Using Wikipedia to Help Word Sense Disambiguation

Wei Qiu

June 5, 2012

Outline

1 Word Sense Disambiguation

2 Wikipedia

3 Mihalcea2007: Use Wikipedia to Generate Training Corpus

4 Navigli2010: Extend WordNet with Semantic Relations Extracted from Wikipedia

5 Summaries and Conclusions

Outline

1 Word Sense Disambiguation

2 Wikipedia

3 Mihalcea2007: Use Wikipedia to Generate Training Corpus

4 Navigli2010: Extend WordNet with Semantic Relations Extracted from Wikipedia

5 Summaries and Conclusions

Motivation

(1) I hardly water the plant that grows in my yard.

(2) The government decided to build a new power plant.

Motivation

(1) I hardly water the plant that grows in my yard.

(2) The government decided to build a new power plant.

(3) The eastern bank of Saar river is a great place for barbecue.

(4) There is a bank in Saarland University.

Motivation

(1) I hardly water the plant that grows in my yard.

(2) The government decided to build a new power plant.

(3) The eastern bank of Saar river is a great place for barbecue.

(4) There is a bank in Saarland University.

(5) I went fishing for some some sea bass. (a type of fish)

(6) The bass line of the song is too weak. (tones of low frequency)

...

Motivation

(1) I hardly water the plant that grows in my yard.

(2) The government decided to build a new power plant.

(3) The eastern bank of Saar river is a great place for barbecue.

(4) There is a bank in Saarland University.

(5) I went fishing for some some sea bass. (a type of fish)

(6) The bass line of the song is too weak. (tones of low frequency)

...

Ambiguities are inherent to human language!

Definition of WSD

Automatically assigning the most appropriate meaning to a polysemous word within a given context.

Definition of WSD

Automatically assigning the most appropriate meaning to a polysemous word within a given context.

Key Questions of WSD

1. What's the meaning?
2. How can computers represent the meanings?
3. What's the appropriate granularity of meanings for WSD?

Approximation: systems must choose one of discrete number of senses known to the systems in advance. A predefined inventory is often used for the senses.(Dictionaries and Thesauruses(WordNet))

A Very Short Overview On WSD

1. Knowledge based: rely on man-made lexical resources
2. Corpus based:
   1. Supervised: sense-tagged training data
   2. Unsupervised(Word Sense Induction)
   3. Semi-supervised, bootstrapping

A Very Short Overview On WSD

1. Knowledge based: rely on man-made lexical resources
2. Corpus based:
   1. Supervised: sense-tagged training data
   2. Unsupervised(Word Sense Induction)
   3. Semi-supervised, bootstrapping

Notes

Supervised learning methods have been the most successful WSD technique. However,

1. Cost of human labeling is very high.
2. Limited due to the strong connection between senses and context.

Outline

1 Word Sense Disambiguation

2 Wikipedia

3 Mihalcea2007: Use Wikipedia to Generate Training Corpus

4 Navigli2010: Extend WordNet with Semantic Relations Extracted from Wikipedia

5 Summaries and Conclusions

Wikipedia

Introduction

1. Online encyclopedia
2. Collaborative
3. Large quantity
4. High quality

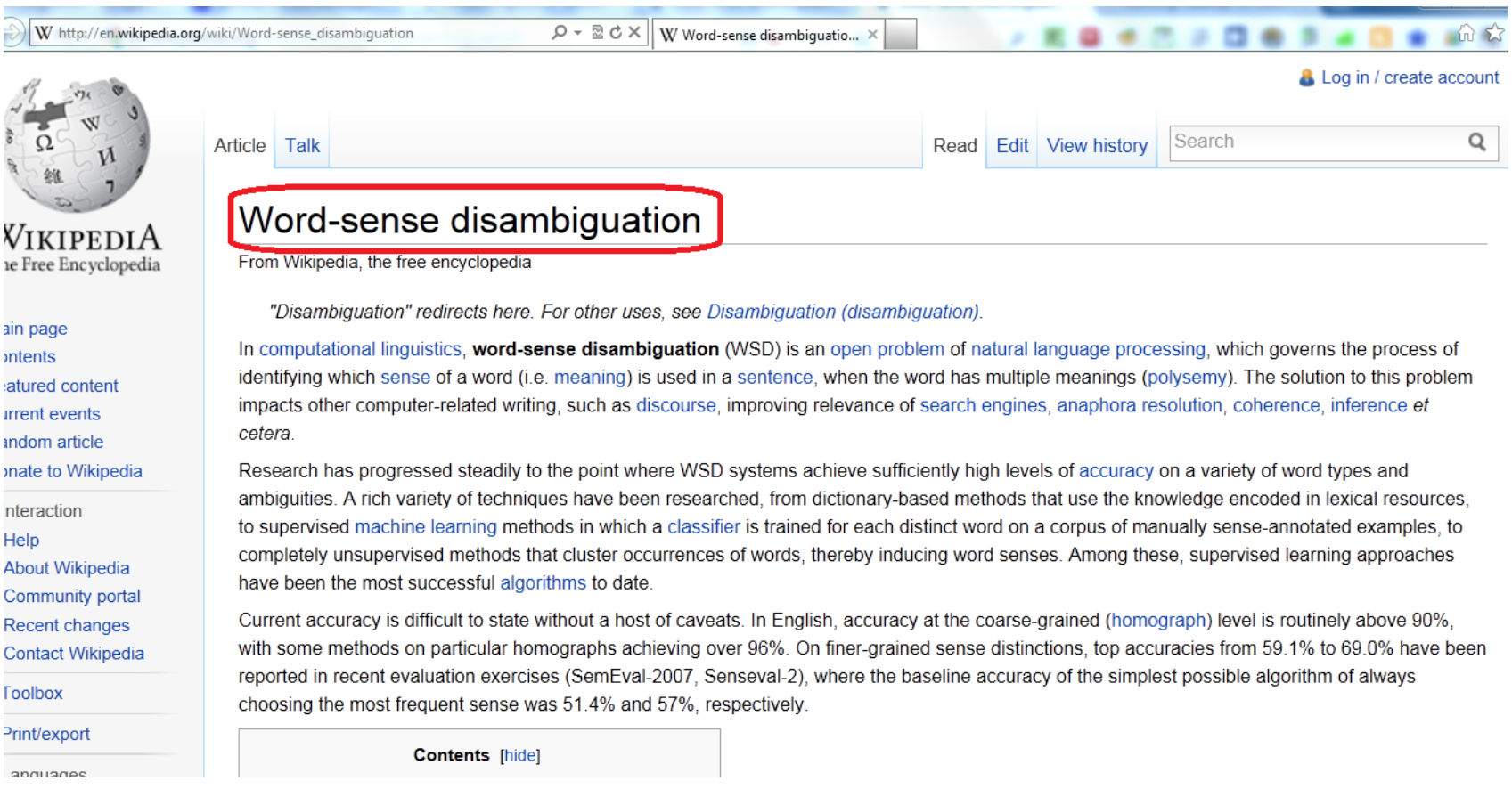
Structure and Terminology

Basic Entry: Article/Page

1. Defines and describes an entity or an event
2. Hypertext

Structure and Terminology

Identifier of Page

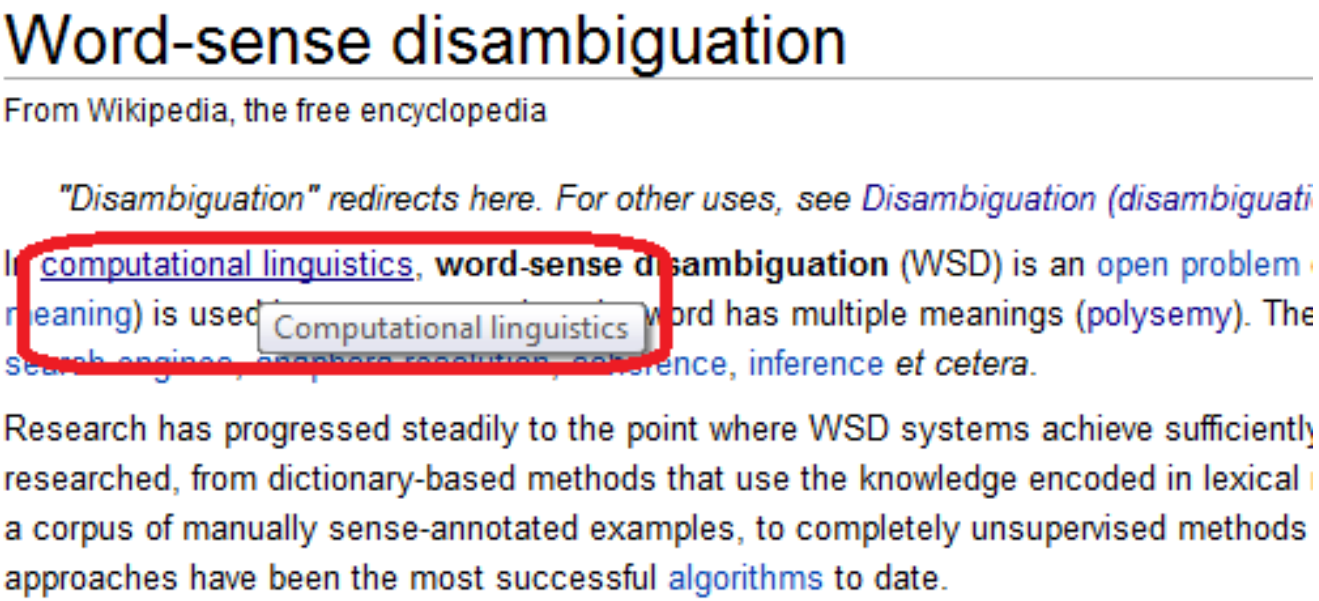


Structure and Terminology

Hyperlinks

1. created with identifiers
2. direct link & piped link

Direct link: the surface(reference) is the same with the identifier.



Structure and Terminology

Piped link:

which governs the process of identifying which sense of a word (i.e.
ts other computer-related writing, such as dis ours elevance of
Word sense

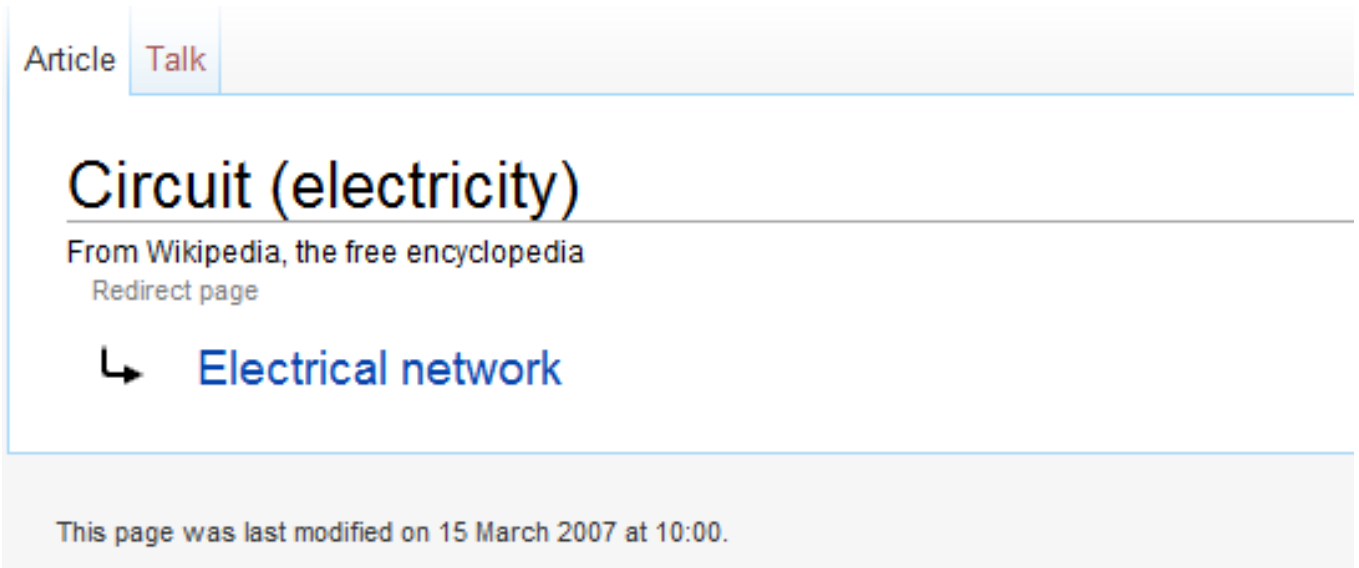
ariety of word types and ambiguities. A rich variety of techniques have been
ne learning methods in which a classifier is trained for each distinct word on
ds, thereby inducing word senses. Among these, supervised learning

level is routinely above 90%, with some methods on particular homographs
cent evaluation exercises (SemEval-2007, Senseval-2), where the baseline
ively.

Structure and Terminology

Redirect Pages

Different contributors may use different names for the same entity. There is only one page which actually containing the description.



Structure and Terminology

Disambiguation Pages

For ambiguous words, there are several pages describing different meanings of words.

Article | Talk | Read | Edit

Circuit

From Wikipedia, the free encyclopedia

Circuit may mean:

- Circuit theory, the theory of accomplishing work by routing electrons, gas, fluids, or other matter through loops (e.g., electrical, hydraulic or pneumatic circuits)
  Further information: Hydraulic analogy and Fluidics

In **electrical engineering**

- Electrical circuit, an electrical network that has a closed loop giving a return path for the current
  - Circuit analysis, the process of finding the voltages across, and the currents through, every component in an electrical circuit
  - Electronic circuit, active electronic components connected in a circuit
  - Analog circuit or Digital circuit
  - Integrated circuit
    - Mixed-signal integrated circuit
  - Asynchronous circuit or Synchronous circuit
  - Printed circuit board (PCB)
  - Series and parallel circuits
  - Telecommunication circuit
  - Circuit diagram
  - Balanced circuit
  - LC circuit

In **fluid power** and **fluid mechanics**

Outline

1 Word Sense Disambiguation

2 Wikipedia

3 Mihalcea2007: Use Wikipedia to Generate Training Corpus

4 Navigli2010: Extend WordNet with Semantic Relations Extracted from Wikipedia

5 Summaries and Conclusions

Motivation & Idea

Motivation

1. To address *sense-tagged data bottleneck problem*

Motivation & Idea

Motivation

1. To address *sense-tagged data bottleneck problem*

Idea

Hypelinks(links & piped links) can be regarded as annotations.

Examples

(7) In 1834, Sumner was admitted to the **[[bar (law)|bar]]** at the age of twenty-three, and entered private practice in Boston.

(8) It is danced in 3/4 time (like most waltzes), with the couple turning approx. 180 degrees every **[[bar (music)|bar]]**.

(9) Vehicles of this type may contain expensive audio players, televisions, video players, and **[[bar (counter)|bar]]**s, often with refrigerators.

(10) Jenga is a popular beer in the **[[bar (establishment)|bar]]**s of Thailand.

(11) This is a disturbance on the water surface of a river or estuary, often cause by the presence of a **[[bar (landform)|bar]]** or dune on the riverbed.

Why not use disambiguation pages directly?

Disambiguation pages list possible meanings of words, why not use them directly?

1. Occurrences of ambiguous words are not linked to the articles mentioned on disambiguation pages.
2. Inconsistencies.

Why not use disambiguation pages directly?

Disambiguation pages list possible meanings of words, why not use them directly?

1. Occurrences of ambiguous words are not linked to the articles mentioned on disambiguation pages.
2. Inconsistencies.

Example

bar is a disambiguation page
paper is not, but *paper_(disambiguation)* is.

Building Sense Tagged Corpora

1. Extract all paragraphs in Wikipedia that contain ambiguous word as part of a link(piped link)
2. Collect all possible labels by extracting the leftmost component of links.
3. Labels are manually mapped to corresponding WordNet sense.

Example

Bar

1217 examples extracted from Wikipedia.
All *bar* in single links are removed because *bar* is disambguiation page.
Labels are mapped to 9 senses in WordNet.

Example

Word sense	Labels in Wikipedia	Wikipedia definition	WordNet definition
bar (establishment)	bar_(establishment), nightclub gay_club, pub	a retail establishment which serves alcoholic beverages	a room or establishment where alcoholic drinks are served over a counter
bar (counter)	bar_(counter)	the counter from which drinks are dispensed	a counter where you can obtain food or drink
bar (unit)	bar_(unit)	a scientific unit of pressure	a unit of pressure equal to a million dynes per square centimeter
bar (music)	bar_(music), measure_music musical_notation	a period of music	musical notation for a repeating pattern of musical beats
bar (law)	bar_association, bar_law law_society_of_upper_canada state_bar_of_california	the community of persons engaged in the practice of law	the body of individuals qualified to practice law in a particular jurisdiction
bar (landform)	bar_(landform)	a type of beach behind which lies a lagoon	a submerged (or partly submerged) ridge in a river or along a shore
bar (metal)	bar_metal, pole_(object)	-	a rigid piece of metal or wood
bar (sports)	gymnastics_uneven_bars, handle_bar	-	a horizontal rod that serves as a support for gymnasts as they perform exercises
bar (solid)	candy_bar, chocolate_bar	-	a block of solid substance

Figure: Word senses for the word bar, based on annotation labels used in Wikipedia

Disambiguation System

Preprocess: tokenize and POS tag

Naive Bayes classifer:

1. local features
    1. current word
    2. POS of current word
    3. 3-gram
    4. POS of 3-gram
2. topical features
    1. list of at most 5 words occurring at least 3 times in the context defining a certain word sense

Experiments

Settings

1. subset of ambiguous words used in SENSEVAL-2 and SENSEVAL-3
2. focused on nouns
3. only nouns with more than 1 sense selected
4. baselines:
    1. MFS: always select most frequent sense
    2. LeskC(Kilgarriff and Rosenzweig, 2000)

Results

			baselines		word sense
word	#s	#ex	MFS	LeskC	disambig.
argument	2	114	70.17%	73.63%	**89.47%**
arm	3	291	61.85%	69.31%	**84.87%**
atmosphere	3	773	54.33%	56.62%	**71.66%**
bank	3	1074	**97.20%**	**97.20%**	**97.20%**
bar	10	1108	47.38%	68.09%	**83.12%**
chair	3	194	67.57%	65.78%	**80.92%**
channel	5	366	51.09%	52.50%	**71.85%**
circuit	4	327	85.32%	85.62%	**87.15%**
degree	7	849	58.77%	73.05%	**85.98%**
difference	2	24	**75.00%**	**75.00%**	**75.00%**
disc	3	73	52.05%	52.05%	**71.23%**
dyke	2	76	77.63%	82.00%	**89.47%**
fatigue	3	123	66.66%	70.00%	**93.22%**
grip	3	34	44.11%	**77.00%**	70.58%
image	2	84	69.04%	74.50%	**80.28%**
material	3	223	**95.51%**	**95.51%**	**95.51%**
mouth	2	409	94.00%	94.00%	**95.35%**
nature	2	392	**98.72%**	**98.72%**	98.21%
paper	5	895	**96.98%**	**96.98%**	**96.98%**
party	3	764	68.06%	68.28%	**75.91%**
performance	2	271	**95.20%**	**95.20%**	**95.20%**
plan	3	83	77.10%	81.00%	**81.92%**
post	5	33	54.54%	**62.50%**	51.51%
restraint	2	9	**77.77%**	**77.77%**	**77.77%**
sense	2	183	**95.10%**	**95.10%**	**95.10%**
shelter	2	17	**94.11%**	**94.11%**	**94.11%**
sort	2	11	81.81%	**90.90%**	**90.90%**
source	3	78	55.12%	81.00%	**92.30%**
spade	3	46	60.86%	**81.50%**	80.43%
stress	3	565	53.27%	54.28%	**86.37%**
AVERAGE	3.31	316	72.58%	78.02%	**84.65%**

Outline

1 Word Sense Disambiguation

2 Wikipedia

3 Mihalcea2007: Use Wikipedia to Generate Training Corpus

4 Navigli2010: Extend WordNet with Semantic Relations Extracted from Wikipedia

5 Summaries and Conclusions

Motivation

1. Though supervised methods are good, human annotation is costly.
2. In contrast, knowledge-based approaches exploit wide-coverage lexical resources such as WordNet.
3. WordNet is typically insufficient.
4. Previous work showed that *manually-developed extension of WordNet would help*

Motivation

1. Though supervised methods are good, human annotation is costly.
2. In contrast, knowledge-based approaches exploit wide-coverage lexical resources such as WordNet.
3. WordNet is typically insufficient.
4. Previous work showed that *manually-developed extension of WordNet would help*

Idea

Extend WordNet automatically using semantic relations extracted from Wikipedia

Extending WordNet

Two steps:

1. Map Wikipedia pages to WordNet senses
2. Transfer relations connecting Wikipedia pages to WordNet.

Mapping Wikipedia Pages to WordNet Senses

Disambiguation Context:

1. Wikipedia
   1. Sense labels; e.g. SODA(SOFT DRINK) $\rightarrow$ SOFT, DRINK
   2. Links: title's lemmas of outgoing links: e.g. The links in Wikipage SODA include soda, lemonade, sugar, etc
   3. Categories; e.g. SODA(SOFT DRINK) $\rightarrow$ SOFT DRINKS. (only syntactic heads used)
2. WordNet sense
   1. Synonymy
   2. Hypernymy/Hoponymy
   3. Sisterhood
   4. Gloss

Mapping Algorithm: Link Wikipedia Page to WordNet Sense

Input: $Senses_{\text{Wiki}}$, $Senses_{\text{WN}}$
Output: a mapping $\mu : Senses_{\text{Wiki}} \rightarrow Senses_{\text{WN}}$

1: **for each** $w \in Senses_{\text{Wiki}}$
2: $\quad \mu(w) := \epsilon$
3: **for each** $w \in Senses_{\text{Wiki}}$
4: $\quad$ **if** $|Senses_{\text{Wiki}}(w)| = |Senses_{\text{WN}}(w)| = 1$ **then**
5: $\quad\quad \mu(w) := w_n^1$
6: **for each** $w \in Senses_{\text{Wiki}}$
7: $\quad$ **if** $\mu(w) = \epsilon$ **then**
8: $\quad\quad$ **for each** $d \in Senses_{\text{Wiki}}$ s.t. d redirects to w
9: $\quad\quad\quad$ **if** $\mu(d) \neq \epsilon$ and $\mu(d)$ is in a synset of w **then**
10: $\quad\quad\quad\quad \mu(w) :=$ sense of w in synset of $\mu(d)$; **break**
11: **for each** $w \in Senses_{\text{Wiki}}$
12: $\quad$ **if** $\mu(w) = \epsilon$ **then**
13: $\quad\quad$ **if** no tie occurs **then**
14: $\quad\quad\quad \mu(w) := \underset{s \in Senses_{\text{WN}}(w)}{\text{argmax}} \; p(s|w)$
15: **return** μ

1. Step 1: Initialize

Mapping Algorithm: Link Wikipedia Page to WordNet Sense

Input: $Senses_{\text{Wiki}}$, $Senses_{\text{WN}}$
Output: a mapping $\mu : Senses_{\text{Wiki}} \rightarrow Senses_{\text{WN}}$

1: **for each** $w \in Senses_{\text{Wiki}}$
2: $\quad \mu(w) := \epsilon$
3: **for each** $w \in Senses_{\text{Wiki}}$
4: $\quad$ **if** $|Senses_{\text{Wiki}}(w)| = |Senses_{\text{WN}}(w)| = 1$ **then**
5: $\quad\quad \mu(w) := w_n^1$
6: **for each** $w \in Senses_{\text{Wiki}}$
7: $\quad$ **if** $\mu(w) = \epsilon$ **then**
8: $\quad\quad$ **for each** $d \in Senses_{\text{Wiki}}$ s.t. d redirects to w
9: $\quad\quad\quad$ **if** $\mu(d) \neq \epsilon$ and $\mu(d)$ is in a synset of w **then**
10: $\quad\quad\quad\quad \mu(w) :=$ sense of w in synset of $\mu(d)$; **break**
11: **for each** $w \in Senses_{\text{Wiki}}$
12: $\quad$ **if** $\mu(w) = \epsilon$ **then**
13: $\quad\quad$ **if** no tie occurs **then**
14: $\quad\quad\quad \mu(w) := \underset{s \in Senses_{\text{WN}}(w)}{\text{argmax}} \; p(s|w)$
15: **return** μ

1. Step 1: Initialize
2. Step 2: Map all monosemous words(both in Wikipedia and WordNet)

Mapping Algorithm: Link Wikipedia Page to WordNet Sense

Input: $Senses_{\text{Wiki}}$, $Senses_{\text{WN}}$
Output: a mapping $\mu : Senses_{\text{Wiki}} \rightarrow Senses_{\text{WN}}$

1: **for each** $w \in Senses_{\text{Wiki}}$
2: $\quad \mu(w) := \epsilon$
3: **for each** $w \in Senses_{\text{Wiki}}$
4: $\quad$ **if** $|Senses_{\text{Wiki}}(w)| = |Senses_{\text{WN}}(w)| = 1$ **then**
5: $\quad\quad \mu(w) := w_n^1$
6: **for each** $w \in Senses_{\text{Wiki}}$
7: $\quad$ **if** $\mu(w) = \epsilon$ **then**
8: $\quad\quad$ **for each** $d \in Senses_{\text{Wiki}}$ s.t. d redirects to w
9: $\quad\quad\quad$ **if** $\mu(d) \neq \epsilon$ and $\mu(d)$ is in a synset of w **then**
10: $\quad\quad\quad\quad \mu(w) :=$ sense of w in synset of $\mu(d)$; **break**
11: **for each** $w \in Senses_{\text{Wiki}}$
12: $\quad$ **if** $\mu(w) = \epsilon$ **then**
13: $\quad\quad$ **if** no tie occurs **then**
14: $\quad\quad\quad \mu(w) := \underset{s \in Senses_{\text{WN}}(w)}{\text{argmax}} \; p(s|w)$
15: **return** μ

1. Step 1: Initialize
2. Step 2: Map all monosemous words(both in Wikipedia and WordNet)
3. Step 3: Map all words whose redirection pages are already mapped.

Mapping Algorithm: Link Wikipedia Page to WordNet Sense

Input: $Senses_{\text{Wiki}}$, $Senses_{\text{WN}}$
Output: a mapping $\mu : Senses_{\text{Wiki}} \rightarrow Senses_{\text{WN}}$

1: **for each** $w \in Senses_{\text{Wiki}}$
2: $\quad \mu(w) := \epsilon$
3: **for each** $w \in Senses_{\text{Wiki}}$
4: $\quad$ **if** $|Senses_{\text{Wiki}}(w)| = |Senses_{\text{WN}}(w)| = 1$ **then**
5: $\quad\quad \mu(w) := w_n^1$
6: **for each** $w \in Senses_{\text{Wiki}}$
7: $\quad$ **if** $\mu(w) = \epsilon$ **then**
8: $\quad\quad$ **for each** $d \in Senses_{\text{Wiki}}$ s.t. d redirects to w
9: $\quad\quad\quad$ **if** $\mu(d) \neq \epsilon$ and $\mu(d)$ is in a synset of w **then**
10: $\quad\quad\quad\quad \mu(w) :=$ sense of w in synset of $\mu(d)$; **break**
11: **for each** $w \in Senses_{\text{Wiki}}$
12: $\quad$ **if** $\mu(w) = \epsilon$ **then**
13: $\quad\quad$ **if** no tie occurs **then**
14: $\quad\quad\quad \mu(w) := \underset{s \in Senses_{\text{WN}}(w)}{\text{argmax}} \; p(s|w)$
15: **return** μ

1. Step 1: Initialize
2. Step 2: Map all monosemous words(both in Wikipedia and WordNet)
3. Step 3: Map all words whose redirection pages are already mapped.
4. Step 4: Map other pages.

Mapping Algorithm: Step 4

$$
\begin{aligned}
\mu(w) &= \arg\max_{s \in Senses_{WN}(w)} p(s|w) \\
&= \arg\max_{s} \frac{p(s,w)}{p(w)} \\
&= \arg\max_{s} p(s,w) \\
p(s,w) &= \frac{score(s,w)}{\sum_{\substack{s' \in Senses_{WN}(w), \\ w' \in Senses_{Wiki}(w)}} score(s',w')} \\
score(s,w) &= |Ctx(s) \bigcap Ctx(w)| + 1
\end{aligned}
$$

Transferring Semantic Relations: WordNet++

Idea

Transfer the link between pages of Wikipedia to WordNet senses.

Transferring Semantic Relations: WordNet++

Idea

Transfer the link between pages of Wikipedia to WordNet senses.

Example

Wikipage SODA(SOFT DRINK) contains a link to the Wikipage SYRUP.
if $\mu(SODA(SOFTDRINK)) = soda_n^2$ and $\mu(SYRUP) = syrup_n^1$
Then $(soda_n^2, syrup_n 1)$ will be added to WordNet.

Experiments

1. Evaluation of Mapping
2. Coarse-grained WSD
3. Domain WSD

Experiment 1: Evaluation of Mapping

Settings

1. 1000 Wikipages are random selected as sample.
2. Compared with human annotated data.
3. Tow baseline systems, Most Frequent Sense & Random.

Results of Experiment 1

	P	R	F_1	A
Structure	**82.2**	68.1	74.5	81.1
Gloss	81.1	64.2	71.7	78.8
Structure + Gloss	81.9	**77.5**	**79.6**	**84.4**
MFS BL	24.3	47.8	32.2	24.3
Random BL	23.8	46.8	31.6	23.9

Figure: Performance of the mapping algorithm. Structure: Disambiguation context contains only synonymy.etc; Gloss: Disambiguation context contains only gloss.

Experiment 2: Coarse-grained WSD

Semeval-2007 coarse-grained all-words WSD task.

Why "coarse-grained"

1. Fined-grained WSD is much harder.
2. Meanings of Wikipedia are coarser.

Disambiguation Algorithm

1. Simplified Extended Lesk(ExtLesk): overlap of context of target word and glosses of sense and synsets who have semantic relation to it.
2. Degree Centrality(Degree): walk from target word to other words in context. Sense with highest vertex degree is selected.

Results of Experiment2

Resource	Algorithm	Nouns only P	R	F_1
WordNet	ExtLesk	83.6	57.7	68.3
	Degree	86.3	65.5	74.5
Wikipedia	ExtLesk	82.3	64.1	72.0
	Degree	**96.2**	40.1	57.4
WordNet++	ExtLesk	82.7	69.2	75.4
	Degree	87.3	72.7	**79.4**
	MFS BL	77.4	**77.4**	77.4
	Random BL	63.5	63.5	63.5

Figure: Performance on Semeval-2007 coarse-grained all-words WSD(nouns only subset)

Comparison with other state-of-the-art systems

Algorithm	Nouns only P/R/F_1	All words P/R/F_1
ExtLesk	81.0	79.1
Degree	**85.5**	**81.7**
SUSSX-FR	81.1	77.0
TreeMatch	N/A	73.6
NUS-PT	82.3	**82.5**
SSI	**84.1**	**83.2**
MFS BL	77.4	78.9
Random BL	63.5	62.7

Figure: Performance on Semeval-2007 coarse-grained all-words WSD with MFS as a back-off strategy when no sense assignment is attempted.

Experiment 3: Domain WSD

Whether can the WSD benefit from the wide coverage of Wikipedia?

Settings

1. Data: Sports and Finance sections of the domain corpora form Koeling2005.
2. Baseline systems: Argirre2009
    1. Personalized PageRank
    2. Static PageRank
    3. -NN supervised WSD

Results of Experiment 3

Algorithm	Sports P/R/F_1	Finance P/R/F_1
k-NN †	30.3	43.4
Static PR †	20.1	39.6
Personalized PR †	35.6	46.9
ExtLesk	40.1	45.6
Degree	**42.0**	**47.8**
MFS BL	19.6	37.1
Random BL	19.5	19.6

Figure: Performance on the Sports and Finance sections of the dataset from Koeling2005

Outline

1 Word Sense Disambiguation

2 Wikipedia

3 Mihalcea2007: Use Wikipedia to Generate Training Corpus

4 Navigli2010: Extend WordNet with Semantic Relations Extracted from Wikipedia

5 Summaries and Conclusions

Summaries and Conclusions

1. Two different ways to use Wikipedia data
   1. As corpus(supervised methods)
   2. As source of knowledge(knowledge-based methods)
2. Knowledge-based system can also achieve state-of-the-art performance.
3. With high quality enrichment of WordNet, simple algorithms can achieve competitive performance.

Thank you!