# Measuring Semantic Distance

## using Distributional Profiles of Concepts

Saif Mohammad

Department of Computer Science

University of Toronto

**Grateful acknowledgments:** Graeme Hirst (advisor and co-author);
Iryna Gurevych, Torsten Zesch, and Philip Resnik (co-authors);
Rada Mihalcea, Renee Miller, Gerald Penn, Suzanne Stevenson,
University of Toronto (especially the CL group), and NSERC.

Graeme Hirst
University of Toronto

# Semantic Distance

SALSA DANCE

CLOWN BRIDGE

A measure of how close or distant two units of language are in terms of their meaning

# Why measure semantic distance?

- Natural language processing is teeming with semantic-distance problems:
  - Machine translation

*You know a person by the company they keep*

bag of hypotheses

*Das Wesen eines Menschen erkennt man an der Gesellschaft, mitder er sich umgibt*

# Why measure semantic distance?

- Natural language processing is teeming with semantic-distance problems:
  - Word sense disambiguation

*Hermione cast a bewitching spell*

CHARM OR INCANTATION

bag of hypotheses

# Why measure semantic distance?

- Natural language processing is teeming with semantic-distance problems:
  - Speech recognition, real-word spelling correction

*. . . interest . . . money . . . band . . . loan . . .*

bag of hypotheses

*bank* or *bond*

# Why are some pairs semantically close?

Two words (more precisely, two lexical units) are considered semantically close (or, semantically related) if there is a lexical-semantic relation between them.

- Lexical unit: a combination of surface form and word sense

# Lexical-semantic relations: classical

- Near-synonymy
  - *error–blunder, taxi–cab*
- Hypernymy–hyponymy (is a, a kind of, subsumes)
  - *mammal–elephant, furniture–table*
- Common subsumer
  - *elephant–hippo, table–chair*

Only these relations make a term pair semantically similar.

# Lexical-semantic relations: classical

- Meronymy–holonymny (part of, has a)
  - *boat–rudder, tree–forest*
- Antonymy (opposite of)
  - *hot–cold, dull–interesting*

# Lexical-semantic relations: non-classical

- Typical agent–instrument, agent–action, agent–cause, action–patient pairs
  - *surgeon–scalpel*, *dog–bark*, *virus–disease*
- Adhoc relationships
  - *matches*, *swiss-knife*, *tent*, *rope*

# Knowledge source–based semantic measures

- Structure of a network or resource
  - The nodes represent senses or concepts
  - Examples: Resnik (1995), Jiang and Conrath (1997)
- Drawbacks
  - Resource bottleneck
  - Not easily domain-adaptable
  - Accuracy on pairs other than noun–noun is poor
  - Relatedness estimation is poor

# Corpus-based distributional measures

- Words in similar contexts are close.
  - **Distributional profile (DP)** of a word: strength of association of the word with co-occurring words in text

# DP of a word


## DP of *fusion*

*heat* 0.16
*hydrogen* 0.16
*energy* 0.13
*hot* 0.09
*light* 0.09
*space* 0.04
*gravity* 0.03
*pressure* 0.03

# DPs of words


**DP of *star***

*space* 0.21
*movie* 0.16
*famous* 0.15
*light* 0.12
*rich* 0.11
*heat* 0.08
*planet* 0.07
*hydrogen* 0.07

**DP of *fusion***

*heat* 0.16
*hydrogen* 0.16
*energy* 0.13
*hot* 0.09
*light* 0.09
*space* 0.04
*gravity* 0.03
*pressure* 0.03

# Distance between two words


**DP of *star***

*space* 0.21
*movie* 0.16
*famous* 0.15
*light* 0.12
*rich* 0.11
*heat* 0.08
*planet* 0.07
*hydrogen* 0.07

**DP of *fusion***

*heat* 0.16
*hydrogen* 0.16
*energy* 0.13
*hot* 0.09
*light* 0.09
*space* 0.04
*gravity* 0.03
*pressure* 0.03

# Distance between two words


**DP of *star***

*space* 0.21
*movie* 0.16
*famous* 0.15
*light* 0.12
*rich* 0.11
*heat* 0.08
*planet* 0.07
*hydrogen* 0.07

**DP of *fusion***

*heat* 0.16
*hydrogen* 0.16
*energy* 0.13
*hot* 0.09
*light* 0.09
*space* 0.04
*gravity* 0.03
*pressure* 0.03

# Distance between two words


**DP of *star***

*space* 0.21
*movie* 0.16
*famous* 0.15
*light* 0.12
*rich* 0.11
*heat* 0.08
*planet* 0.07
*hydrogen* 0.07

**DP of *fusion***

*heat* 0.16
*hydrogen* 0.16
*energy* 0.13
*hot* 0.09
*light* 0.09
*space* 0.04
*gravity* 0.03
*pressure* 0.03

# Distance between two words

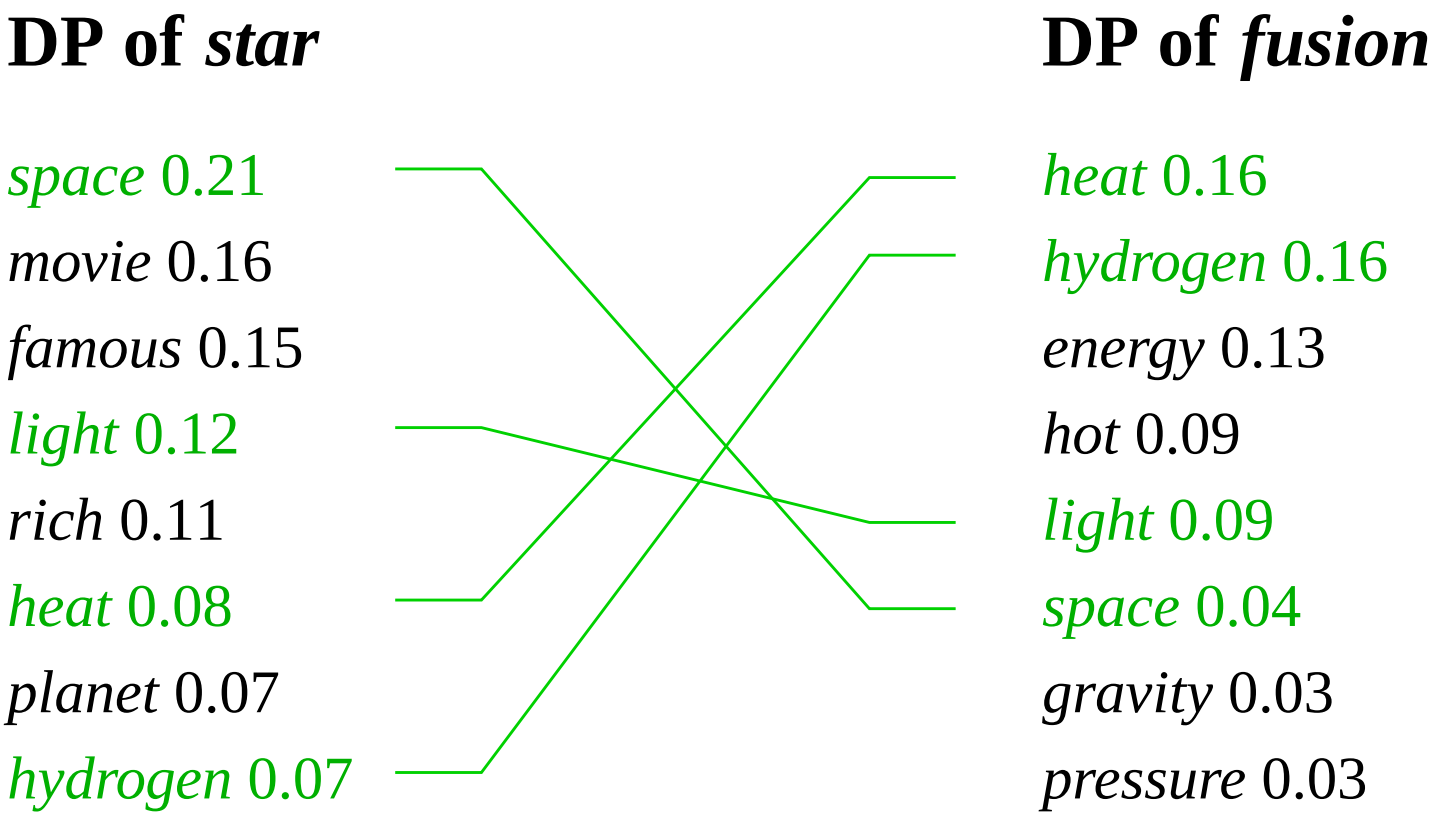

**DP of *star***

*space* 0.21
*movie* 0.16
*famous* 0.15
*light* 0.12
*rich* 0.11
*heat* 0.08
*planet* 0.07
*hydrogen* 0.07

**DP of *fusion***

*heat* 0.16
*hydrogen* 0.16
*energy* 0.13
*hot* 0.09
*light* 0.09
*space* 0.04
*gravity* 0.03
*pressure* 0.03

# Distance between two words



# Distributional measures of word-distance

- Words in similar contexts are close.
  - Distributional profile (DP) of a word: strength of association of the word with co-occurring words (text)
  - Distributional measure: distance between DPs

    Cosine, Lin, $\alpha$-skew divergence
- Drawback
  - Poor accuracy (albeit higher coverage)
    - Conflation of word senses

# Problem with distributional word-distance measures

**DP of *star***

*space* 0.21
*movie* 0.16
*famous* 0.15
*light* 0.12
*rich* 0.11
*heat* 0.08
*planet* 0.07
*hydrogen* 0.07

**DP of *fusion***

*heat* 0.16
*hydrogen* 0.16
*energy* 0.13
*hot* 0.09
*light* 0.09
*space* 0.04
*gravity* 0.03
*pressure* 0.03

# Problem with distributional word-distance measures

| DP of *star* | DP of *fusion* |
|---|---|
| *space* 0.21 | *heat* 0.16 |
| *movie* 0.16 ← | *hydrogen* 0.16 |
| *famous* 0.15 ← | *energy* 0.13 |
| *light* 0.12 | *hot* 0.09 |
| *rich* 0.11 ← | *light* 0.09 |
| *heat* 0.08 | *space* 0.04 |
| *planet* 0.07 | *gravity* 0.03 |
| *hydrogen* 0.07 | *pressure* 0.03 |

Word sense ambiguity reduces accuracy of distance measures

# Shared limitations


- Precomputing all distances is computationally expensive
  - WordNet-based measures:

    $117,000 \times 117,000$ sense–sense distance matrix
  - Distributional measures:

    $100,000 \times 100,000$ word–word distance matrix

- Monolingual

# A new hybrid approach


- Combines a knowledge source with text
  - Thesaurus categories: concepts/coarse senses
  - Most published thesauri: around 1000 categories

- Profiles concepts (rather than words)
  - Uses sets of words to represent each concept
  - Creates profiles using bootstrapping

# Features


- Can be used in real-time applications
  - Concept–concept distance matrix: only $1000 \times 1000$
- Accurate for all pos–pos pairs
  - Not just noun–noun
- Capable of giving both similarity and relatedness values
- Easily domain adaptable
- Cross-lingual

# Problem with distributional word-distance measures

**DP of *star***

*space* 0.21

*movie* 0.16 ⬅

*famous* 0.15 ⬅

*light* 0.12

*rich* 0.11 ⬅

*heat* 0.08

*planet* 0.07

*hydrogen* 0.07

Word sense ambiguity reduces accuracy of distance measures

# Solution:
# tease out the senses

***star***

*space*

*movie* ⟵

*famous* ⟵

*light*

*rich* ⟵

*heat*

*planet*

*hydrogen*

# Solution:
# tease out the senses

***star***

*space*

*movie* ←

*famous* ←

*light*

*rich* ←

*heat*

*planet*

*hydrogen*

Profile the senses separately.

# Distributional profiles of concepts

DPs of the concepts referred to by *star*:

| **DP of CELESTIAL BODY** | **DP of CELEBRITY** |
|---|---|
| *space* 0.36 | *famous* 0.24 |
| *light* 0.27 | *movie* 0.14 |
| *heat* 0.11 | *rich* 0.14 |
| *planet* 0.07 | *fan* 0.10 |
| *hydrogen* 0.06 | *hot* 0.04 |
| *hot* 0.01 | *fashion* 0.01 |

# Distributional profiles of concepts

DPs of the concepts referred to by *star*:

| **DP of CELESTIAL BODY** | **DP of CELEBRITY** |
|---|---|
| *(celestial body, star, sun,...)* | *(celebrity, hero, star,...)* |
| *space* 0.36 | *famous* 0.24 |
| *light* 0.27 | *movie* 0.14 |
| *heat* 0.11 | *rich* 0.14 |
| *planet* 0.07 | *fan* 0.10 |
| *hydrogen* 0.06 | *hot* 0.04 |
| *hot* 0.01 | *fashion* 0.01 |

# Distance: *star* and *fusion*


**DP of FUSION**

*(atomic reaction, fusion, thermonuclear reaction,...)*

*heat* 0.16

*hydrogen* 0.16

*energy* 0.13

*hot* 0.09

*light* 0.09

*space* 0.04

# Distance: *star* and *fusion*

| **DP of CELEBRITY** | **DP of FUSION** |
|---|---|
| *(celebrity, hero, star,...)* | *(atomic reaction, fusion, thermonuclear reaction,...)* |
| *famous* 0.24 | *heat* 0.16 |
| *movie* 0.14 | *hydrogen* 0.16 |
| *rich* 0.14 | *energy* 0.13 |
| *fan* 0.10 | *hot* 0.09 |
| *hot* 0.04 | *light* 0.09 |
| *fashion* 0.01 | *space* 0.04 |

First, consider the CELEBRITY sense of *star*.

# Distance: *star* and *fusion*


| **DP of CELEBRITY** | **DP of FUSION** |
|---|---|
| *(celebrity, hero, star,...)* | *(atomic reaction, fusion, thermonuclear reaction,...)* |
| *famous* 0.24 | *heat* 0.16 |
| *movie* 0.14 | *hydrogen* 0.16 |
| *rich* 0.14 | *energy* 0.13 |
| *fan* 0.10 | *hot* 0.09 |
| *hot* 0.04 | *light* 0.09 |
| *fashion* 0.01 | *space* 0.04 |

First, consider the CELEBRITY sense of *star*.

- Distributionally NOT close

# Distance: *star* and *fusion*


**DP of FUSION**

*(atomic reaction, fusion, thermonuclear reaction,...)*

*heat* 0.16

*hydrogen* 0.16

*energy* 0.13

*hot* 0.09

*light* 0.09

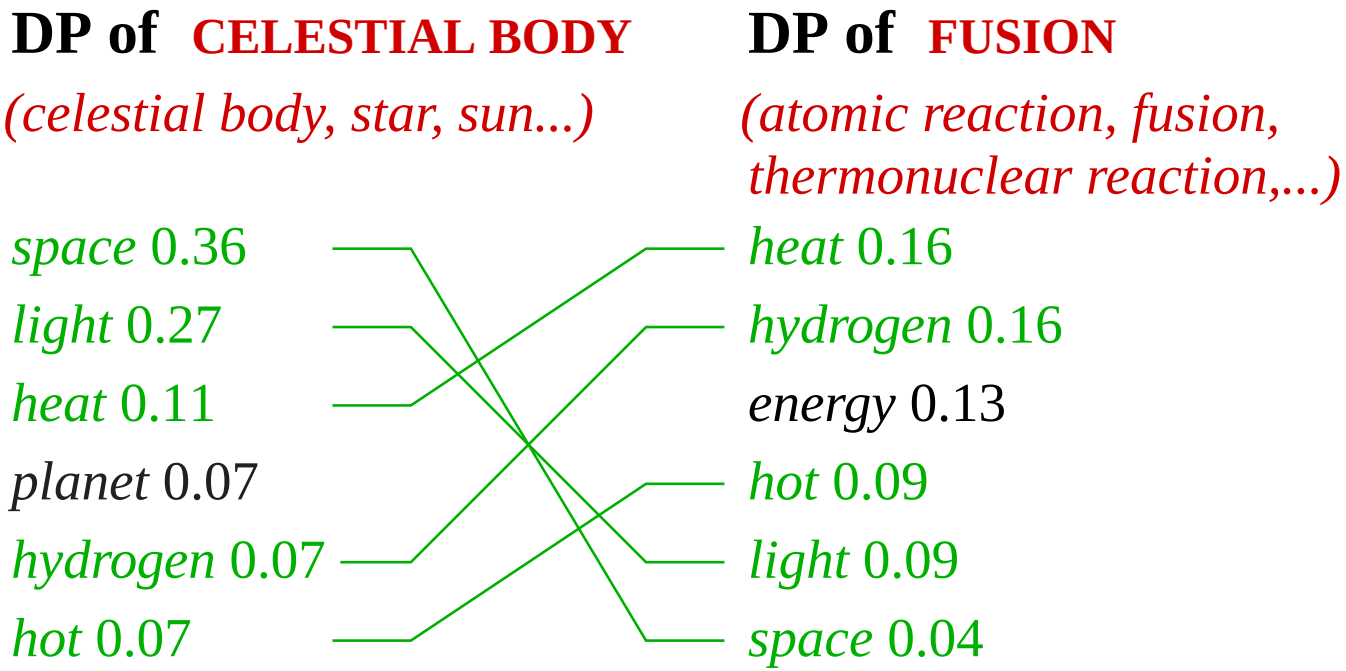
*space* 0.04

# Distance: *star* and *fusion*

| **DP of CELESTIAL BODY** | **DP of FUSION** |
|---|---|
| *(celestial body, star, sun...)* | *(atomic reaction, fusion, thermonuclear reaction,...)* |
| *space* 0.36 | *heat* 0.16 |
| *light* 0.27 | *hydrogen* 0.16 |
| *heat* 0.11 | *energy* 0.13 |
| *planet* 0.07 | *hot* 0.09 |
| *hydrogen* 0.07 | *light* 0.09 |
| *hot* 0.07 | *space* 0.04 |

Then, consider the CELESTIAL BODY sense of *star*.

# Distance: *star* and *fusion*

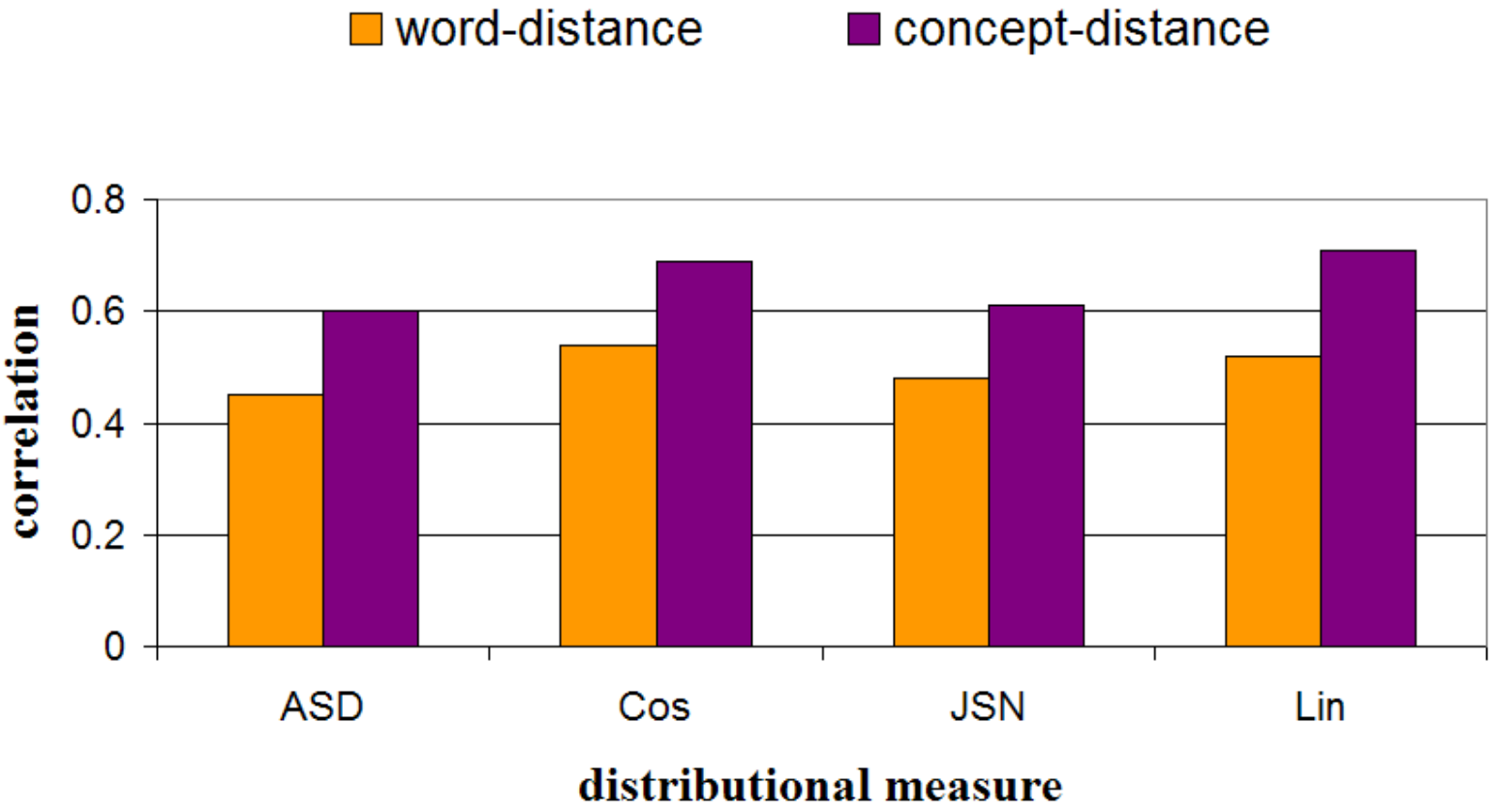



Then, consider the CELESTIAL BODY sense of *star*.

- Distributionally close
- Word sense ambiguity NOT a problem

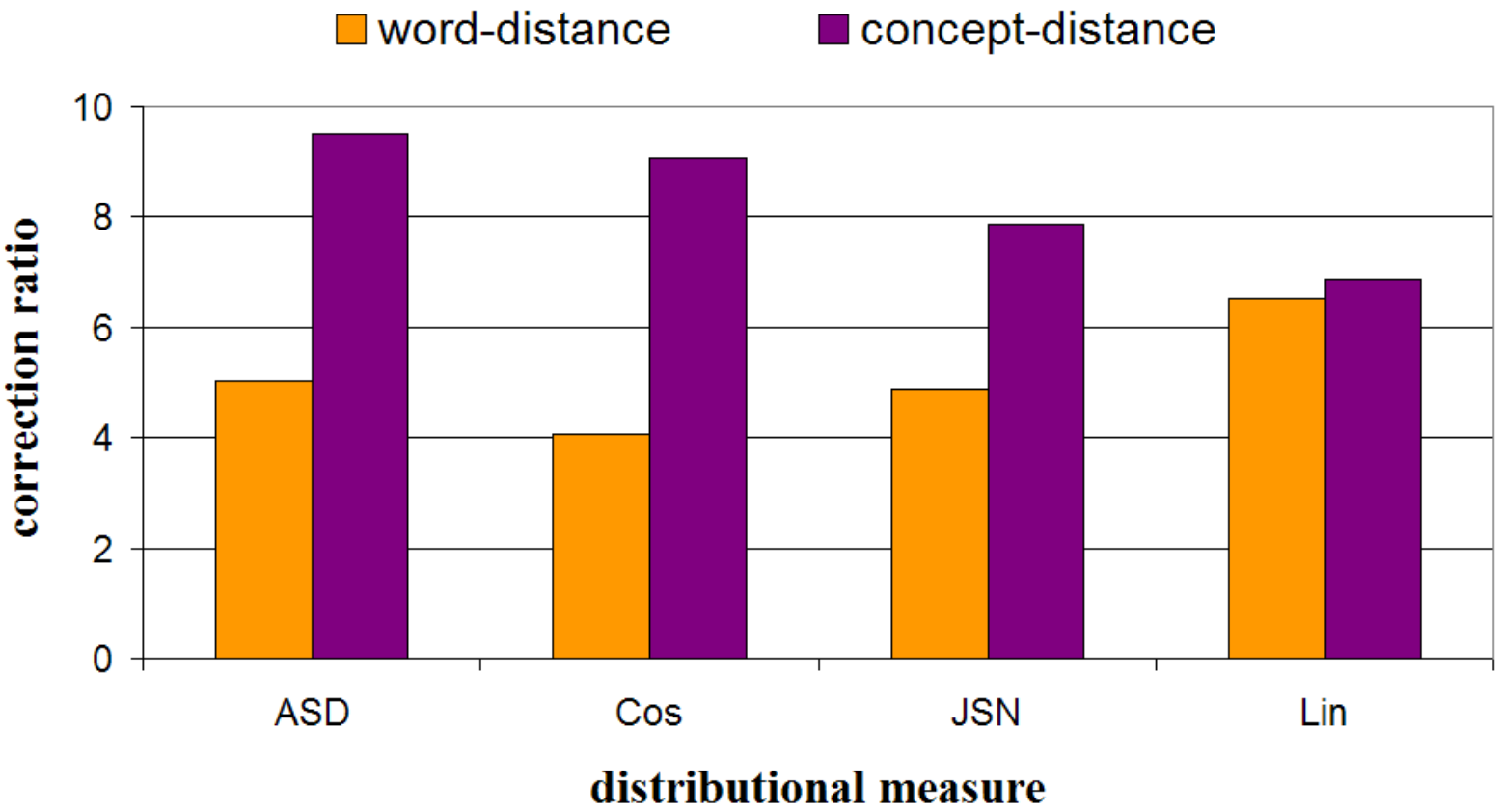
# Ranking word pairs

(Monolingual)



# Correcting spelling errors

(Monolingual)



# But. . .

Application of distance algorithms in most languages is hindered by a lack of high-quality linguistic resources.

# So: Make it cross-lingual

# So: Make it cross-lingual

Torsten Zesch

Darmstadt University of Technology

Iryna Gurevych

Darmstadt University of Technology

# So: Make it cross-lingual

Philip Resnik
University of Maryland

# So: Make it cross-lingual


- Determining distance in a resource-poor language
  - Combine its text with a thesaurus from a (possibly resource-rich) language
    - Largely alleviates the knowledge-source bottleneck
  - Use a bilingual lexicon
  - Without parallel corpora or sense-annotated data

- Experiments: German as a “resource-poor” language

# Cross-lingual links


*Stern* *Bank* $\mapsto w^{de}$

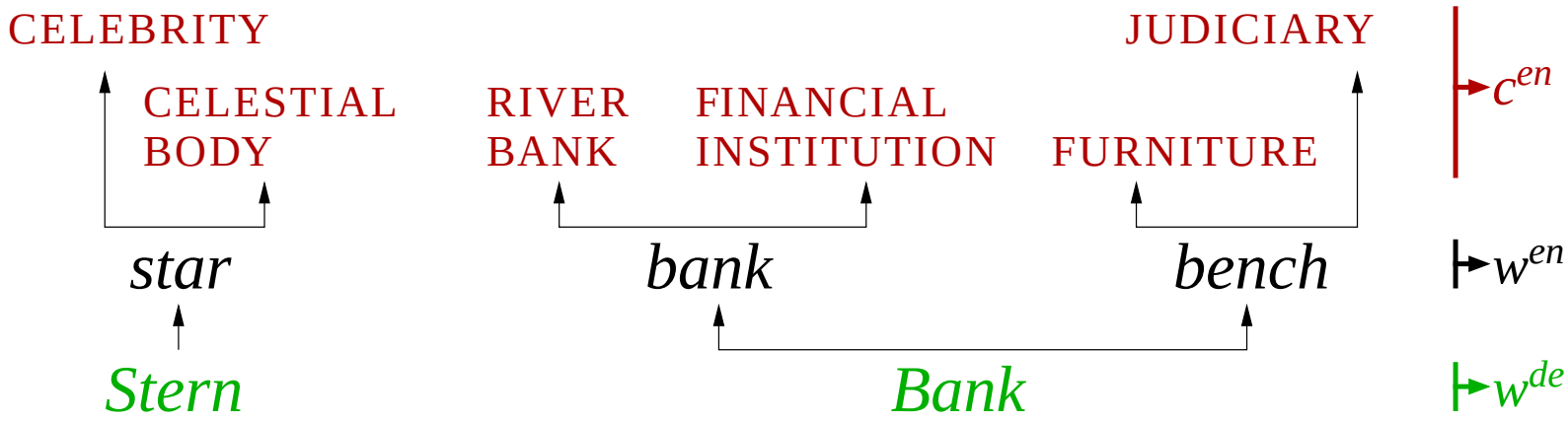
German words $w^{de}$

# Cross-lingual links


German words $w^{de}$

English translations $w^{en}$ (German–English lexicon)

# Cross-lingual links

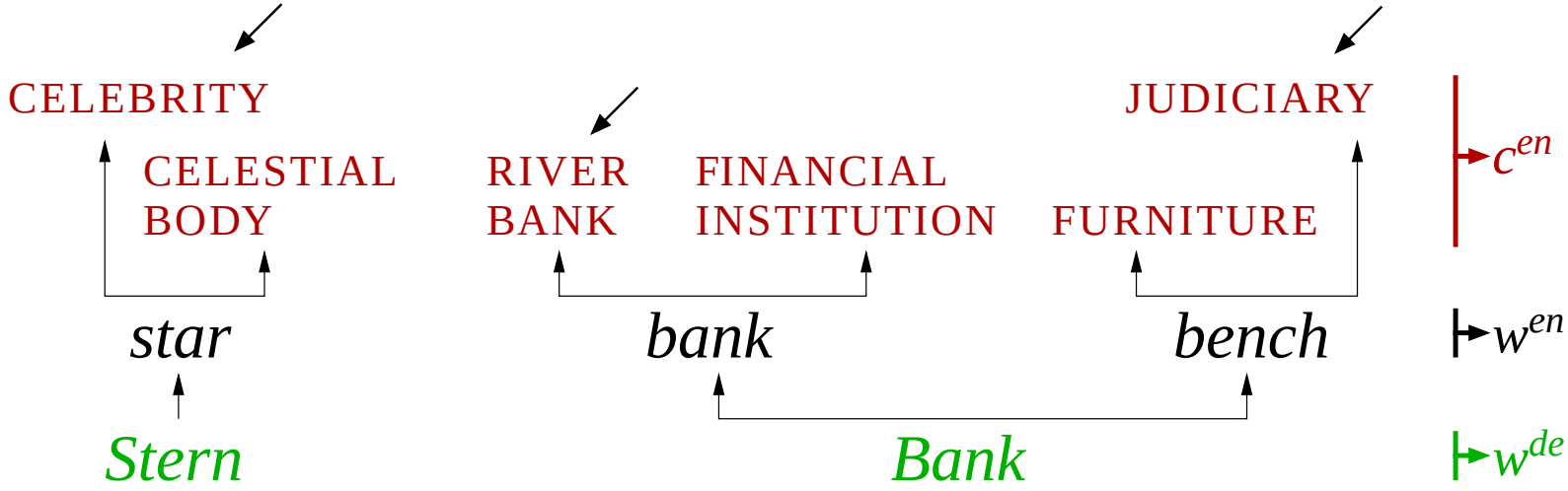



German words $w^{de}$

English translations $w^{en}$ (German–English lexicon)

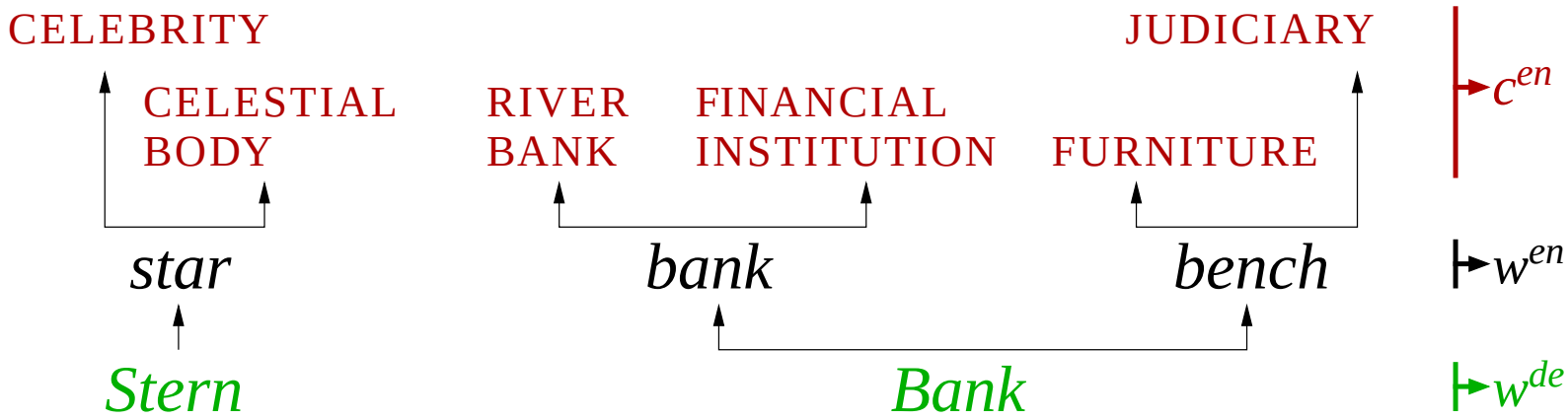
English concepts $c^{en}$ (English thesaurus)

# Dealing with ambiguity

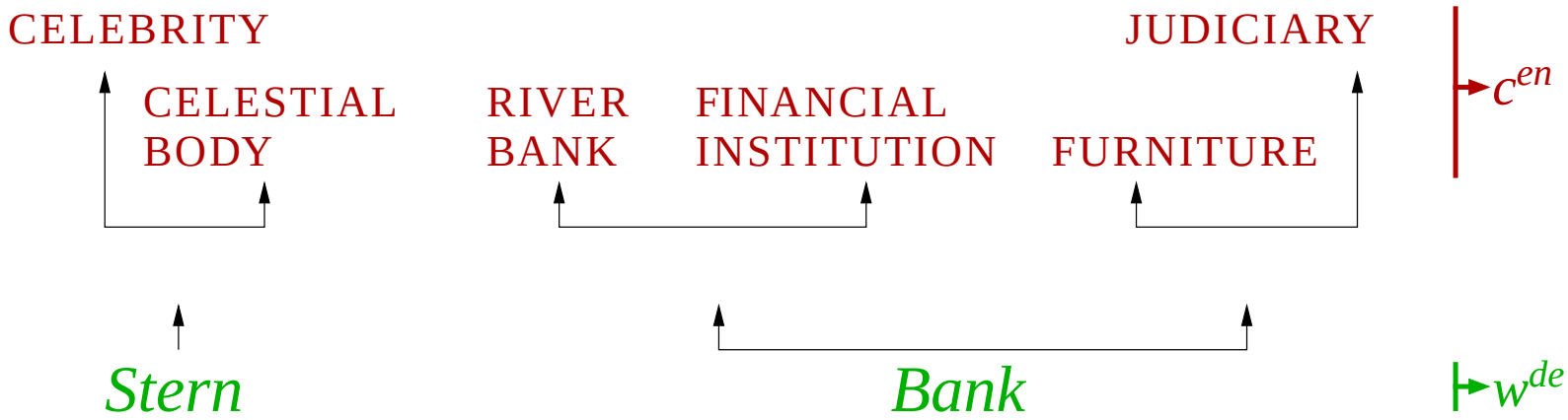



The concepts of CELEBRITY, RIVER BANK and JUDICIARY are semantically unrelated to *Stern* and *Bank*.

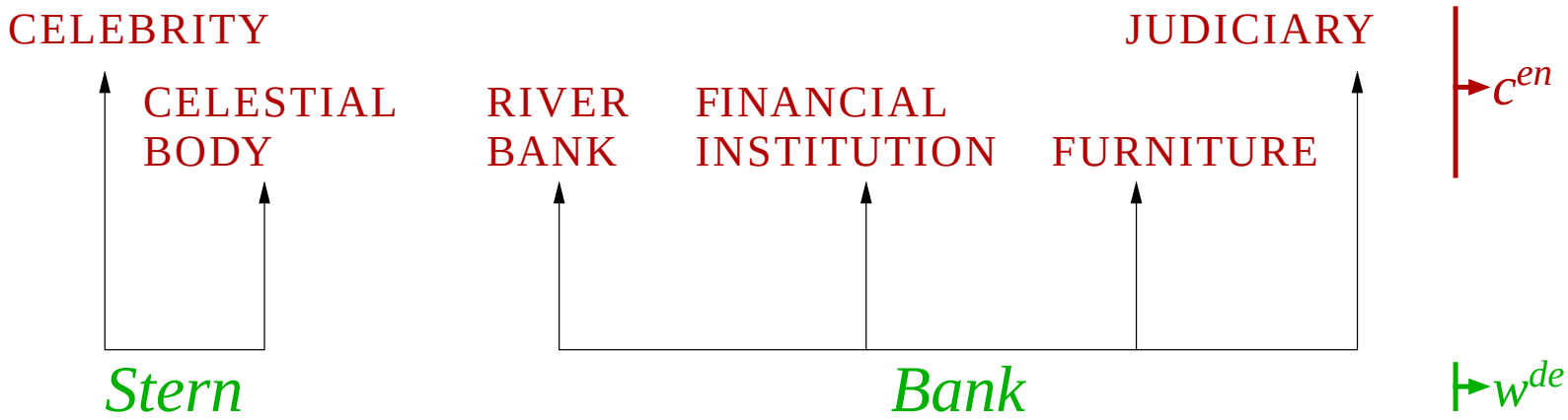
# Losing the English words



# Losing the English words



# Losing the English words




**Cross-lingual candidate senses** of German words *Stern* and *Bank*

# Cross-lingual DPCs

Cross-lingual DPs of the concepts referred to by *star*:

**DP of CELESTIAL BODY**
*(celestial body, star, sun,...)*

**DP of CELEBRITY**
*(celebrity, hero, star,...)*

→ English

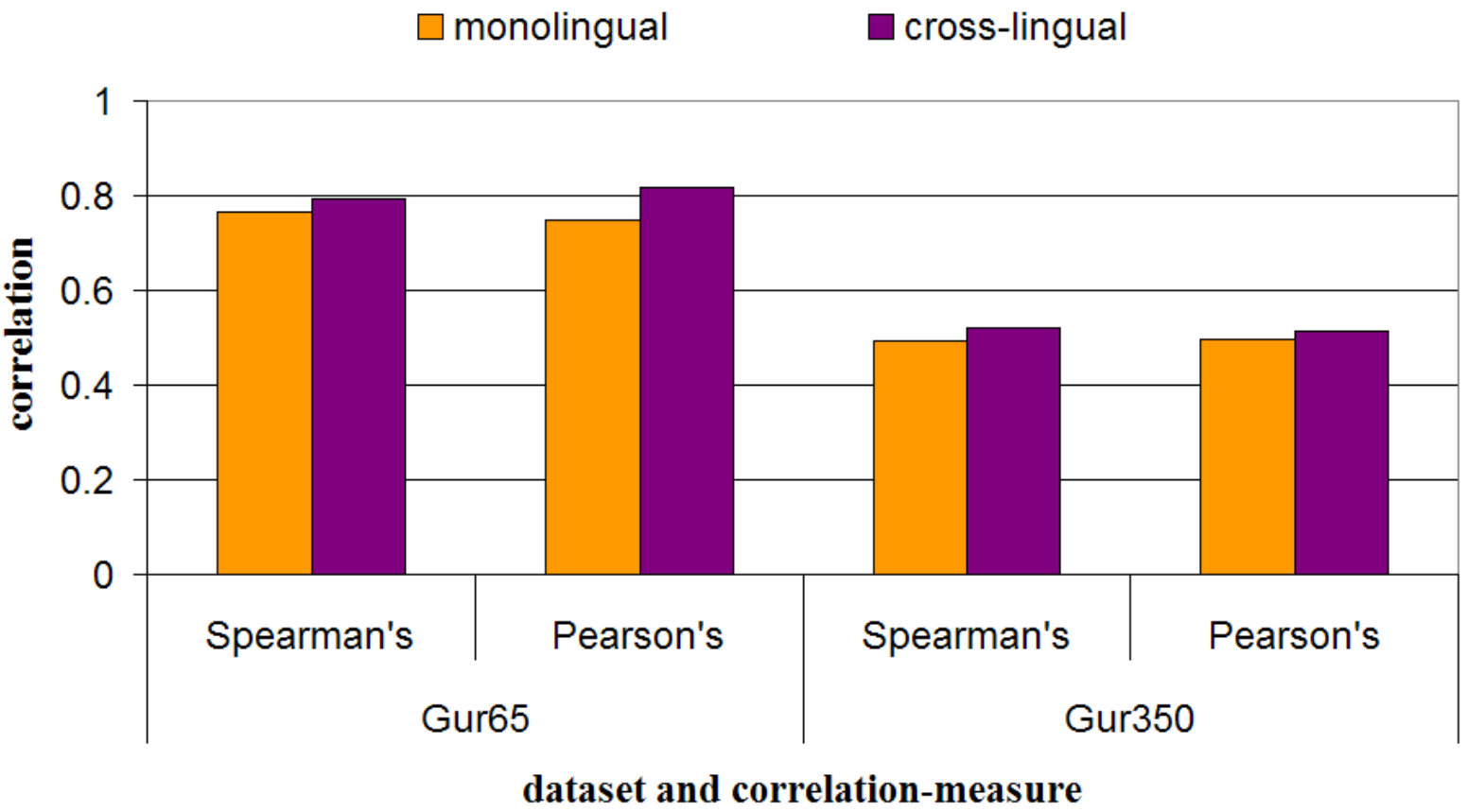
# Cross-lingual DPCs

Cross-lingual DPs of the concepts referred to by *star*:

| **DP of CELESTIAL BODY** | **DP of CELEBRITY** | |
|---|---|---|
| *(celestial body, star, sun,...)* | *(celebrity, hero, star,...)* | English |
| *Raum* 0.36 | *berühmt* 0.24 | German |
| *Licht* 0.27 | *Film* 0.14 | |
| *Hitze* 0.11 | *reich* 0.14 | |
| *Planet* 0.07 | *Fan* 0.10 | |
| *Wasserstoff* 0.06 | *heiß* 0.04 | |
| *heiß* 0.01 | *Mode* 0.01 | |

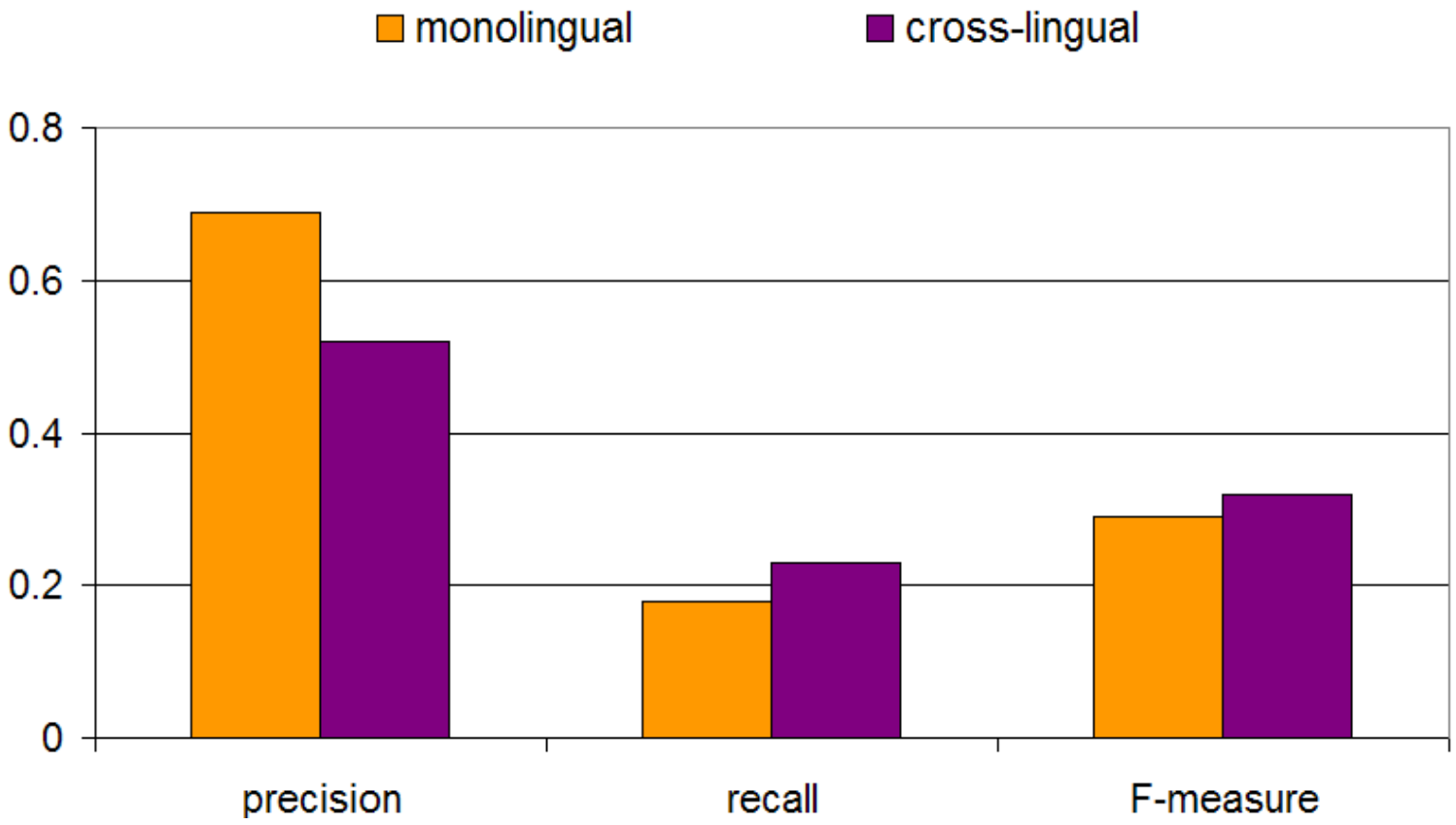
# Ranking word pairs

(Cross-lingual)



# Solving word choice problems

(Cross-lingual)



# Distance between a concept and its context

*word ... word ... target word ... word ... word*

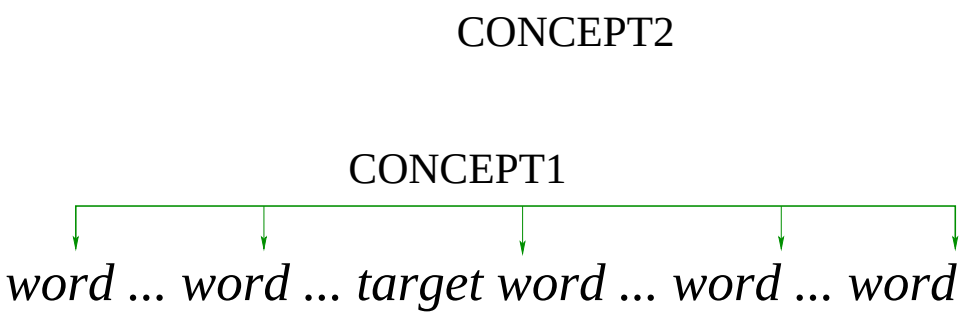
# Distance between a concept and its context

CONCEPT2

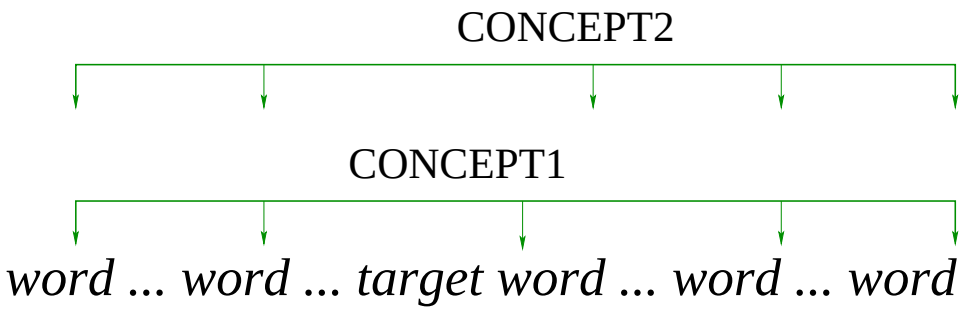
CONCEPT1

*word ... word ... target word ... word ... word*

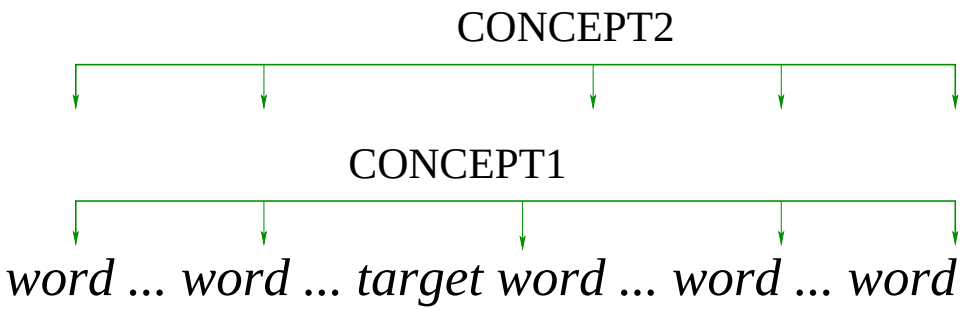
# Distance between a concept and its context



# Distance between a concept and its context



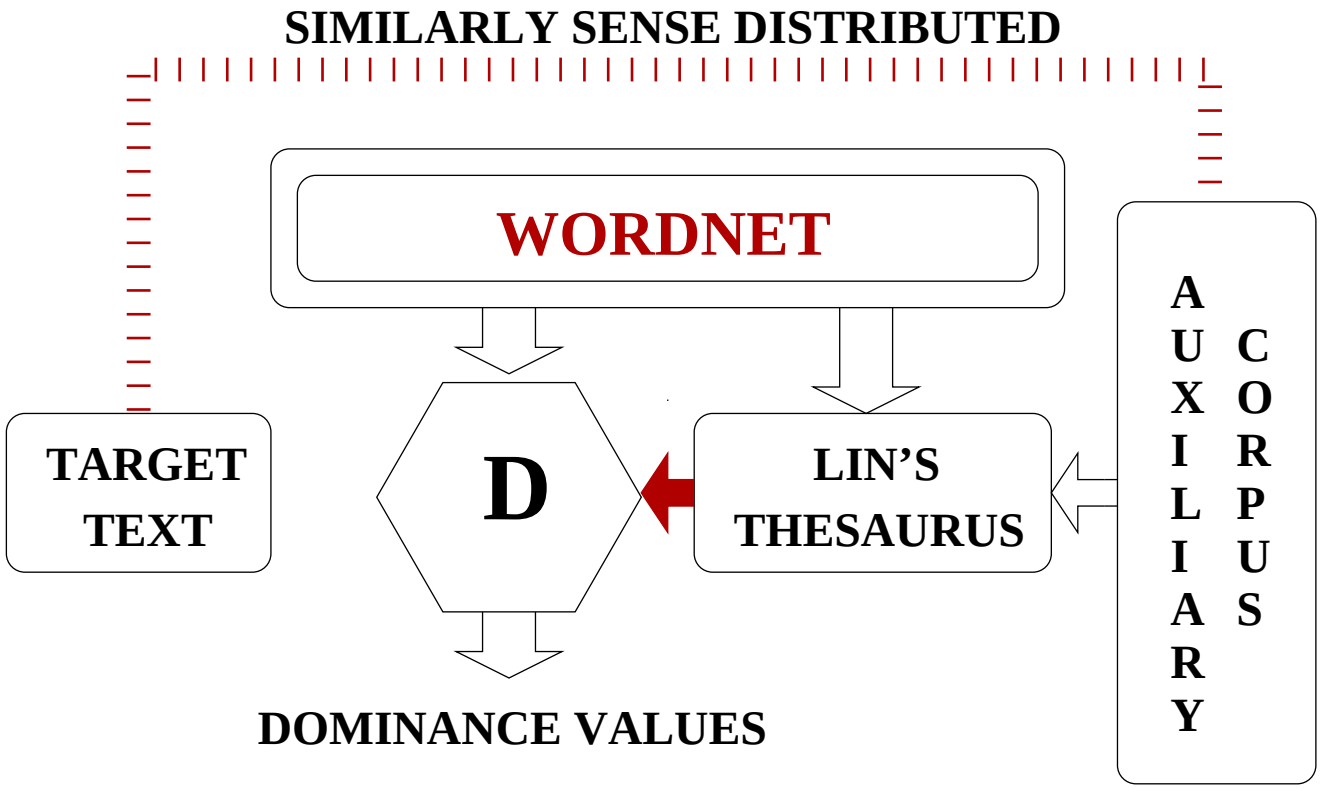
# Distance between a concept and its context




Word sense dominance and word sense disambiguation:

- Obviate the need of sense-annotated data

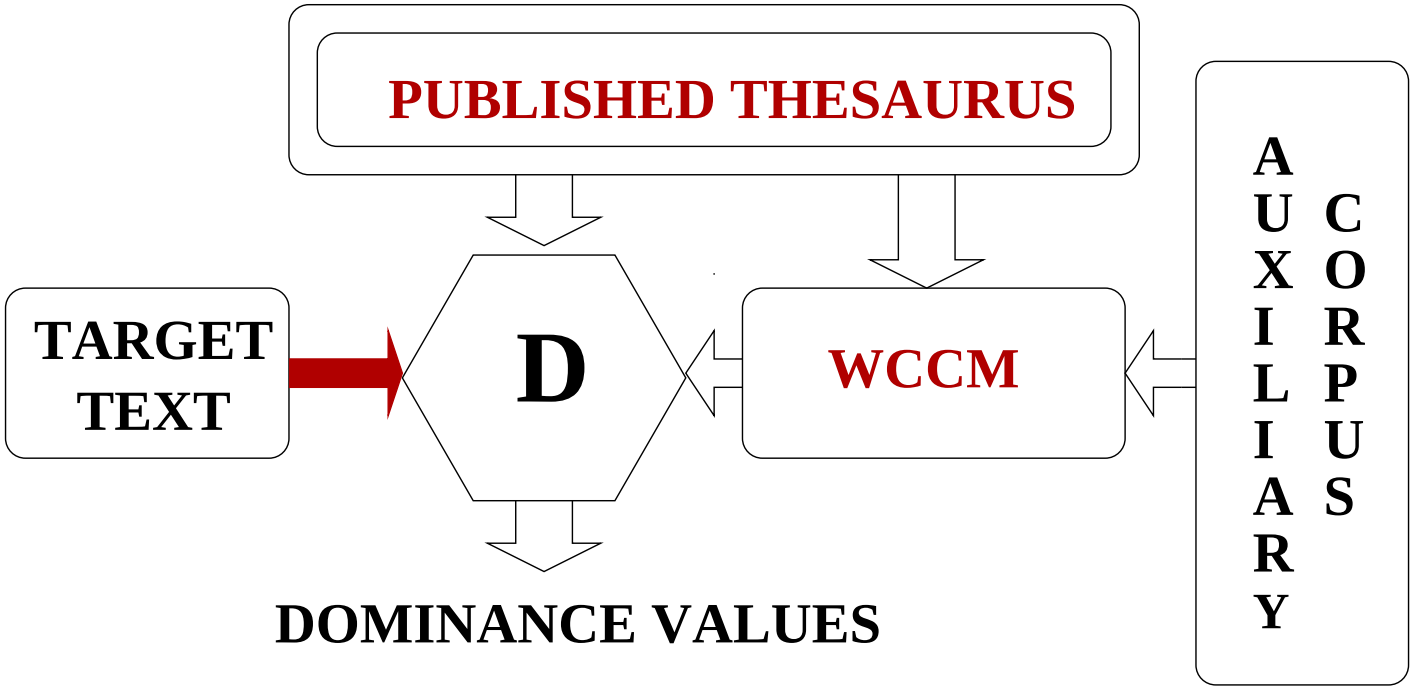
# McCarthy et al.'s Method




D = dominance method

- Requires WordNet and works well only for nouns.
- Needs auxiliary text with similar sense distribution.
- Requires retraining (Lin's thesaurus).

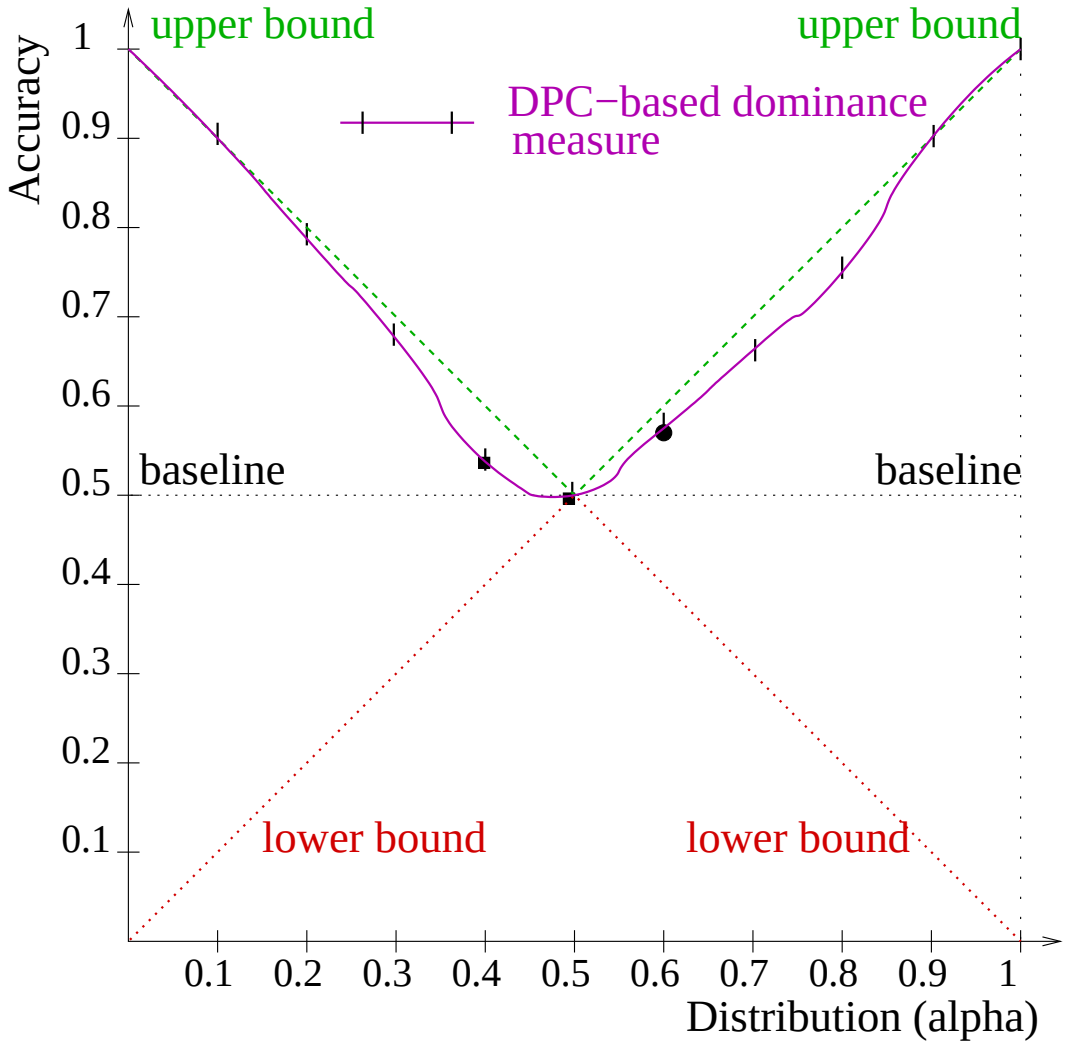
# Our Method




WCCM = word–category co-occurrence matrix

- Uses a published thesaurus and works well for all pos.
- Auxiliary text need not have similar sense distribution.
- No retraining is needed (WCCM created just once).

# Word sense dominance




Mean distance below upper bound = 0.02

# Unsupervised Naïve Bayes word sense classifier

- Estimated probabilities from the DPC
- Took part in SemEval-07's:
  - English Lexical Sample Task
    - Only one percentage point behind the best unsupervised system
  - Multilingual Chinese–English Lexical Sample Task
    - Placed clear first among unsupervised systems

# Accomplishments (1)


- Performed a qualitative and quantitative comparison of WordNet-based and distributional measures

- Identified significant limitations of state-of-the-art approaches to measuring semantic distance
  - Word sense ambiguity
    - A hurdle for distributional measures

# Accomplishments (2)


- Proposed a new hybrid approach to semantic distance
  - Combines text with a thesaurus
  - Models concepts (rather than words)
  - Uses thesaurus categories as very coarse senses

# Accomplishments (3)


- Extensive evaluation
  - Monolingual
    - By combining English text with an English thesaurus
      - Ranked word pairs
      - Corrected real-word spelling errors
      - Determined word sense dominance
      - Did word sense disambiguation

# Accomplishments (4)


- Extensive evaluation (continued)
  - Cross-lingual
    - By combining German text with an English thesaurus
      - Ranked word pairs and solving word-choice problems in German
    - By combining Chinese text with an English thesaurus
      - Identified the English translations of Chinese words from their contexts

# Future work


- Adding cross-lingual semantic distance as a feature to a state-of-the-art MT system (with Philip Resnik)
- Cross-lingual document clustering
- Cross-lingual information retrieval
- Cross-lingual summarization (with Bonnie Dorr)
- Determining paraphrases, lexical entailment, and contradictions (with Bonnie Dorr)
- Determining cognates using semantic distance between words in different languages (with Greg Kondrak)
- Porting the approach to Wikipedia (with Torsten Zesch and Iryna Gurevych)

# Conclusions (1)


- Distributional profiles of concepts can be used to infer their semantic properties, and indeed estimate semantic distance.
- Cross-lingual DPCs allow for a seamless transition from words in one language to concepts in another.

# Conclusions (2)


- Distributional measures of concept-distance are markedly superior to previous approaches.

# Conclusions (2)


- Distributional measures of concept-distance are markedly superior to previous approaches.
  - Works well for all pos pairs

# Conclusions (2)


- Distributional measures of concept-distance are markedly superior to previous approaches.
  - Works well for all pos pairs
  - Gives both relatedness and similarity

# Conclusions (2)


- Distributional measures of concept-distance are markedly superior to previous approaches.
  - Works well for all pos pairs
  - Gives both relatedness and similarity
  - Domain adaptable

# Conclusions (2)


- Distributional measures of concept-distance are markedly superior to previous approaches.
  - Works well for all pos pairs
  - Gives both relatedness and similarity
  - Domain adaptable
  - Can be used in real-time systems

# Conclusions (2)


- Distributional measures of concept-distance are markedly superior to previous approaches.
  - Works well for all pos pairs
  - Gives both relatedness and similarity
  - Domain adaptable
  - Can be used in real-time systems
  - Cross-lingual
    - Solve problems in a one language using a knowledge source from another
    - Solve problems that involve multiple languages

# A computationl model of Antonymy

Antonyms simultaneously convey a sense of both distance and closeness (Cruse, 86).


Bonnie Dorr

University of Maryland