Determining Word Sense Dominance Using a Thesaurus

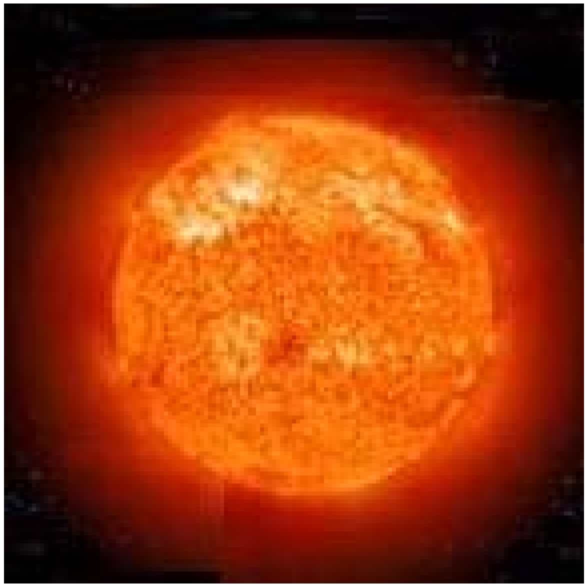
Saif Mohammad and Graeme Hirst

Department of Computer Science

University of Toronto

EACL, Trento, Italy (5th April, 2006)

Word Sense Dominance

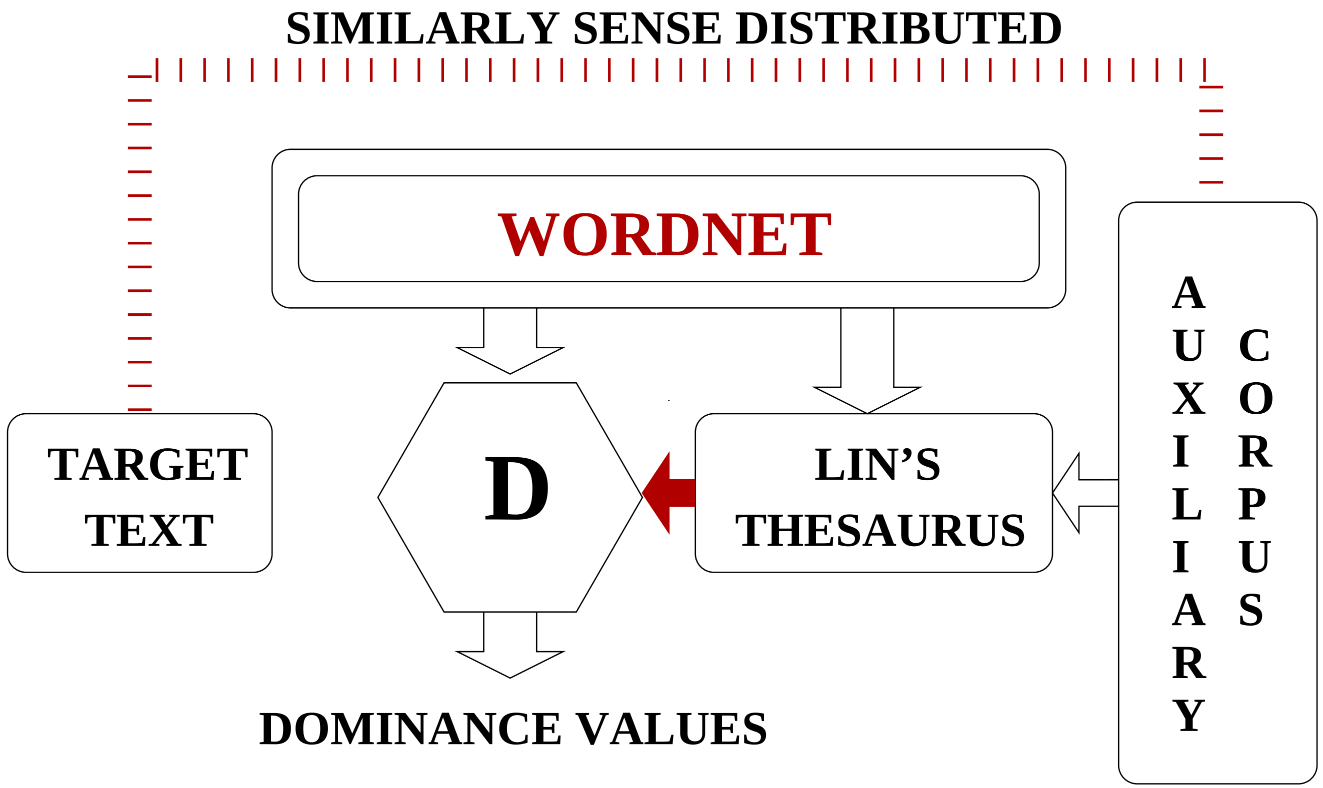


star **(CELESTIAL BODY)** ***star*** **(CELEBRITY)**

The **degree of dominance of a sense** of a word is the proportion of occurrences of that sense in text.

- Applications:
  - Sense disambiguation, document clustering, . . .

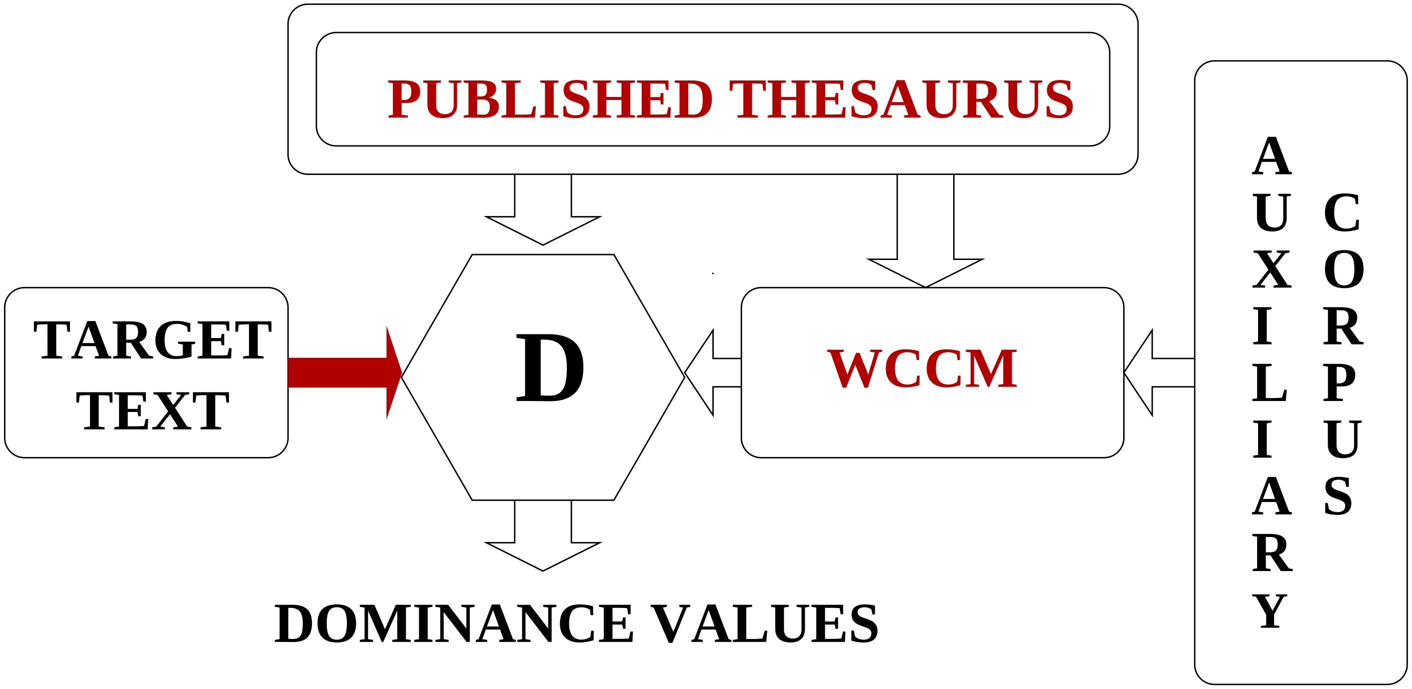
McCarthy et al.'s Method



D = dominance method

- Requires WordNet.
- Needs auxiliary text with similar sense distribution.
- Requires retraining (Lin's thesaurus).

Our Method



WCCM = word–category co-occurrence matrix

- We use a published thesaurus.
- Auxiliary text need not have similar sense distribution.
- No retraining is needed (WCCM created just once).

Published Thesauri

- E.g., *Roget's* (English), *Macquarie* (English), *Cilin* (Chinese), *Bunrui Goi Hyou* (Japanese)

- Vocabulary divided into about 1000 categories
  - Words in a category (category terms or c-terms) are closely related.
  - A category very roughly corresponds to a sense (Yarowsky, 1992).

- One word, more than one category
  - *bark* in ANIMAL NOISES and MEMBRANE.

Why a Thesaurus?

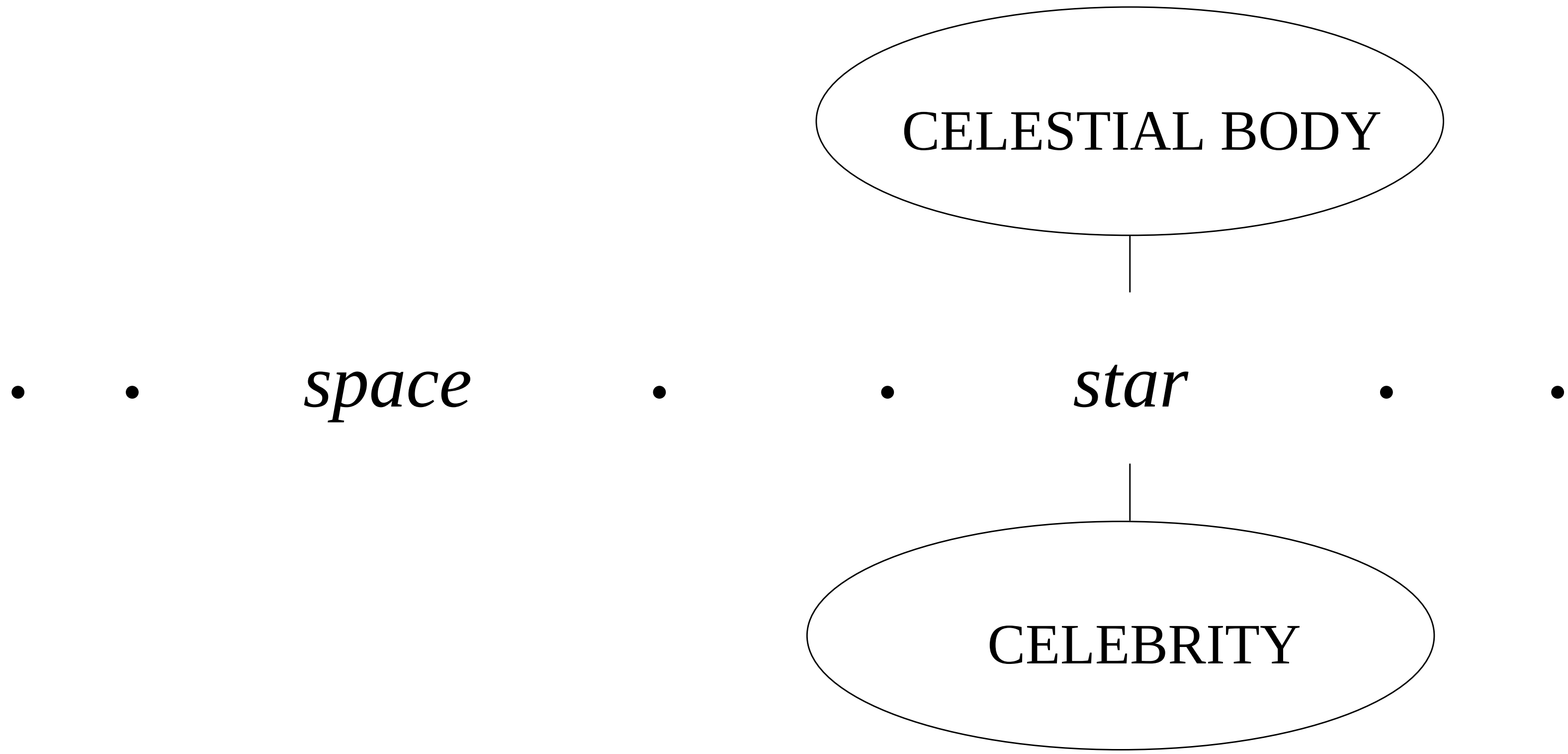
- Coarse senses: WordNet is much too fine grained.
- Computational ease: With just 1000 categories, the word–category co-occurrence matrix is of manageable size.
- Availability: Thesauri are available in many languages.
- Words for a sense: Each sense can be represented unambiguously with a set of (possibly ambiguous) words.

Word–Category Matrix

	c_1	c_2	$\ldots$	c_j	$\ldots$
w_1	m_{11}	m_{12}	$\ldots$	m_{1j}	$\ldots$
w_2	m_{21}	m_{22}	$\ldots$	m_{2j}	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ldots$	$\ldots$
w_i	m_{i1}	m_{i2}	$\ldots$	m_{ij}	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

- WCCM: categories (thesaurus) vs. words (vocabulary)
- Cell m_{ij}: number of times word w_i co-occurs with a c-term listed in category c_j
- Text: most of the *BNC*

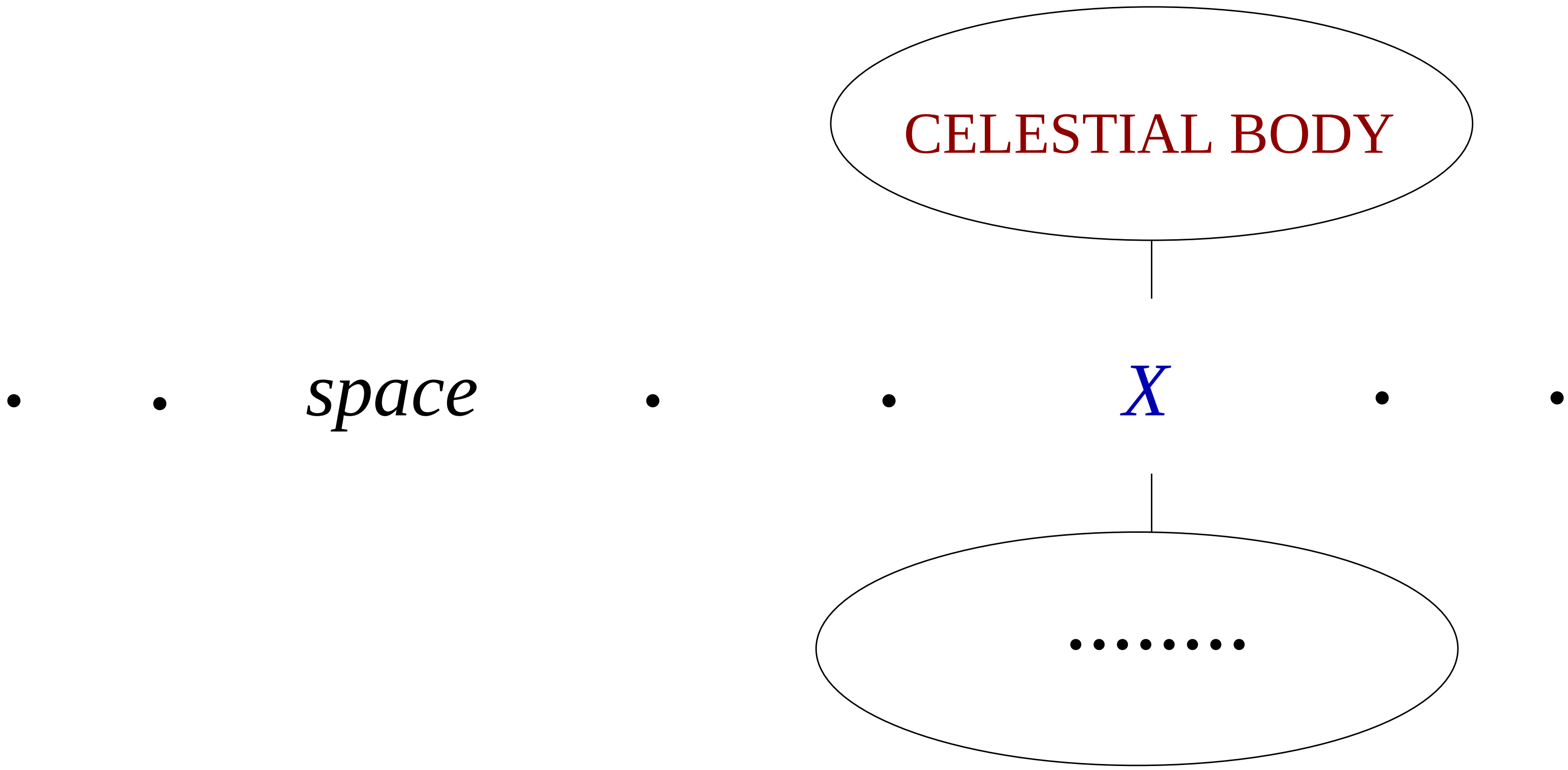
Example



cell (space, CELESTIAL BODY) incremented by 1

cell (space, CELEBRITY) incremented by 1

Example (continued)



X: star, nova, constellation, sun, empty, web

Word–Category Matrix

	c_1	c_2	$\ldots$	CELESTIAL BODY	$\ldots$
w_1	m_{11}	m_{12}	$\ldots$	m_{1j}	$\ldots$
w_2	m_{21}	m_{22}	$\ldots$	m_{2j}	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ldots$	$\ldots$
space	m_{i1}	m_{i2}	$\ldots$	$m \uparrow\uparrow$	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

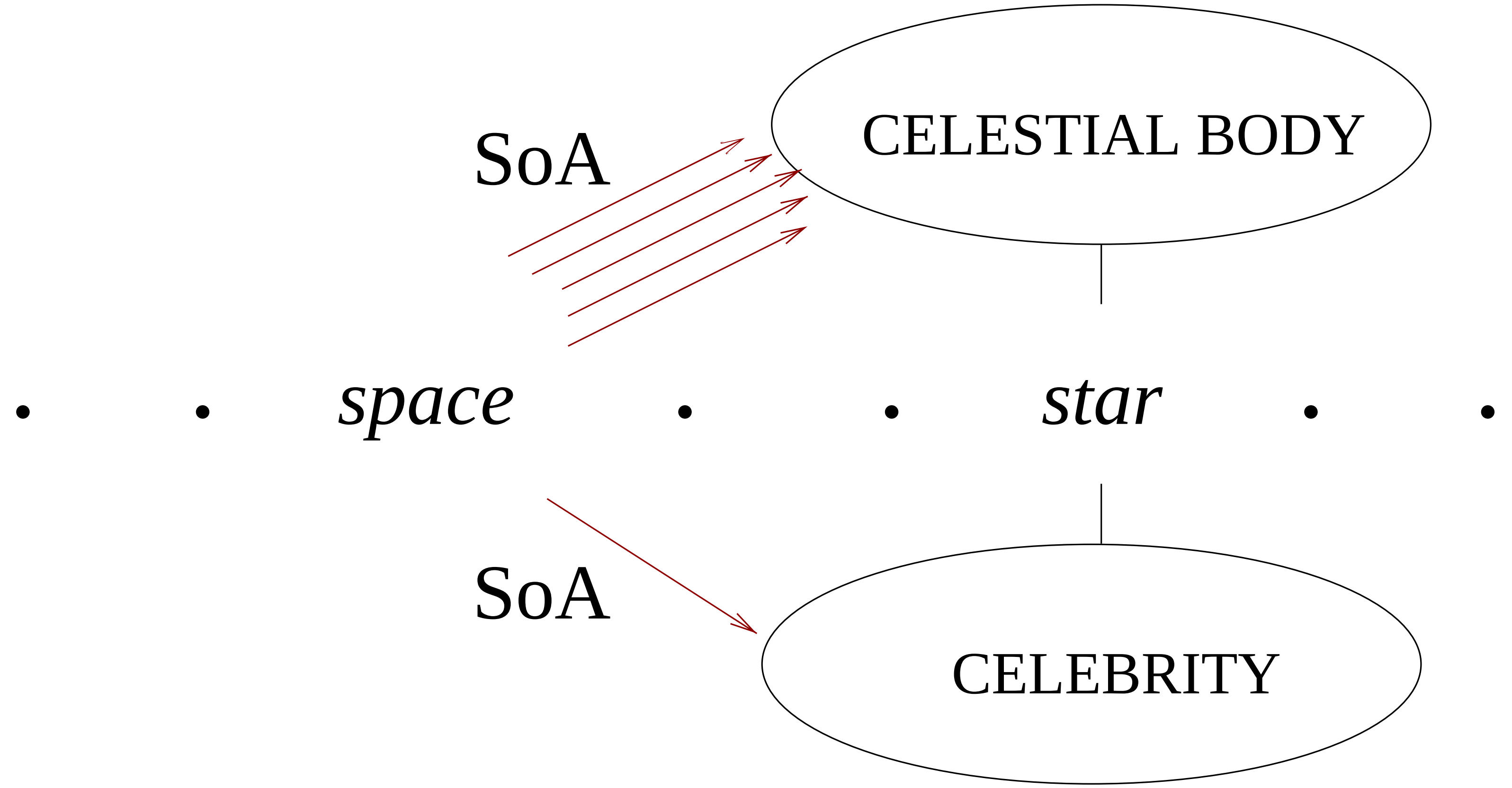
Contingency Table for w and c

	c	$\neg c$
w	n_{wc}	$n_{w\neg}$
$\neg w$	$n_{\neg c}$	$n_{\neg\neg}$

Applying a statistic gives the strength of association (SoA)

- cosine
- Dice
- odds ratio
- pointwise mutual information
- Yule's coefficient of colligation

Evidence for the Senses



Base WCCM

- Matrix created after the first pass of unannotated text
  - noisy
  - captures strong associations

- Words that occur close to a target word
  - Good indicators of intended sense
  - Co-occurrence frequency used as evidence

Bootstrapping the WCCM

- Second pass of the auxiliary corpus
  - Word sense disambiguation: using co-occurring words and evidence from base WCCM

- New, more accurate, WCCM
  - Cell m_{ij}: number of times word used in sense c_j co-occurs with w_i

Four Methods

	Weighted voting	Unweighted voting
Implicit sense disambiguation	$D_{I,W}$	$D_{I,U}$
Explicit sense disambiguation	$D_{E,W}$	$D_{E,U}$

The stronger the association of a sense with its co-occurring words, the higher is its dominance.

- Weighted vote (SoA) to each sense or unweighted vote to sense with the highest SoA
- Explicit word sense disambiguation or not

Method: $D_{I,W}$

- Each word that co-occurs with the target word t gives a weighted vote (SoA) to each sense.
- Dominance of a sense c is the proportion of votes it gets.

$$D_{I,W}(t,c) = \frac{\sum_{w \in T} SoA(w,c)}{\sum_{c' \in senses(t)} \sum_{w \in T} SoA(w,c')}$$

T is the set of all words that co-occur with t.

Method: $D_{I,U}$

- Each word that co-occurs with the target word gives an unweighted, equal vote to a winner sense.
  - Sense with highest strength of association with co-occurring words

- Dominance of a sense is the proportion of votes it gets.

$$D_{I,U}(t,c) = \frac{|\{w \in T : \text{argmax}_{c' \in \text{senses}(t)} SoA(w,c') = c\}|}{|T|}$$

Methods: $D_{E,W}$ and $D_{E,U}$

- Explicit sense disambiguation
  - Votes from co-occurring words
  - Votes can be weighted or unweighted

- Dominance of a sense
  - Proportion of occurrences pertaining to that sense

Experimental Setup

- Naïve sense disambiguation system
  - Gives predominant sense as output

- Test datasets
  - Different sense distributions of the two most dominant senses of each target word

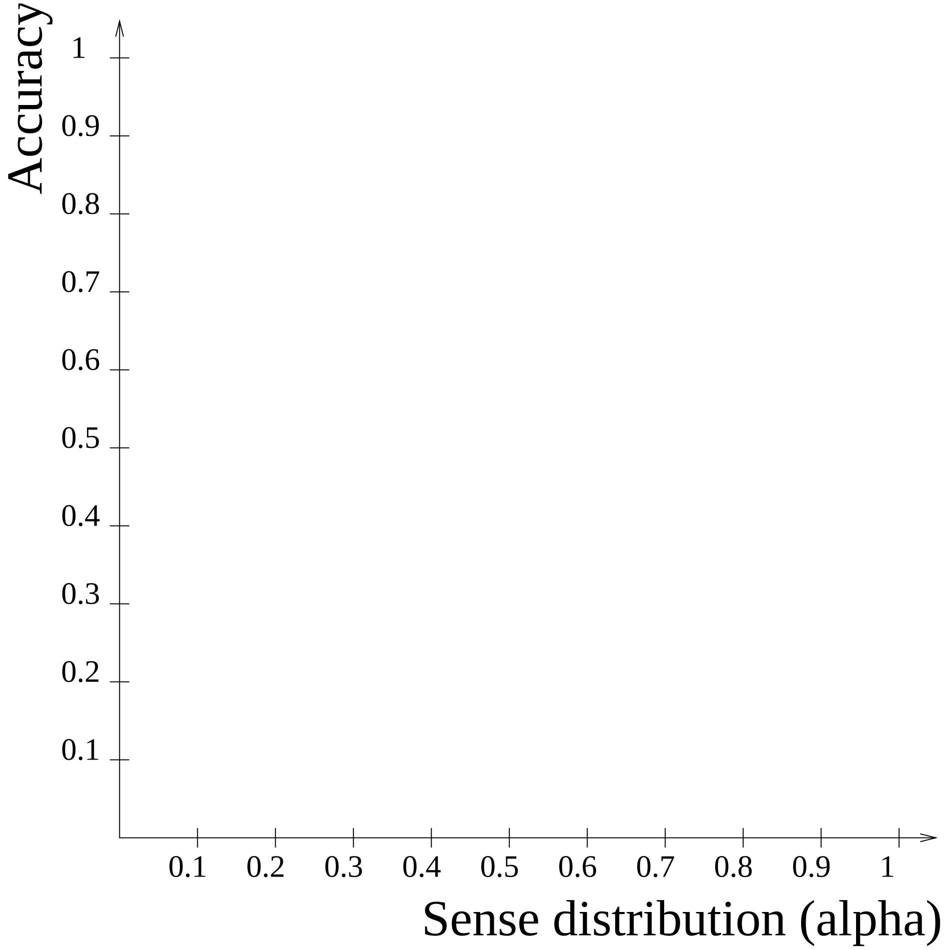
Sense-tagged Data

We created pseudo-thesaurus-sense-tagged data for the 27 head words in SENSEVAL-1 English Sample Space using the held out subset of *BNC*.

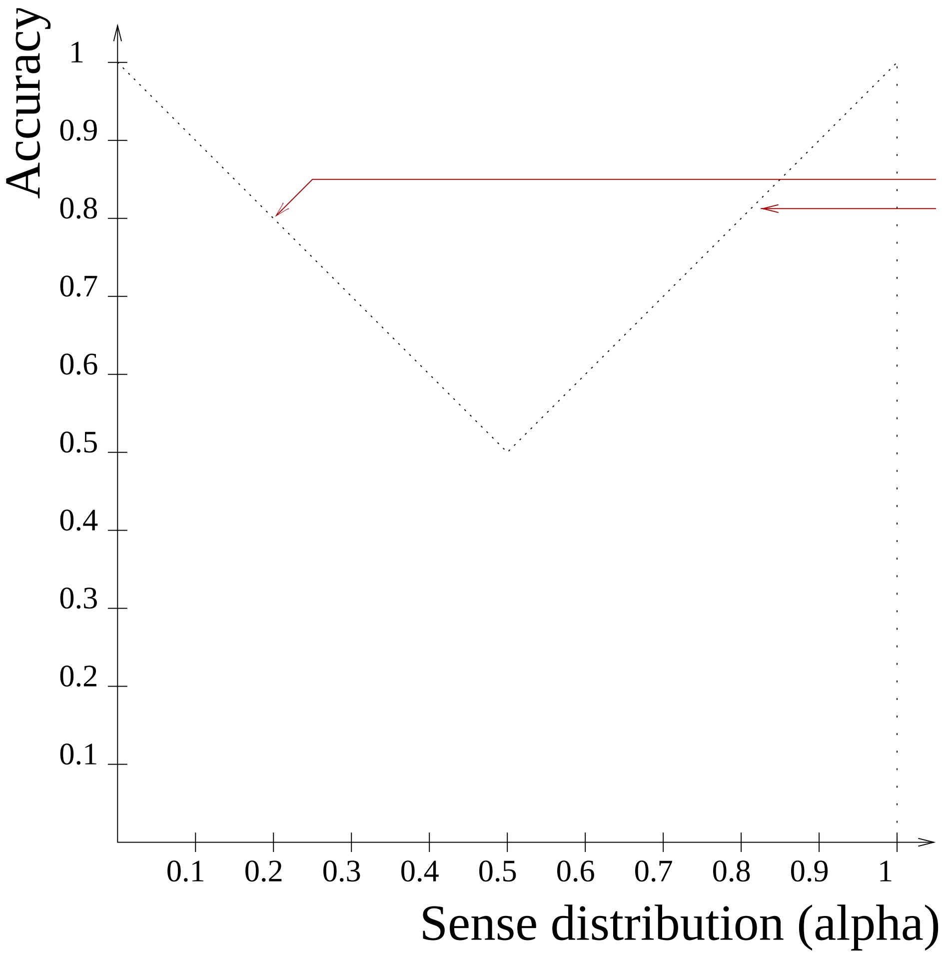
Non-monosemous target word: *brilliant*
Category: INTELLIGENCE

Monosemous c-term: *clever*
Sentence from auxiliary text:
Hermione had a clever plan.
Sense annotated sentence:
Hermione had a brilliant//INTELLIGENCE plan.

Best Results

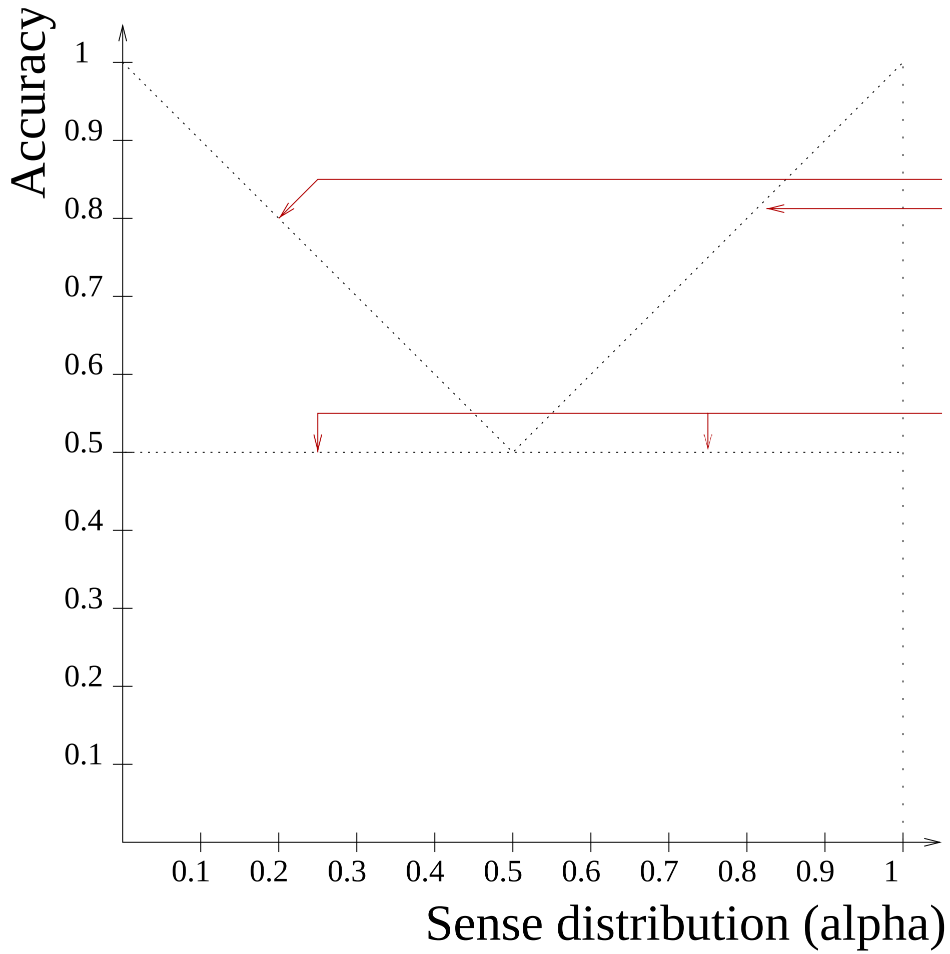


Best Results



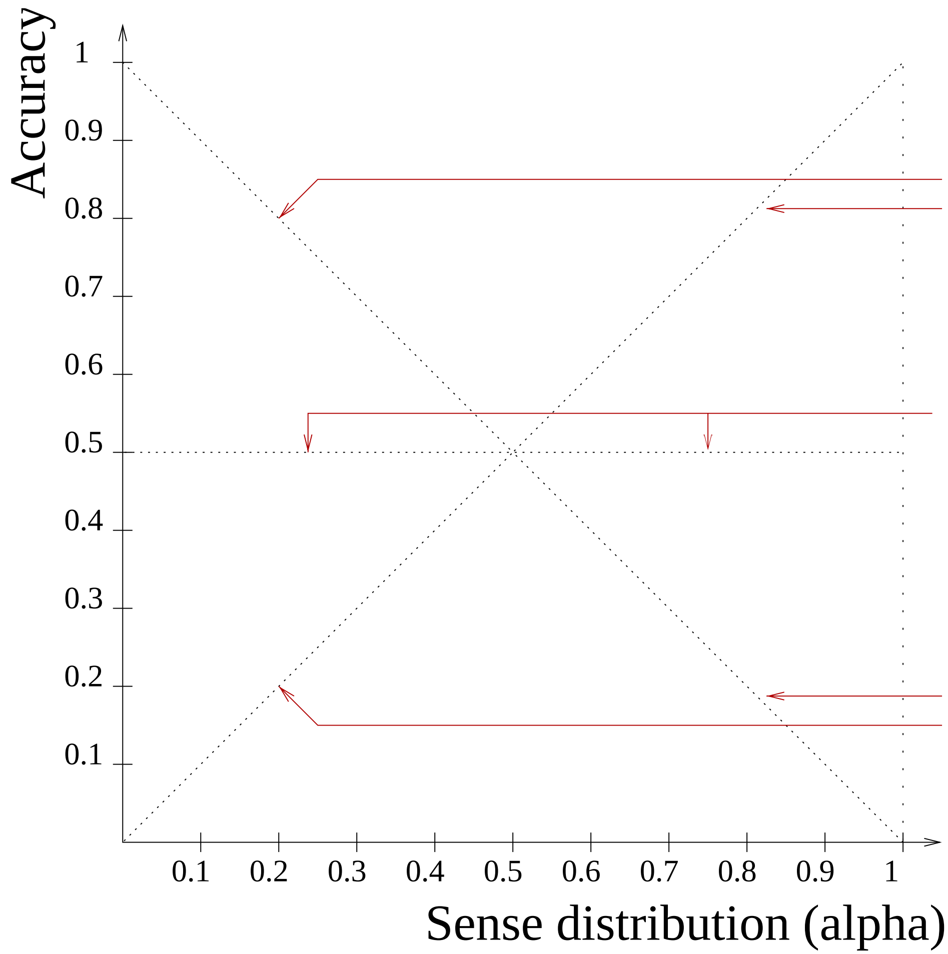
- Upper bound

Best Results



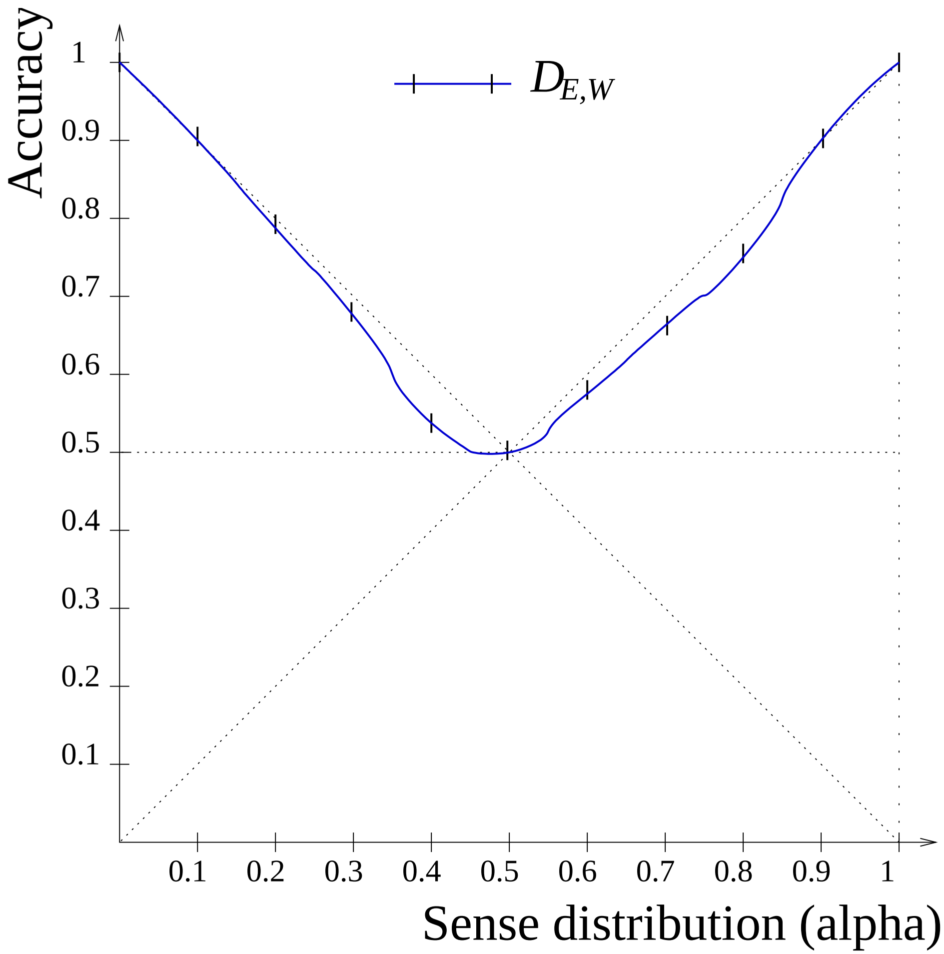
- Upper bound
- Random baseline

Best Results



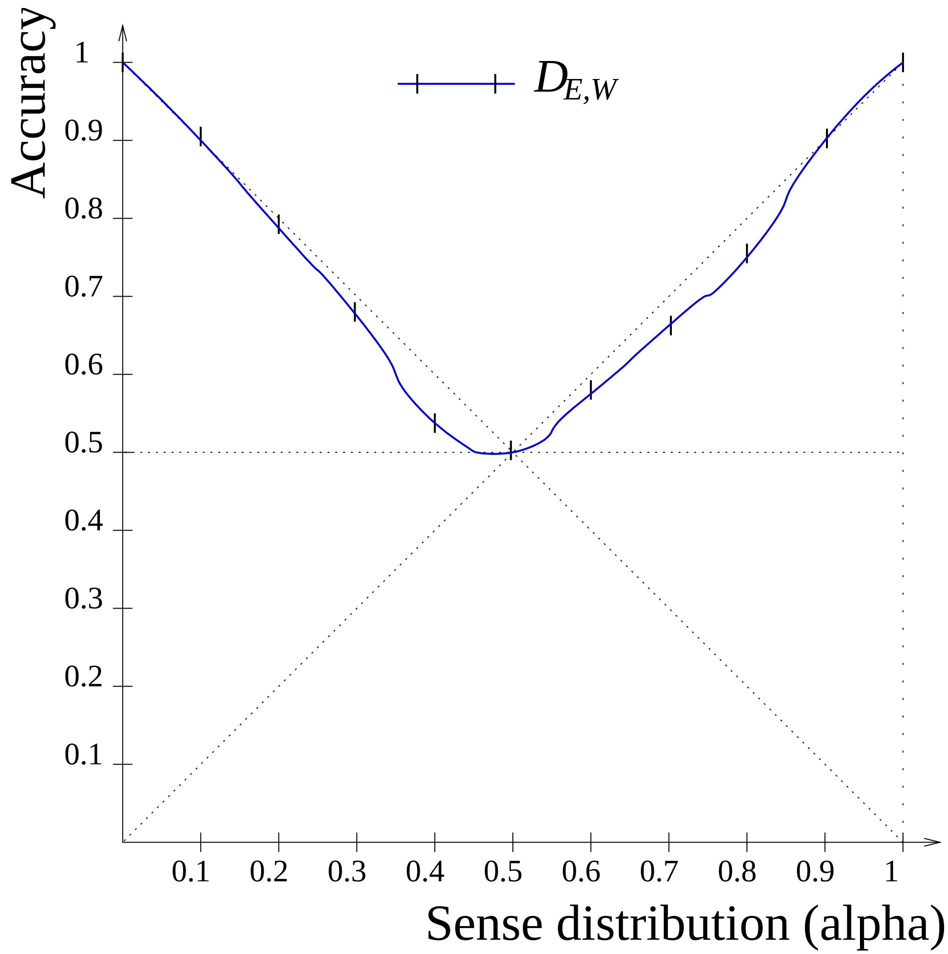
- Upper bound
- Random baseline
- Lower bound

Best Results



Mean distance below upper bound

Best Results

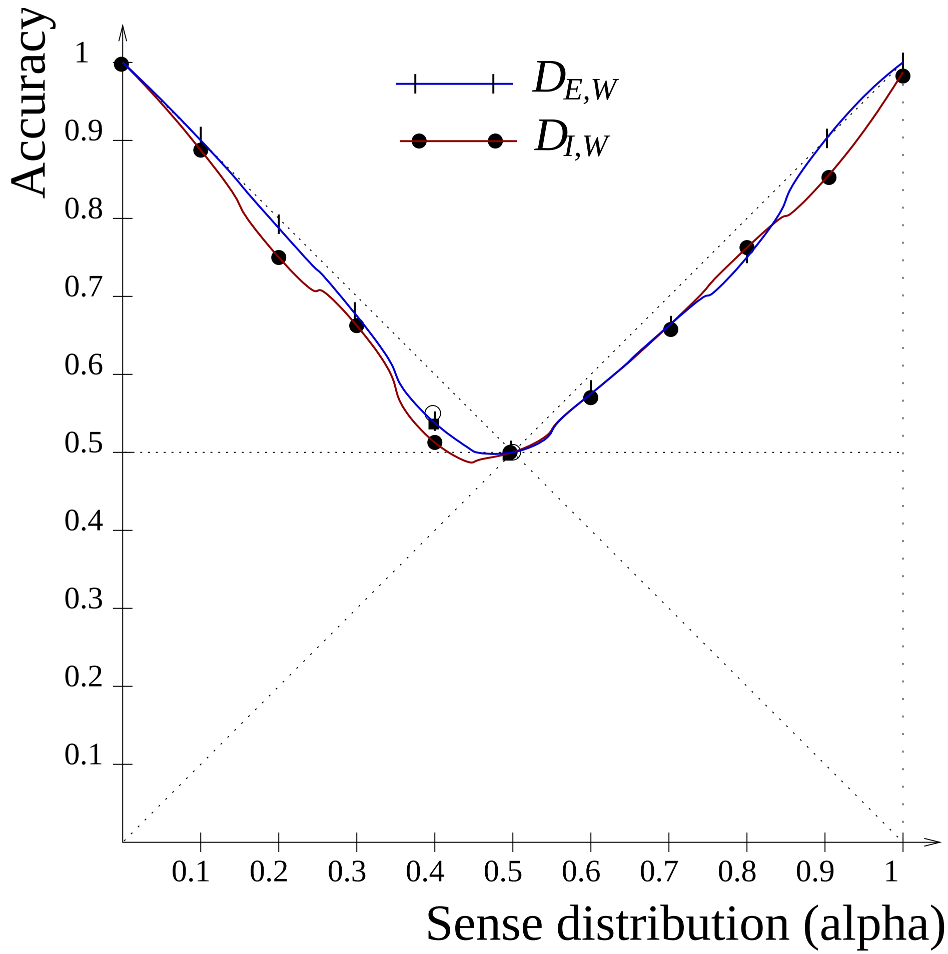


Mean distance below upper bound

- $D_{E,W}$: .02

  pmi, odds, Yule

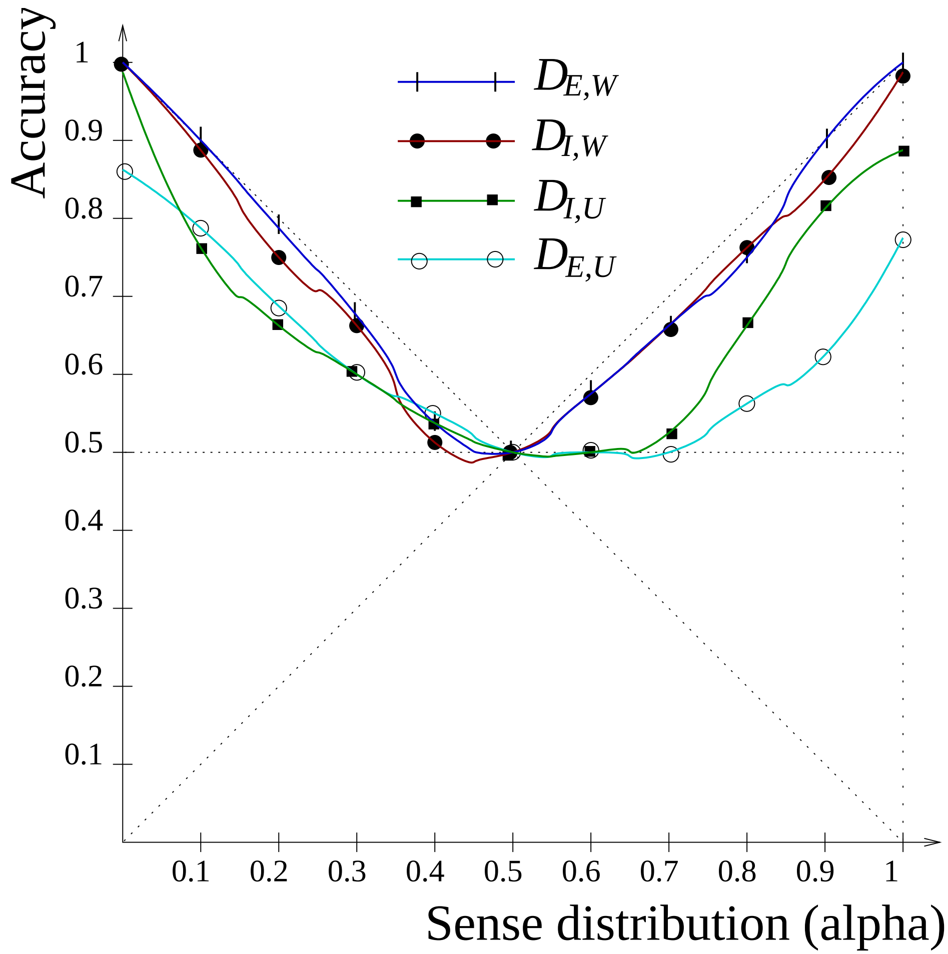
Best Results



Mean distance below upper bound

- $D_{E,W}$: .02
  pmi, odds, Yule
- $D_{I,W}$: .03
  pmi

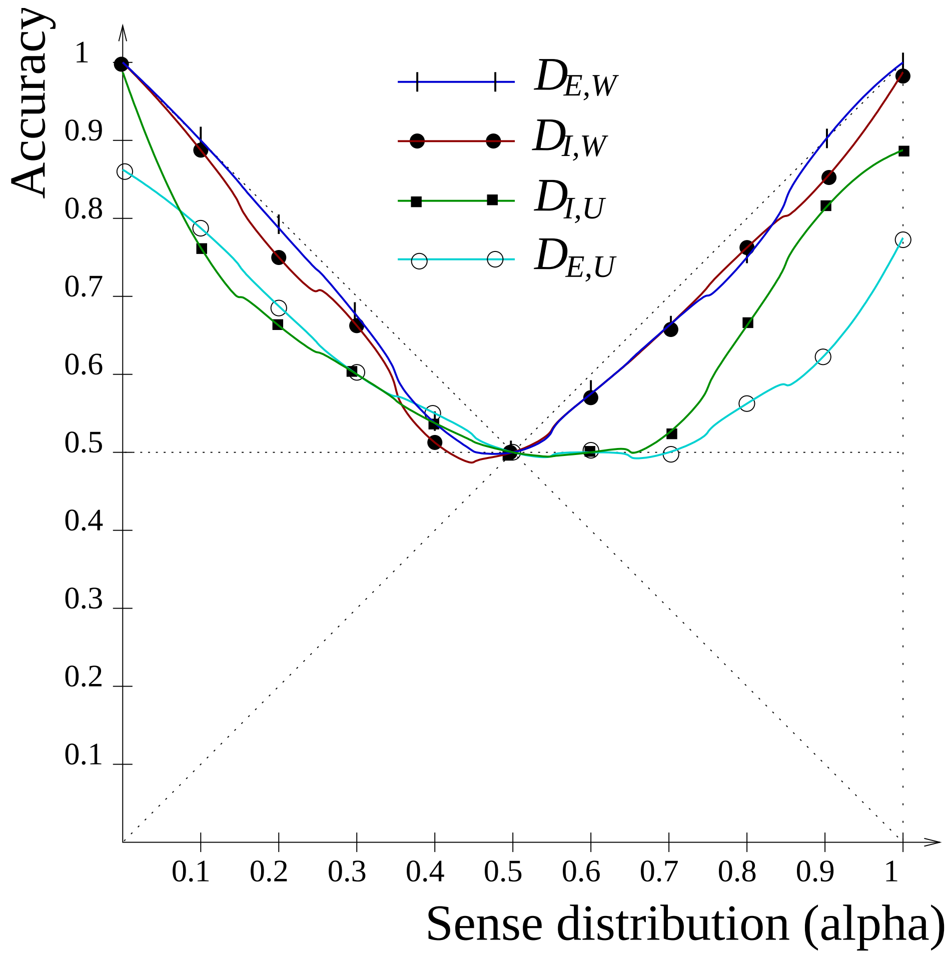
Best Results



Mean distance below upper bound

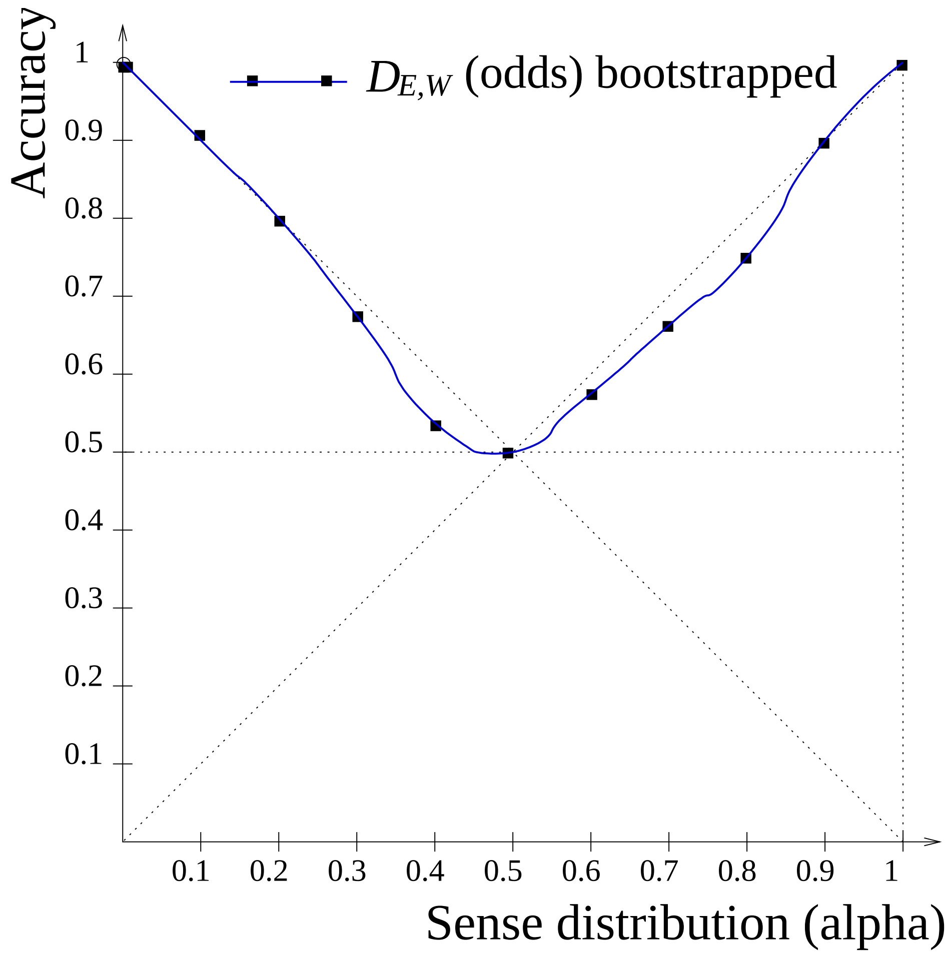
- $D_{E,W}$: .02
  pmi, odds, Yule
- $D_{I,W}$: .03
  pmi
- $D_{I,U}$: .11
  phi, pmi, odds, Yule
- $D_{E,U}$: .16
  phi, pmi, odds, Yule

Observations

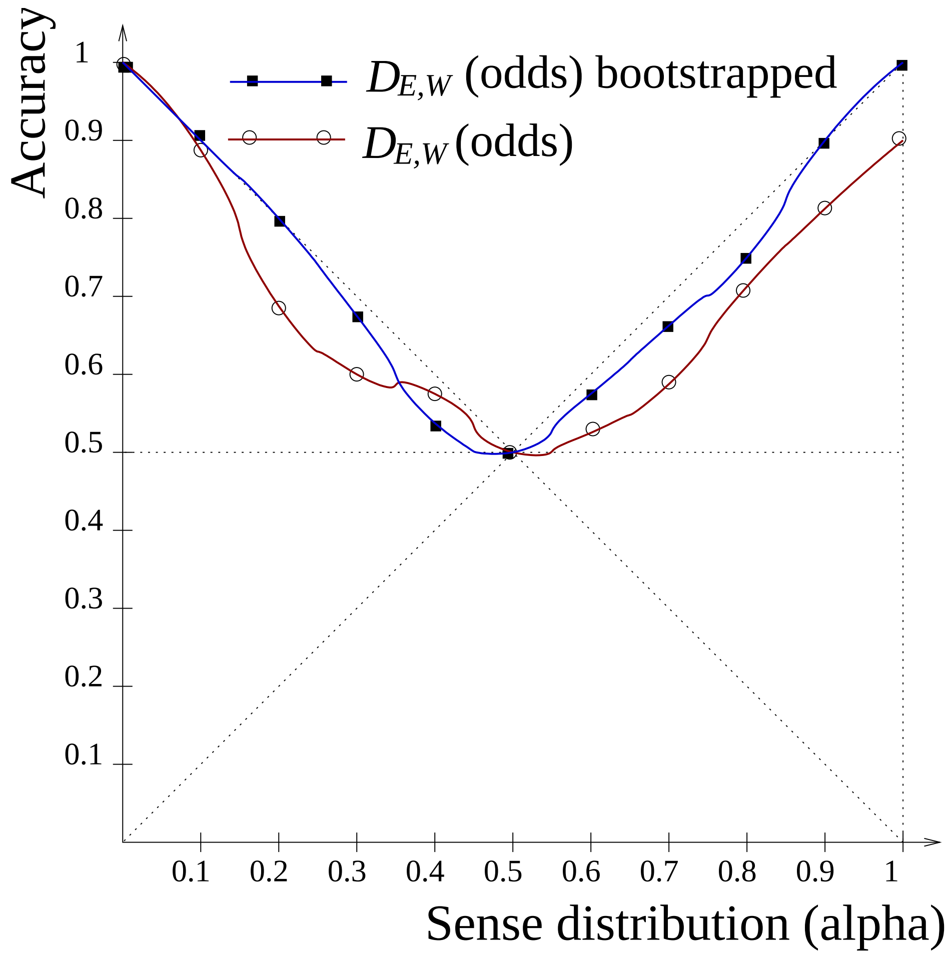


- Weighted methods are better
- Explicit or implicit disambiguation does not matter
- Odds, pmi, and Yule are better

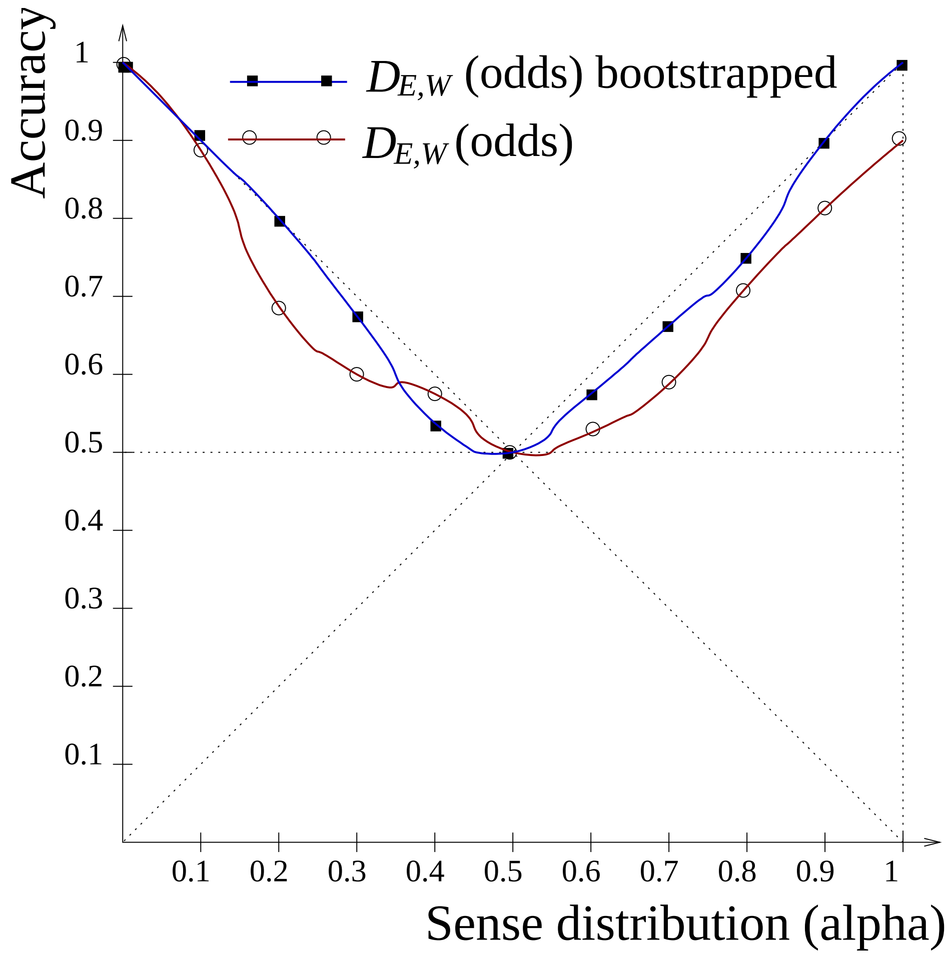
Effect of Bootstrapping



Effect of Bootstrapping



Effect of Bootstrapping



- Most of the gain given by the first iteration.
- Relative behavior of measures more or less the same.

In Summary

- New methods for determining sense dominance
  - Raw text and a published thesaurus
  - No similarly-sense-distributed text or re-training
- Extensive experiments
  - Synthetically created thesaurus-sense-tagged data
- Results are close to the upper bound

Future Work

WCCM has applications beyond sense dominance.

- Linguistic distances

  Distributional distance of concepts

- Word sense disambiguation

  Unsupervised naïve Bayes classifier

- Machine translation

  Domain-specific translational dominance

- Document clustering

  Represent document in concept space