosen most frequently by Personalized PageRank.

PPRank.all-in-one:
Concatenate contexts from all instances of target word to form one large instance and then use Personalised Page Rank.

Variants of Personalised Page Rank

Personalised PageRank using Related Words:

- Instead of allotting initial probability mass to context words, this method allots initial probability mass to related words from the distributional thesaurus as seen in Koeling et. al.
- Thus, we would end up predicting the same sense for all instances of the target word in the given text.

Dataset Used for Evaluation

- Three datasets were used [Koeling et al. 2005]:
  - General - BNC
  - Domain Specific - Sports(Reuters) and Finances(Reuters)

Dataset Used for Evaluation

- Three datasets were used [Koeling et al. 2005]:
  - General - BNC
  - Domain Specific - Sports(Reuters) and Finances(Reuters)

Dataset Used for Evaluation

- Three datasets were used [Koeling et al. 2005]:
  - General - BNC
  - Domain Specific - Sports(Reuters) and Finances(Reuters)
- It was ensured that the dataset was fairly polysemous. Each word had an average polysemy of 6.7 senses, ranging from 2 to 13 senses.

Dataset Used for Evaluation

- Three datasets were used [Koeling et al. 2005]:
  - General - BNC
  - Domain Specific - Sports(Reuters) and Finances(Reuters)
- It was ensured that the dataset was fairly polysemous. Each word had an average polysemy of 6.7 senses, ranging from 2 to 13 senses.
- The 3 datasets were semantically annotated by three reviewers(Inter-tagger agreement - 65%).

Baselines:

- Random Sense.
- SemCor Most Frequent Sense(MFS).
- Static PageRank.

Upperbound - Test MFS:

- Most Frequent Sense on annotated test data.
- Not practical as we cannot always obtain annotated test data.

Results

- For each model a confidence interval was also computed using bootstrap resampling.

	Systems	BNC	Sports	Finances
Baselines	Random	*19.7	*19.2	*19.5
	SemCor MFS	*34.9 [33.60, 36.20]	*19.6 [18.40, 20.70]	*37.1 [35.70, 38.00]
	Static PRank	*36.6 [35.30, 38.00]	*20.1 [18.90, 21.30]	*39.6 [38.40, 41.00]
Supervised	SVM	*38.7 [37.30, 39.90]	*25.3 [24.00, 26.30]	*38.7 [37.10, 40.10]
	k-NN	42.8 [41.30, 44.10]	*30.3 [29.00, 31.20]	*43.4 [42.00, 44.80]
Context	PPRank	**43.8** [42.40, 44.90]	*35.6 [34.30, 37.00]	*46.9 [45.39, 48.10]
	PPRank.maxsense	*39.3 [38.00, 40.60]	*36.0 [34.70, 37.40]	*53.1 [51.70, 54.40]
	PPRank.all-in-one	*39.6 [38.20, 40.90]	*42.5 [41.20, 43.90]	*46.4 [44.90, 47.80]
Related words	[Koeling *et al.*, 2005]	*40.7 [39.20, 42.00]	*43.3 [42.00, 44.60]	*49.7 [48.00, 51.10]
	PPRank	*37.7 [36.30, 39.00]	**51.5** [50.00, 52.90]	**59.3** [57.80, 60.70]
	PPRank.th+ctx	*38.2 [36.70, 39.50]	49.9 [48.50, 51.60]	57.8 [56.40, 59.20]
Upperbound	Test MFS	*52.0 [50.60, 53.30]	*77.8 [76.60, 79.00]	*82.3 [81.00, 83.30]

Performance of Baseline Models

- As expected, Random performed poorly due to high polysemy of words.

Performance of Baseline Models

- As expected, Random performed poorly due to high polysemy of words.
- SemCor MFS performed badly because it highly depends on the "sense distributions" of the words in training corpus. (More on this later). SemCor MFS performed close to Random in the case of domain-specific data.

Performance of Baseline Models

- As expected, Random performed poorly due to high polysemy of words.
- SemCor MFS performed badly because it highly depends on the "sense distributions" of the words in training corpus. (More on this later). SemCor MFS performed close to Random in the case of domain-specific data.
- Static PageRank faces the same issues as SemCor MFS as it does not consider any context words and it simply chooses the most important sense in the Knowledge Base. The authors hypothesize that Static PR is almost identical to SemCor MFS in theory, and this can be verified from the results.

Performance of Supervised Models

- Since the models are highly dependent on SemCor for training, they face the same issues as SemCor MFS and only provide a slight improvement

Performance of Supervised Models

- Since the models are highly dependent on SemCor for training, they face the same issues as SemCor MFS and only provide a slight improvement
- Clearly none of the methods discussed so far can be deployed to any different corpora, even if its in a general domain.

Performance of Context based PageRank Models

- Performs better than supervised methods as the context words influence the initial probabilities during PR.

Performance of Context based PageRank Models

- Performs better than supervised methods as the context words influence the initial probabilities during PR.
- The maxsense and all-in-one variants perform better in case of the domain-specific data as there would be lesser polysemy and within the domain each word will predominantly follow a sense most related to that domain. However in case of the BNC corpus there would be words with multiple senses in use, so maxsense and all-in-one wouldn't work well.

Performance of Related Words based PageRank Models

- PR using related words: The best performing model in case of domain-specific data. This is because it uses information such as co-occurence counts from the entire corpus to find related words, which is more relevant to a word's sense if we consider the word to have only 1 sense in the data. As in the case of maxsense and all-in-one, since there is lesser polysemy for the domain-specific data, this performs well

Performance of Related Words based PageRank Models

- PR using related words: The best performing model in case of domain-specific data. This is because it uses information such as co-occurence counts from the entire corpus to find related words, which is more relevant to a word's sense if we consider the word to have only 1 sense in the data. As in the case of maxsense and all-in-one, since there is lesser polysemy for the domain-specific data, this performs well
- The model using combination of context and related words performs poorly, probably because the initial probability mass that is provided on the graph is suboptimally distributed due to the presence of the mix of words. If a context suggests one sense and the related words suggests another it could lead to bad disambiguation.

Effect of Sense Distributions

- The authors divided words into 2 groups based on their sense occurrence/distribution in the SemCor corpus and the test corpor: Similar group and Different group

Effect of Sense Distributions

- The authors divided words into 2 groups based on their sense occurrence/distribution in the SemCor corpus and the test corpor: Similar group and Different group
- It was observed for BNC the two groups had almost the same number of words(19,22), whereas for the domain specific data the similar group was half the size of the different group.

Effect of Sense Distributions

- The authors divided words into 2 groups based on their sense occurrence/distribution in the SemCor corpus and the test corpor: Similar group and Different group
- It was observed for BNC the two groups had almost the same number of words(19,22), whereas for the domain specific data the similar group was half the size of the different group.
- This explains why SemCor MFS performed fairly well on the BNC corpus as compared to the domain-specific data. The authors then proceeded to test Page Rank, MFS and Supervised methods on these two separate groups.

Effect of Sense Distributions

	Similar			Different		
Systems	BNC	Sp.	Fin.	BNC	Sp.	Fin.
Semcor MFS	54.7	**65.5**	**79.0**	9.7	3.8	8.4
k-NN	**57.1**	64.6	69.9	24.6	18.5	25.4
Context PPR	50.0	34.9	64.2	**36.0**	35.9	35.0
Related PPR	38.1	53.1	73.7	24.8	**50.9**	**49.5**

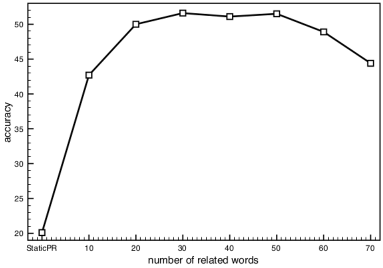
- As observed in the above table, SemCor MFS and supervised methods performed better on the similar data,(as it is close to the training set), while PageRank performed better on the different group words.

Effect of Sense Distributions

	Similar			Different		
Systems	BNC	Sp.	Fin.	BNC	Sp.	Fin.
Semcor MFS	54.7	**65.5**	**79.0**	9.7	3.8	8.4
k-NN	**57.1**	64.6	69.9	24.6	18.5	25.4
Context PPR	50.0	34.9	64.2	**36.0**	35.9	35.0
Related PPR	38.1	53.1	73.7	24.8	**50.9**	**49.5**

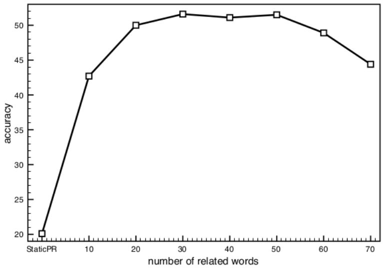
- As observed in the above table, SemCor MFS and supervised methods performed better on the similar data,(as it is close to the training set), while PageRank performed better on the different group words.
- Since PR based models use WordNet knowledge and context/related words in the test data, they perform better on the different words group. For example, a particular sense for a word may not have even been present in the SemCor data, however all the word senses would be present in WordNet and given the correct words which relate to this, the sense will be given a high page rank.

Number of Related Words



- As observed in the above plot, with increase in number of related words the performance first increases, and then drops.

Number of Related Words



- As observed in the above plot, with increase in number of related words the performance first increases, and then drops.
- Initially the accuracy increases as we need enough related words to capture the predominant sense. However adding too many related words has a noisy effect, and some of the words may be more relevant to different senses of the word. For example, bank can have related words pertaining to financial sense or to the river-bank sense.

References I

Eneko Agirre and Aitor Soroa.
Knowledge-Based WSD on Specific Domains: Performing better than Generic Supervised WSD

Eneko Agirre and Aitor Soroa.
Personalizing PageRank for Word Sense Disambiguation