Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph Based Algorithms for Sequence Data Labeling

Rada Mihalcea (2005)

Presentation by Clayton Greenberg

January 13, 2014

0. Contents

What is a word sense?

- The meaning of a word is affected by its context (Kilgarriff 2006).
- Construction of discrete senses is a “gross simplification” (McCarthy 2009).

WordNet terms

- SynSet – a node comprising "synonymous" tokens. It represents a sense. Searching WordNet on a token returns a list of SynSets, *ordered such that the "most common" sense is listed first*.
- Gloss – a dictionary "definition" associated with a SynSet.
- Hypernym – a more general term.
- Hyponym – a more specific term.

Applying WordNet to WSD

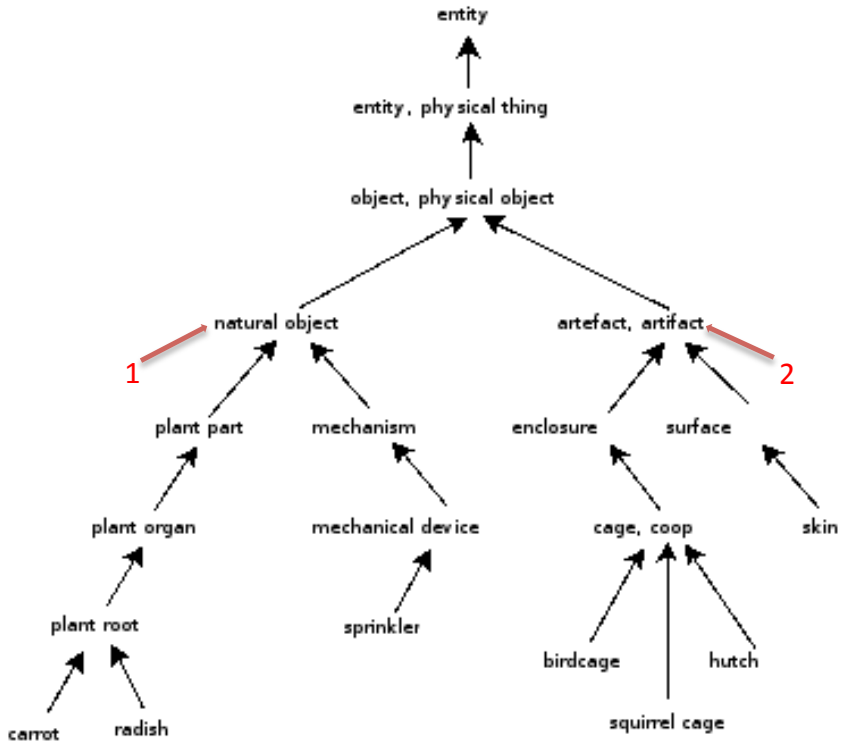


Figure 1. "is a" relation example

Approach typology

Name of WSD approach:	Corpus is manually tagged with senses	System uses a predetermined dictionary of senses
Supervised (Highest "accuracy")	✓	✓
Knowledge-based (Mihalcea 2005)	✗	✓
Unsupervised (Word Sense Induction)	✗	✗

All words task: every word in the corpus is to be tagged with a sense.
Lexical sample task: performance is evaluated on a preselected set of 20-30 words.
First sense heuristic: a baseline in which the most common sense is always chosen.

The state of the art

- Standard corpus SemCor (220k manually sense-tagged words from the Brown corpus).
- Supervised, lexical sample: 65-88% accuracy. All words: 60-82.5%. Outperforms first sense heuristic by ~3% on lexical sample and ~10% on all words.
- Knowledge-based: 37% precision on words with more than one sense, compared to random baseline of 27%.
- Unsupervised, lexical sample (best): 81.6% performance, compared to first sense heuristic at 78.7%.

Main potential applications

- Parsing ambiguities: `Time flies like an arrow; fruit flies like an orange`. But this is an intermediate task. What will you use it for?
- Machine translation: lexical choice task. `match` ➔ `cerilla, partido, duplicado, pareja, puntos, va de conjunto con`. Palmer et al. (2007) assert that every sense distinction that can be made, will be made for some language pair.
- Information retrieval, question-answering, textual entailment: use WSD to prune away documents that appear relevant based on surface terms but refer to different senses.

Problems with the problem

- Supervised data are very expensive to produce.
- Humans do not usually agree on senses, which leads to low ceilings.
- What is the optimal granularity? Coarse-grained gets better results, but fine-grained might be more informative.
- Evidence that WSD can improve the performance of language tasks is thin at best.

Mihalcea (2005): the basic insight

- We should determine all of the senses in a given context concurrently, because the choice of one has an impact on the others.
- Markov models only take the previous sense decisions into account. We want to consider the next ones, too.
- But the problem would get too complex if we considered all senses in the sentence/document at once, so we set a maximum distance.

Runtime complexity

- The MaxDist value reduces runtime complexity.

$$O(C \sum_{i=1}^{n} \sum_{j=i+1}^{i+MaxDist} (N_{w_i} \times N_{w_j}))$$

- Without MaxDist, we could feasibly have an edge connecting each pair of vertices in the graph. In that case, we would have to consider the entire space of possible sense combinations, which grows exponentially with each word.
- C is a constant representing the number of iterations it takes for the scores to converge.
- N is number of labels, n is number of words.

The model

- One graph for the entire sequence.
- A graph is composed of vertices and edges.
- The value of a vertex is a score.
- The value of an edge is a weight.
- Weights are determined first, and then the scores are calculated iteratively from the weights and other scores.

Preprocessing steps

1. Part of speech tagging – This does depend on the sense (c.f. `plant`), but accuracy rates are significantly higher than those for WSD, so it does not cause a problem.
2. Lemmatization (generates the base form of the word, such as `went` ➔ `go`) – This also depends on the sense (c.f. `glasses` as vessels or an ocular device) so the inflected form is usually retained during the disambiguation step.

Calculating edge weights

W_{12} = # of common tokens in their glosses, excluding stop words /

length($gloss_1$) / length($gloss_2$)

Calculating vertex scores

- PageRank: without edge weights:

$$P(V_a) = (1-d) + d * \sum_{V_b \in In(V_a)} \frac{P(V_b)}{|Out(V_b)|}$$

- PageRank: with edge weights:

$$WP(V_a) = (1-d) + d \sum_{V_b \in In(V_a)} \frac{w_{ba}}{\sum_{V_c \in Out(V_b)} w_{bc}} WP(V_b)$$

- Based on random walk.

The algorithm

1. Build the graph, given edge weights.
2. Repeat the score calculation for each vertex until the scores converge.
3. For each word, assign the label corresponding to the vertex with the highest score.

Word senses from WordNet

The **church bells** no longer **rung** on **Sundays**.

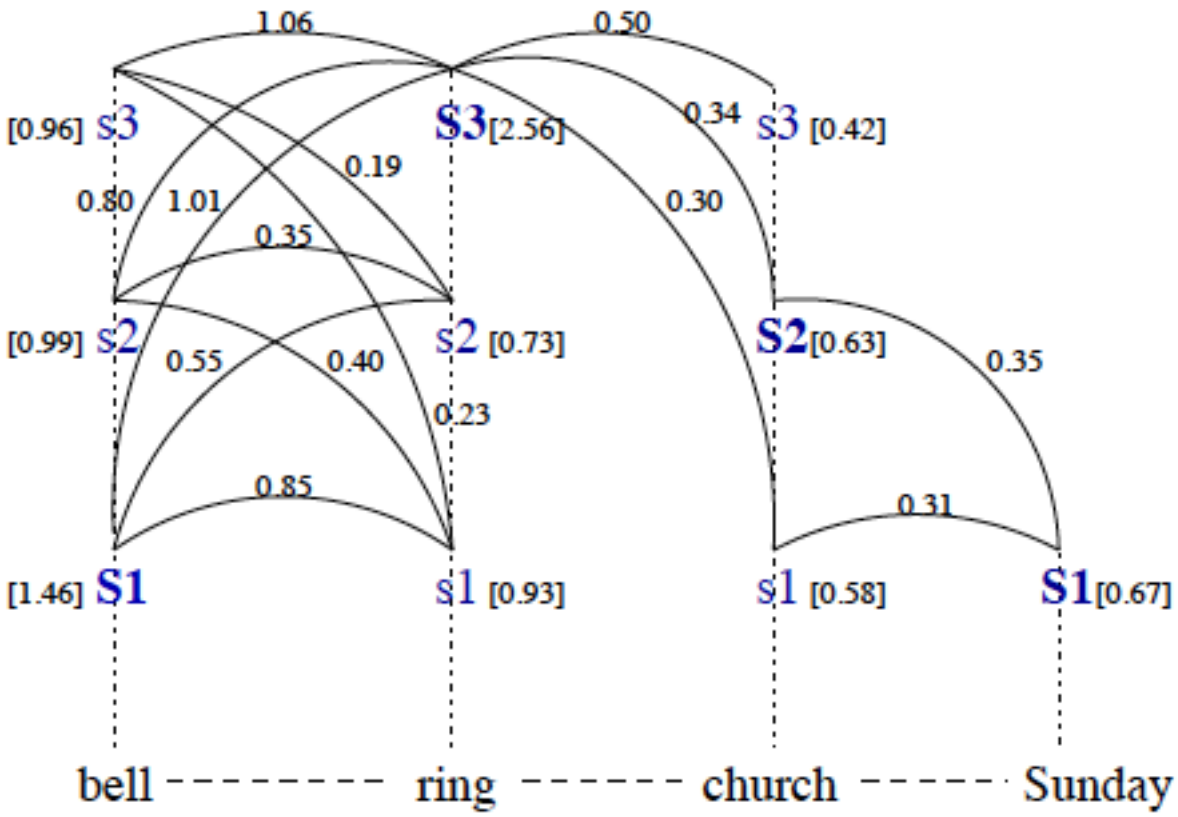
church
1: one of the groups of Christians who have their own beliefs and forms of worship
2: a place for public (especially Christian) worship
3: a service conducted in a church

bell
1: a hollow device made of metal that makes a ringing sound when struck
2: a push button at an outer door that gives a ringing or buzzing signal when pushed
3: the sound of a bell

ring
1: make a ringing sound
2: ring or echo with sound
3: make (bells) ring, often for the purposes of musical edifi ca-tion

Sunday
1: fi rst day of the week; observed as a day of rest and worship by most Christians

The church bells no longer rung on Sundays



Evaluation 1

Part-of speech	Fine-grained sense distinctions						Coarse-grained sense distinctions					
	Random baseline		Individual (Lesk)		Sequence (graph-based)		Random baseline		Individual (Lesk)		Sequence (graph-based)	
	P	R	P	R	P	R	P	R	P	R	P	R
Noun	41.4%	19.4%	50.3%	23.6%	57.5%	27.0%	42.7%	20.0%	51.4%	24.1%	58.8%	27.5%
Verb	20.7%	3.9%	30.5%	5.7%	36.5%	6.9%	22.8%	4.3%	31.9%	6.0%	37.9%	7.1%
Adjective	41.3%	9.3%	49.1%	11.0%	56.7%	12.7%	42.6%	42.6%	49.8%	11.2%	57.6%	12.9%
Adverb	44.6%	5.2%	64.6%	7.6%	70.9%	8.3%	40.7%	4.8%	65.3%	7.7%	71.9%	8.5%
ALL	37.9%	37.9%	48.7%	48.7%	54.2%	54.2%	38.7%	38.7%	49.8%	49.8%	55.3%	55.3%

Table 1: Precision and recall for graph-based sequence data labeling, individual data labeling, and random baseline, for fine-grained and coarse-grained sense distinctions.

Evaluation 2

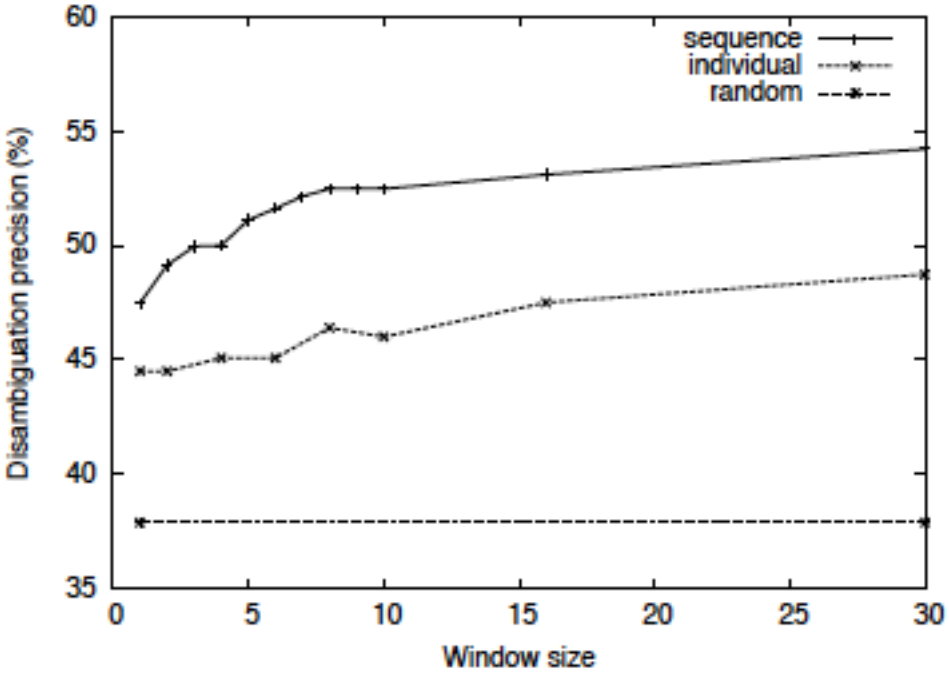


Figure 3: Disambiguation results using sequence data labeling, individual labeling, and random baseline, for various context sizes.

Conclusion

- This system gets good results despite using less expensive training data.
- We consider all sense decisions (past and future) within a given context size, which provides a significant accuracy increase.

Discussion questions

1. What might be a better method for determining the weights of the edges?
2. If we were to represent words in context more fluidly than with discrete senses, how would the model look?
3. Is WSD “worth it?”