Applying a Naive Bayes Similarity Measure to Word Sense Disambiguation

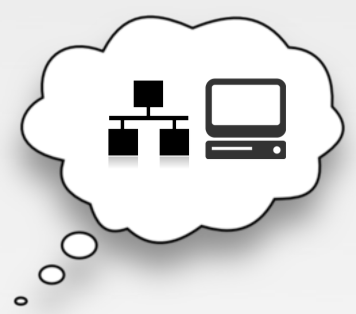
Tong Wang and Graeme Hirst
University of Toronto

Presented at *The 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, June 2014.

Financially supported by NSERC CRSNG

1. WSD

terminal (n.)



The *passengers* *disembarked onto the parking bay in front of a small* ***terminal*** *building*.

She confirmed the tape recorder was working and examined her *computer* ***terminal*** *for instructions*.

2. The (simplified) Lesk Algorithm for WSD [1]

The sense whose dictionary gloss has the highest degree of overlap with the context words is chosen as the correct sense.

ter·mi·nal (n.)

1 station where transport vehicles load or unload passengers or goods.
2 electronic equipment consisting of a device providing access to a computer.

(WordNet 3.1)

3. Limitations of String Matching

“Lesk’s approach is very sensitive to the exact wording of definitions, so the absence of a certain word can radically change the results.” [2]

Even for the two most frequent senses of the word *terminal*, only a small number of contexts actually overlap with the corresponding glosses by exact string matching:

terminal	*passenger* | *terminal*	*computer* | *terminal*
2,235	143 (6.40%)	317 (14.18%)

(Counts in the BNC)

We therefore propose a “softer” measure of gloss-context overlap using a Naive Bayes model:

$$p(\mathbf{f}|\mathbf{e}) = \prod_j p(f_j|\{e_i\}) = \prod_j \frac{p(\{e_i\}|f_j)p(f_j)}{p(\{e_i\})}$$
$$= \frac{\prod_j [p(f_j)\prod_i p(e_i|f_j)]}{\prod_j \prod_i p(e_i)} \qquad (1)$$

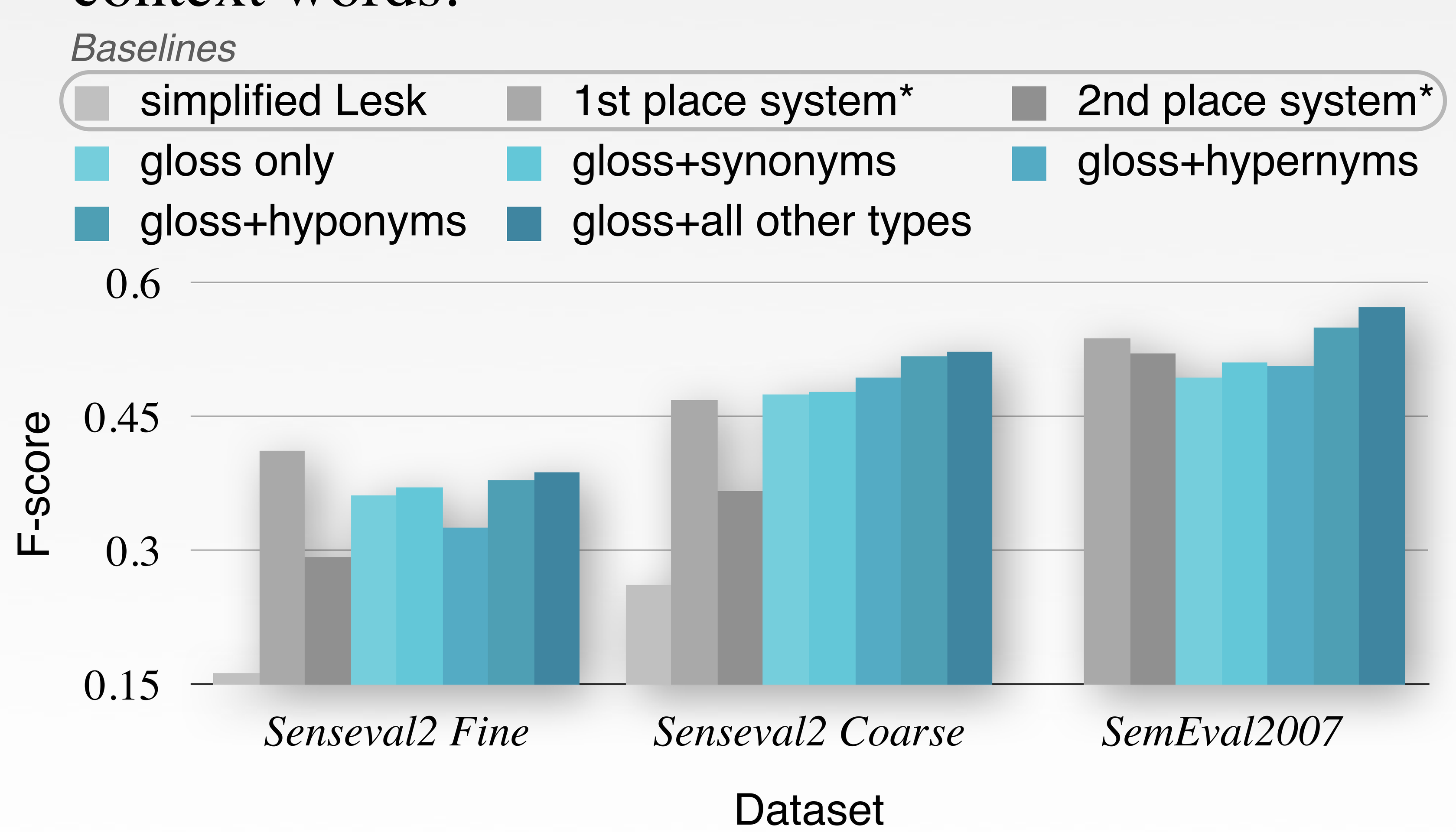
Probability estimation:

$$(1) \approx \sum_i \log\frac{c(f_j)}{c(\cdot)} + \sum_i\sum_j \log\frac{c(e_i,f_j)}{c(f_j)} - |\{f_j\}|\sum_j \log\frac{c(e_i)}{c(\cdot)}$$
$$= (1-|\{e_i\}|)\sum_i \log c(f_j) - |\{f_j\}|\sum_j \log c(e_i)$$
$$+\sum_i\sum_j \log c(e_i,f_j) + |\{f_j\}|(|\{e_i\}|-1)\log c(\cdot),$$

where $c(\cdot)$ is the corpus size.

4. Experiments

WSD accuracy on three datasets. Our probabilistic model (blue bars) uses gloss and various additional knowledge sources to overlap with context words.



* Senseval/SemEval 1st and 2nd place systems.

5. Conclusions

- Probabilistic matching significantly improves WSD accuracy over exact string matching (blue bars vs. left-most grey bars).
- Combining multiple types of lexical knowledge achieves state-of-the-art accuracy (right-most blue bars).
- Hyponyms are the most effective feature when added to gloss texts for WSD (2nd-to-right-most blue bars).

[1] Adam Kilgarriff and Joseph Rosenzweig. Framework and results for English Senseval. *Computers and the Humanities*, 34(1-2):15–48, 2000.

[2] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69, 2009.