Ensemble Methods for Unsupervised Word Sense Disambiguation

Samuel Brody[1] Roberto Navigli[2] Mirella Lapata[1]

[1]School of Informatics
University of Edinburgh

[2]Department of Computer Science
University of Rome "La Sapienza"

ACL 2006

Unsupervised WSD

What's it good for?

- New languages without sense frequency information.
- Incomplete sense frequency information in English.
- Domain sensitive words.
- New sense inventories other than WordNet.
- Annotate automatically, correct manually.

Research Questions

- Many different approaches to unsupervised WSD
  (Lesk, 1986; Yarowsky, 1995; Galley and McKeown 2003; McCarthy *et al.*, 2004; Navigli and Velardi, 2005; Mihalcea, 2005; Mohammad and Hirst, 2006).
- Combination methods help in many tasks, including supervised WSD (Florian *et al.* '02).
- We asked ourselves:
  1. Which approachs should we consider?
  2. Are the different approaches complementary?
  3. Can they be combined?

Our Framework

Multiple purpose framework:

Comparison	A standardized environment and dataset.
Decomposition	Strengths and weaknesses of each method.
Combination	Uniform interface for easy integration.

Context-Definition Overlap (Lesk, 1986)

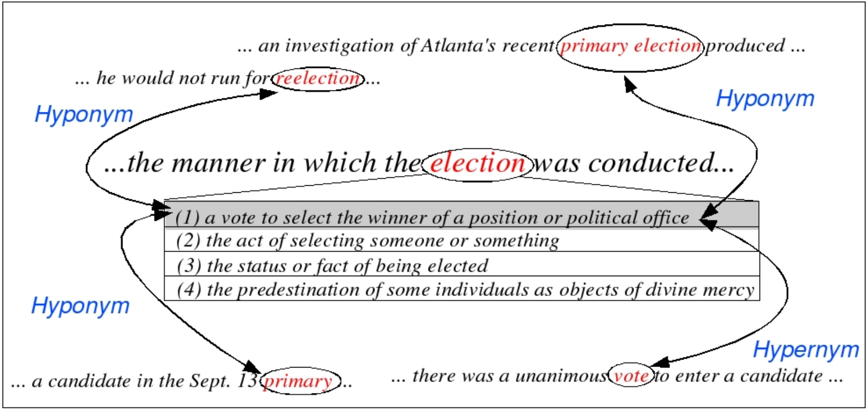
Idea: measure overlap between dictionary glosses and the context of the ambiguous word.

When shooting an arrow with a recurve bow, first adjust your stance.
The two senses for *arrow* in WordNet: 1. A mark to indicate a direction or relation. 2. A projectile with a straight thin shaft and an arrowhead on one end and stabilizing vanes on the other; intended to be shot from a bow.

Extended Glosses: use information from related words (Banerjee and Pedersen, 2003).

Lexical Chains (Galley and McKeown, 2003)

Idea: search for direct WordNet relations between words in the document; use these to disambiguate and form lexical chains.



Similarity-based Ranking (McCarthy *et al.*, 2004)

- The Idea : for each ambiguous word, find words with similar dependency distributions. These are *distributional neighbours.*
- For each neighbour, find the closest sense of the ambiguous word. Increment the score of that sense.
- Select highest scoring sense as the *Predominant Sense* (PS).

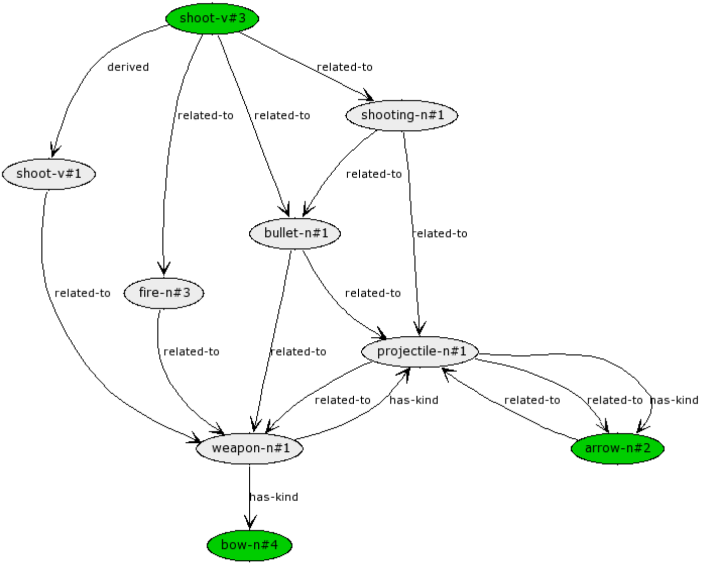
The neighbors of *election* in the BNC:
poll, vote, referendum, ballot, race, campaign, contest, parliament, option, reelection
The <*vote* > sense is more predominant than the <*choice*> sense

SSI (Navigli and Velardi, 2004)

- Graph of WordNet synsets, with weighted connections from extensive lexical knowledge base.
- To disambiguate, use subgraph induced by all nodes (synsets) of words in the sentence.
- Iterative disambiguation.
- Maximize score of weighted connections in the sentence.
- Resulting synsets provide sense labelling.

SSI (Navigli and Velardi, 2004)

I shot an arrow with a bow.



Properties of the Components

Methods offer a range of variability in several aspects.

Method	WSD	Context	Relations
LexChains	types	document	first-order
Overlap	tokens	sentence	first-order
Similarity	types	corpus	higher-order
SSI	tokens	sentence	higher-order

Properties of the Components

Methods offer a range of variability in several aspects.

Method	WSD	Context	Relations
LexChains	types	document	first-order
Overlap	tokens	sentence	first-order
Similarity	types	corpus	higher-order
SSI	tokens	sentence	higher-order

Properties of the Components

Methods offer a range of variability in several aspects.

Method	WSD	Context	Relations
LexChains	types	document	first-order
Overlap	tokens	sentence	first-order
Similarity	types	corpus	higher-order
SSI	tokens	sentence	higher-order

Properties of the Components

Methods offer a range of variability in several aspects.

Method	WSD	Context	Relations
LexChains	types	document	first-order
Overlap	tokens	sentence	first-order
Similarity	types	corpus	higher-order
SSI	tokens	sentence	higher-order

Properties of the Components

Methods offer a range of variability in several aspects.

Method	WSD	Context	Relations
LexChains	types	document	first-order
Overlap	tokens	sentence	first-order
Similarity	types	corpus	higher-order
SSI	tokens	sentence	higher-order

Note: token-based algorithms can assume one-sense-per-discourse, and become type-based (PS).

Experimental Setup

- SemCor corpus.
- 2,595 polysemous nouns (53,674 tokens);
  same data set used by McCarthy *et al.* (2004).
- WordNet 1.7.1 sense inventory.
- Random baseline: uniform distribution over senses.
- Upper bound: first sense heuristic from SemCor.
- Evaluation on tokens and predominant senses (types).

Results: Individual Methods

Method	Acc_{ps}
Baseline	34.5
LexChains	48.3
Overlap	49.4
Similarity	54.9
SSI	53.7
UpperBnd	100

Results: Individual Methods

Method	Acc_{ps}
Baseline	34.5
LexChains	48.3
Overlap	49.4
Similarity	54.9
SSI	53.7
UpperBnd	100

Predominant Sense Detection:

- LexChains and Overlap perform similarly.

Results: Individual Methods

Method	Acc_{ps}
Baseline	34.5
LexChains	48.3
Overlap	49.4
Similarity	54.9
SSI	53.7
UpperBnd	100

Predominant Sense Detection:

- Lexical Chains and Overlap perform similarly.
- So do SSI and Similarity.
- Second pair performs sig. better than the first pair.

Results: Individual Methods

Method	Acc_{ps}
Baseline	34.5
LexChains	48.3
Overlap	49.4
Similarity	54.9
SSI	53.7
UpperBnd	100

Method	$Acc_{wsd/ps}$
Baseline	23.0
LexChains	40.7
Overlap	42.5
Similarity	46.5
SSI	47.9
UpperBnd	68.4

Predominant Sense Detection:

- Lexical Chains and Overlap perform similarly.
- So do SSI and Similarity.
- Second pair performs sig. better than the first pair.

Results: Individual Methods

Method	Acc_{ps}
Baseline	34.5
LexChains	48.3
Overlap	49.4
Similarity	54.9
SSI	53.7
UpperBnd	100

Method	$Acc_{wsd/ps}$
Baseline	23.0
LexChains	40.7
Overlap	42.5
Similarity	46.5
SSI	47.9
UpperBnd	68.4

Predominant Sense Detection:

- Lexical Chains and Overlap perform similarly.
- So do SSI and Similarity.
- Second pair performs sig. better than the first pair.

Word Sense Disambiguation:

- All differences in performance are statistically significant.

Results: Individual Methods

Method	Acc_{ps}
Baseline	34.5
LexChains	48.3
Overlap	49.4
Similarity	54.9
SSI	53.7
UpperBnd	100

Method	$Acc_{wsd/ps}$
Baseline	23.0
LexChains	40.7
Overlap	42.5
Similarity	46.5
SSI	47.9
UpperBnd	68.4

Predominant Sense Detection:

- Lexical Chains and Overlap perform similarly.
- So do SSI and Similarity.
- Second pair performs sig. better than the first pair.

Word Sense Disambiguation:

- All differences in performance are statistically significant.
- SSI best individual method.

Results: Individual Methods

Method	Acc_{ps}	$Acc_{wsd/dir}$	$Acc_{wsd/ps}$
Baseline	34.5	NA	23.0
Overlap	49.4	36.5	42.5
SSI	53.7	42.7	47.9
UpperBnd	100	NA	68.4

The predominant sense sig. outperforms the token-based WSD, in both token-based algorithms!

Why Ensembles?

Method	Overlap	LexChains	Similarity
LexChains	28.05%		
Similarity	35.87%	33.10%	
SSI	30.48%	31.67%	37.14%

- Low overlap between methods.
- Each algorithm correctly labels aprox. 350 words on which the others fail.
- Oracle would achieve 82.4% for PS task, and 58% for WSD.

Why Ensembles?

Method	Overlap	LexChains	Similarity
LexChains	28.05%		
Similarity	35.87%	33.10%	
SSI	30.48%	31.67%	37.14%

- Low overlap between methods.
- Each algorithm correctly labels aprox. 350 words on which the others fail.
- Oracle would achieve 82.4% for PS task, and 58% for WSD.

Conclusion: need an *unsupervised* way to exploit complementary nature of methods.

Equal Voting

- Each ensemble member gets one vote for the PS.
- The sense with the most votes is chosen.
- Ties resolved randomly.

$$Score(Voting(\{M_i\}_{i=1}^{k}), s)) = \sum_{i=1}^{k} eq[s, PS(M_i, w)]$$

$$\text{where } eq[s, PS(M_i, w)] = \begin{cases} 1 & \text{if } s = PS(M_i, w) \\ 0 & \text{otherwise} \end{cases}$$

	Sense 1	Sense 2	Sense 3
Method A			vote
Method B	vote		
Method C			vote
Voting	1	0	**2**

Probability Model

- Each ensemble member provides a probability distribution over the senses.
- These probabilities (normalized scores) are summed.
- The sense with the highest score is chosen.

$$Score(ProbMix(\{M_i\}_{i=1}^{k}), s)) = \sum_{i=1}^{k} \frac{Score(M_i, s)}{\sum_{\hat{s}} Score(M_i, \hat{s})}$$

	Sense 1	Sense 2	Sense 3
Method A	0.30	0.60	0.10
Method B	0.45	0.40	0.15
Method C	0.45	0.30	0.25
ProbMix	1.20	**1.30**	0.50

Ranking

- Each ensemble member provides a ranking of the senses.
- For each sense, the placements are summed.
- The sense with *lowest* total placement (closest to 1st) wins.

$$Score(Ranking(\{M_i\}_{i=1}^{k}), s)) = \sum_{i=1}^{k} Place_i(s)$$

$Place_i(s)$: the number of distinct scores $\geq Score(M_i, s)$.

	Sense 1	Sense 2	Sense 3
Method A	3rd	2nd	1st
Method B	1st	2nd	2nd
Method C	2nd	1st	3rd
Ranking	6	**5**	6

Arbiter

- A single method decides between the senses suggested by the other methods.
- Provides a filter over irrelevant senses, removing distractions.
- Our arbiter was SSI, since it was most accurate, and benefits from a restricted sense inventory.

Ensemble Results

Method	Acc_{ps}	$Acc_{wsd/ps}$
Similarity	54.9	46.5
SSI	53.5	47.9
Arbiter	56.3	48.7
Voting	57.3	49.8
ProbMix	57.2	50.4
Ranking	58.1	50.3

Ensemble Results

Method	Acc_{ps}	$Acc_{wsd/ps}$
Similarity	54.9	46.5
SSI	53.5	47.9
Arbiter	56.3	48.7
Voting	57.3	49.8
ProbMix	57.2	50.4
Ranking	58.1	50.3

- Ensembles perform sig. better than individual methods.

Ensemble Results

Method	Acc_{ps}	$Acc_{wsd/ps}$
Similarity	54.9	46.5
SSI	53.5	47.9
Arbiter	56.3	48.7
Voting	57.3	49.8
ProbMix	57.2	50.4
Ranking	58.1	50.3

- Ensembles perform sig. better than individual methods.
- On WSD, Arbiter is sig. worse than other ensembles.
  - Almost 30% of the time none of the suggested senses was correct.

Ensemble Results

Method	Acc_{ps}	$Acc_{wsd/ps}$
Similarity	54.9	46.5
SSI	53.5	47.9
Arbiter	56.3	48.7
Voting	57.3	49.8
ProbMix	57.2	50.4
Ranking	58.1	50.3

- Ensembles perform sig. better than individual methods.
- On WSD, Arbiter is sig. worse than other ensembles.
  - Almost 30% of the time none of the suggested senses was correct.
- Performance of ProbMix and Ranking are similar.
- Both are sig. better than Voting.

Results on Senseval-3

Method	Precision	Recall	Fscore
Baseline	36.8	36.8	36.8
SSI	62.5	62.5	62.5
IRST-DDD	63.3	62.2	61.2
Ranking	63.9	63.9	63.9
UpperBnd	68.7	68.7	68.7

- Performance on nouns (including monosemous).
- Comparison with IRST-DDD (Strapparava et al. 2004), best unsupervised system.
- Ensemble outperforms SSI and IRST-DDD.

Summary

- Conclusions:
  - Much to be gained from (even) unsupervised combination!
  - Automatically acquired predominant sense outperforms token-based WSD.
- Future Work:
  - Other parts-of-speech.
  - Confidence-based combinations.
  - Integrate other approaches/algorithms.