Cross-lingual Distributional Profiles of Concepts for Measuring Semantic Distance

Saif Mohammad, Iryna Gurevych,
Graeme Hirst, and Torsten Zesch

University of Toronto & Darmstadt University of Technology

Semantic distance

SALSA DANCE

CLOWN BRIDGE

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

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

Example DPs of words

DP of *star*

star : *space* 0.21, *movie* 0.16, *famous* 0.15, *light* 0.12, *constellation* 0.11, *heat* 0.08, *rich* 0.07, *hydrogen* 0.07, ...

DP of *fusion*

fusion : *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, ...

Example DPs of words

DP of *star*

star: *space* 0.21, *movie* 0.16, *famous* 0.15, *light* 0.12, *constellation* 0.11, *heat* 0.08, *rich* 0.07, *hydrogen* 0.07, . . .

DP of *fusion*

fusion: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, . . .

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 (text)
  - Distributional measure: distance between DPs

    Cosine, Lin, α-skew divergence
- Drawbacks
  - Poor accuracy (albeit higher coverage)
  - Conflation of word senses

Problem with distributional word-distance measures

DP of *star*

star: *space* 0.21, *movie* 0.16, *famous* 0.15, *light* 0.12, *constellation* 0.11, *heat* 0.08, *rich* 0.07, *hydrogen* 0.07, ...

DP of *fusion*

fusion: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, ...

Problem with distributional word-distance measures

DP of *star*

star: *space* 0.21, *movie* 0.16, *famous* 0.15, *light* 0.12, *constellation* 0.11, *heat* 0.08, *rich* 0.07, *hydrogen* 0.07, . . .

DP of *fusion*

fusion: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, . . .

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

Our hybrid approach

(Mohammad and Hirst, EMNLP-2006)

- Combines a knowledge source with text
- Profiles concepts (rather than words)
- Uses thesaurus categories as concepts/coarse-grained senses
  - Most published thesauri: around 1000 categories
  - Concept–concept distance matrix: only 1000×1000
- Capable of giving both similarity and relatedness values

Distributional profiles of concepts

DPs of the concepts referred to by *star*:

DP of 'celestial body'

'celestial body' (*celestial body*, *sun*, ...): *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, ...

DP of 'celebrity'

'celebrity' (*celebrity*, *hero*, ...): *famous* 0.24, *movie* 0.14, *rich* 0.14, *fan* 0.10, ...

Distance: *star* and *fusion*

First, consider the ‘celebrity’ sense of *star*:

DP of ‘celebrity’

‘celebrity’*star*: *famous* 0.24, *movie* 0.14, *rich* 0.14, *fan* 0.10, . . .

DP of ‘fusion’

‘fusion’: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, . . .

Distributionally NOT close

Distance: *star* and *fusion*

Then, consider the ‘celestial body’ sense of *star*:

DP of ‘celestial body’

‘celestial body’: *space* 0.21, *light* 0.12, *constellation* 0.11, *heat* 0.08, *hydrogen* 0.07, . . .

DP of ‘fusion’

‘fusion’: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, . . .

Distributionally close

Word sense ambiguity NOT a problem

Our previous results
(Mohammad and Hirst, EMNLP-2006)

- Concept-distance better than word-distance
- Combining text and a knowledge source gives higher accuracies

But. . .

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

So: Make it cross-lingual

- A new way of determining distance in a resource-poor language
  - By combining its text with a thesaurus from a (possibly resource-rich) language
    - Largely eliminates the knowledge-source bottleneck
  - Using a bilingual lexicon and a bootstrapping algorithm
- Without relying on parallel corpora or sense-annotated data
- Experiments: German as a “resource-poor” language

Distance: German concepts

German text
(*taz*)

English thesaurus
(*Macquarie*)

bilingual lexicon
(*BEOLINGUS*)

bootstrapping algorithm

English–German distributional profiles of concepts

Cross-lingual links

Stern *Bank* } w^{de}

German words w^{de}

Cross-lingual links

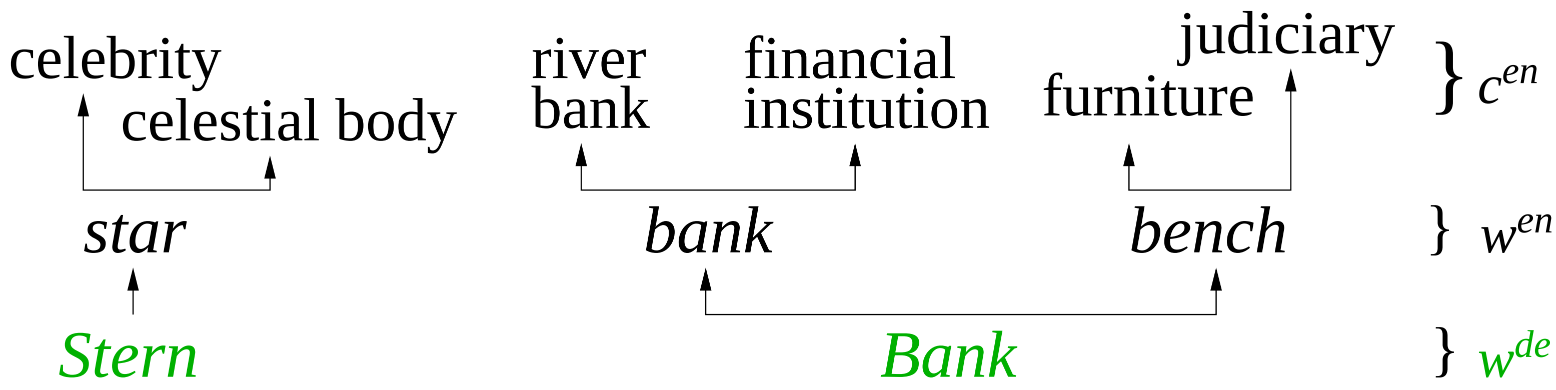
star *bank* *bench* } w^{en}

Stern *Bank* } w^{de}

German words w^{de}

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

Cross-lingual links

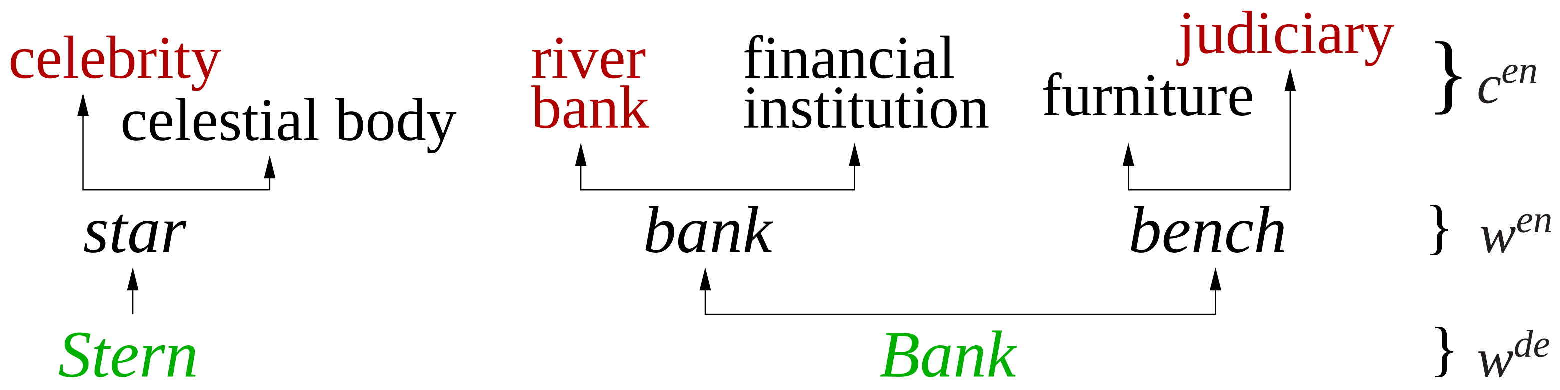


German words w^{de}

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

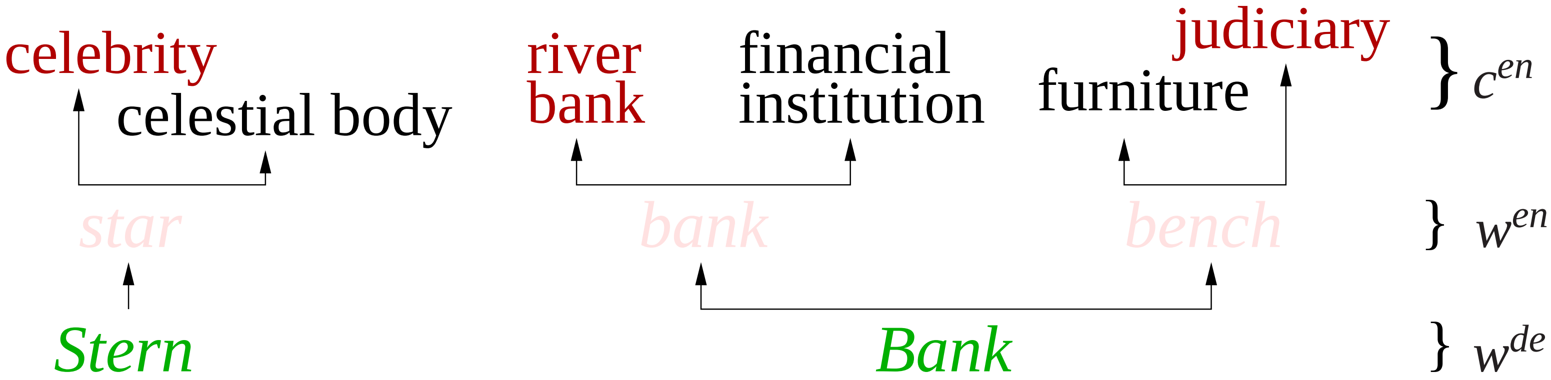
English concepts c^{en} (English thesaurus)

Dealing with ambiguity

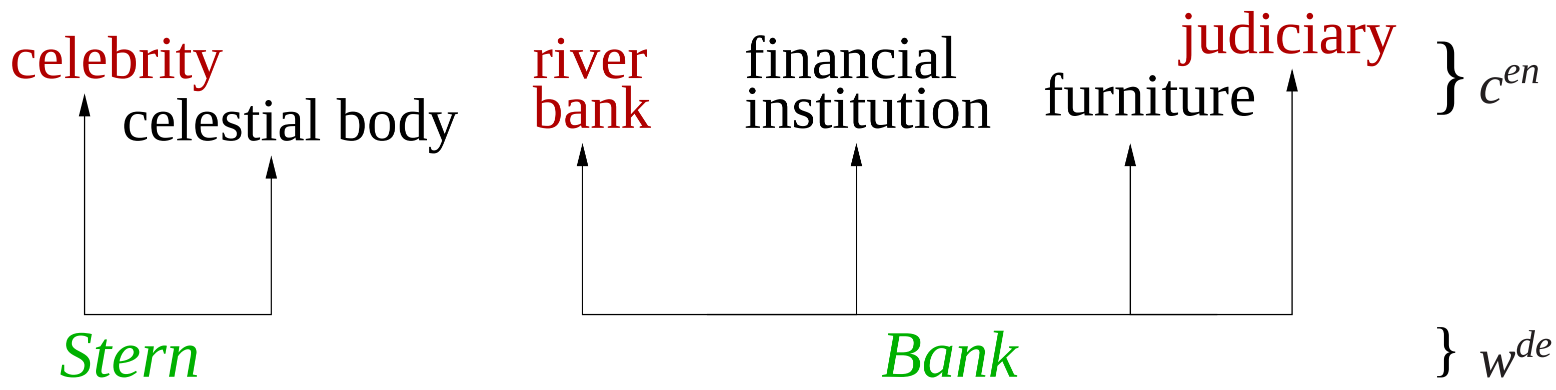


The concepts of 'celebrity' and 'judiciary' are semantically unrelated to *Stern* and *Bank*, respectively.

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*:

Cross-lingual DP of 'celestial body'

'celestial body' (*celestial body*, *sun*, ...): *Raum* 0.36, *Licht* 0.27, *Konstellation* 0.11, ...

Cross-lingual DP of 'celebrity'

'celebrity' (*celebrity*, *hero*, ...): *berühmt* 0.24, *Film* 0.14, *reich* 0.14, ...

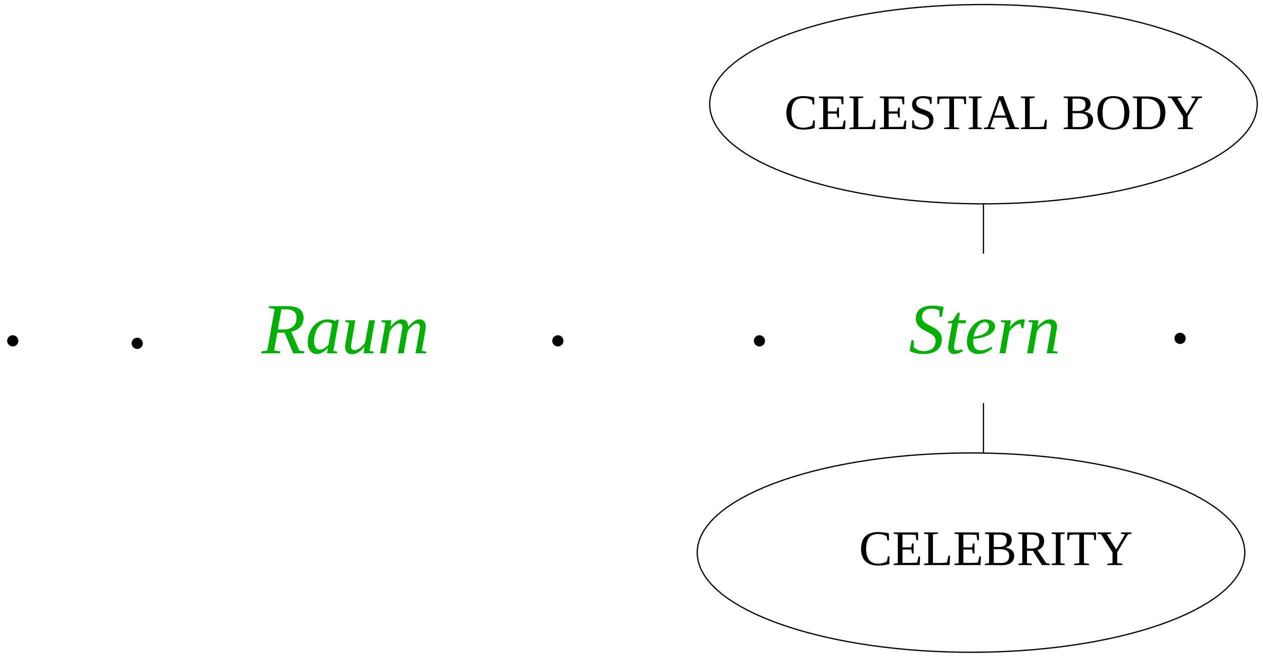
Creating cross-lingual DPCs

Cross-lingual word–category co-occurrence matrix (WCCM)

$$
\begin{array}{c|cccccc}
 & c_1^{en} & c_2^{en} & \ldots & c_j^{en} & \ldots \\
\hline
w_1^{de} & m_{11} & m_{12} & \ldots & m_{1j} & \ldots \\
w_2^{de} & m_{21} & m_{22} & \ldots & m_{2j} & \ldots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
w_i^{de} & m_{i1} & m_{i2} & \ldots & m_{ij} & \ldots \\
\vdots & \vdots & \vdots & \ldots & \vdots & \ddots
\end{array}
$$

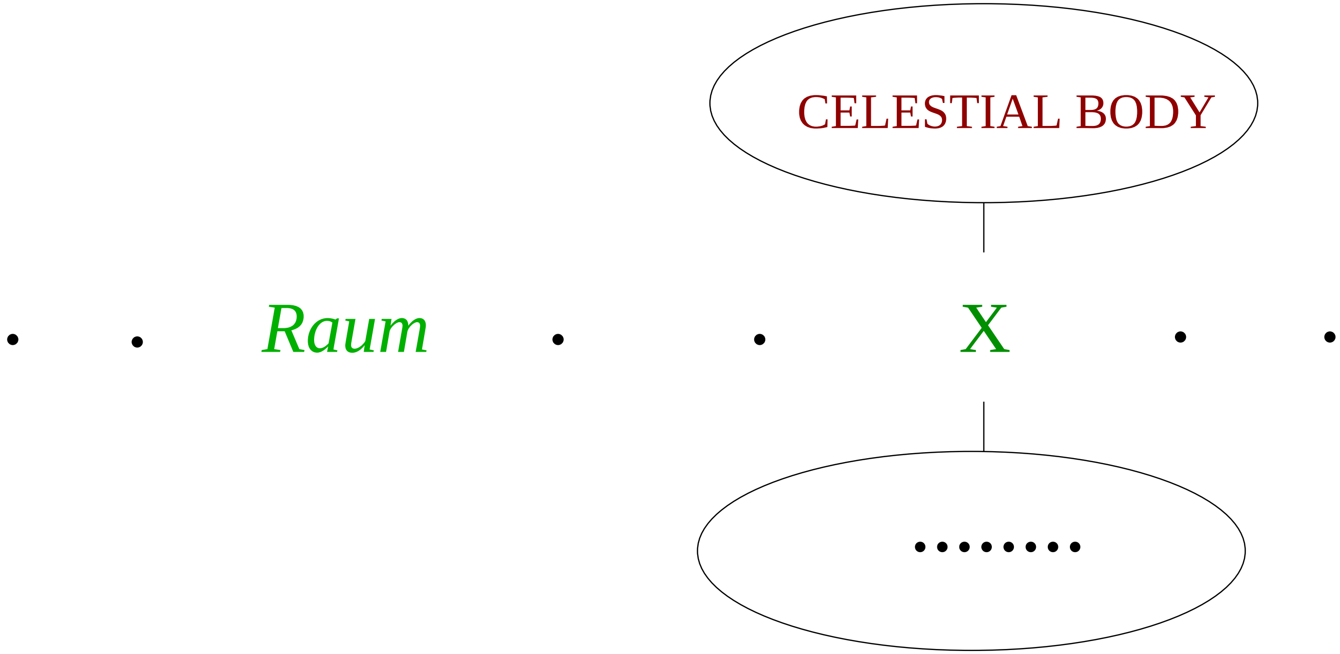
- WCCM: German words vs. English categories
- Cell m_{ij}: number of times word w_i co-occurs with a word having c_j as one of its cross-lingual candidate senses

First pass



- Cell (*Raum*, CELESTIAL BODY) incremented
- Cell (*Raum*, CELEBRITY) incremented

First pass (continued)

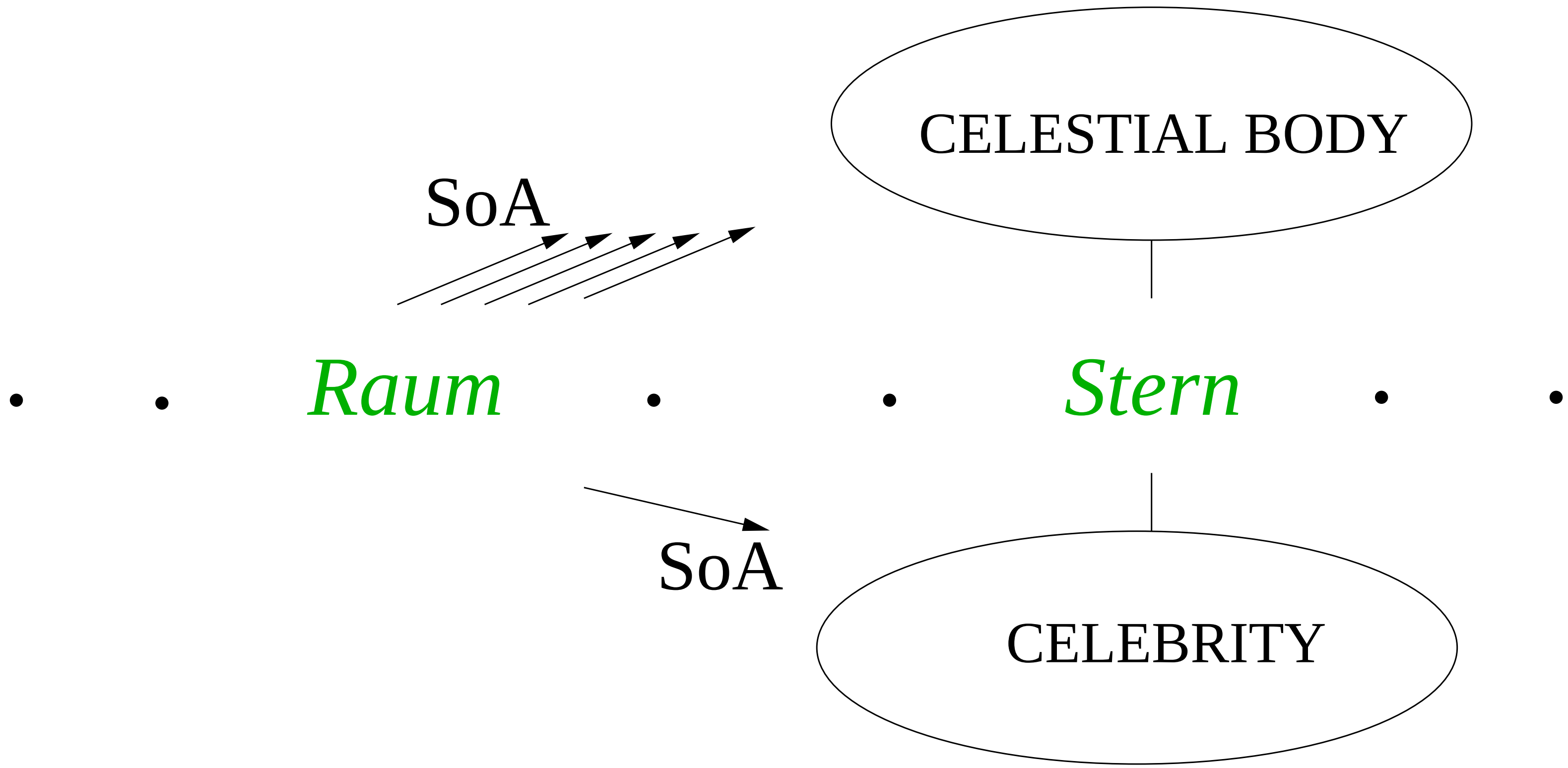


X: *Stern, Sonne, Himmelskörper, Morgensonne, Konstellation*

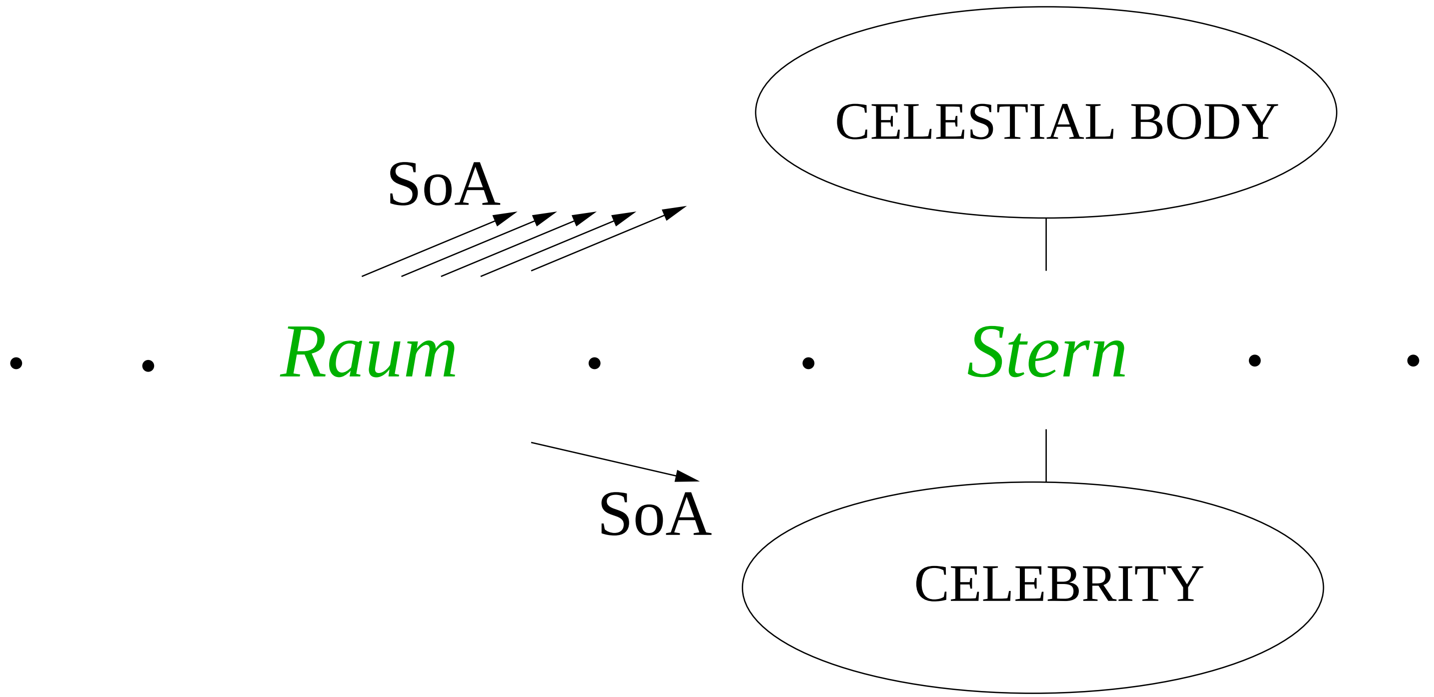
Cross-lingual matrix

$$
\begin{array}{c|ccccc}
 & & & & \text{CELESTIAL} & \\
 & c_1^{en} & c_2^{en} & \dots & \text{BODY} & \dots \\
\hline
w_1^{de} & m_{11} & m_{12} & \dots & m_{1j} & \dots \\
w_2^{de} & m_{21} & m_{22} & \dots & m_{2j} & \dots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
Raum & m_{i1} & m_{i2} & \dots & m_{ij} & \dots \\
\vdots & \vdots & \vdots & \dots & \vdots & \ddots
\end{array}
$$

Evidence for the senses



Second pass



- Cell (*Raum*, CELESTIAL BODY) incremented
- New, more accurate, **bootstrapped WCCM**
  - Word sense dominance (Mohammad and Hirst, EACL-2006)

Cross-lingual DPCs

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

Cross-lingual DP of 'celestial body'

'celestial body' (*celestial body*, *sun*, . . .): *Raum* 0.36, *Licht* 0.27, *Konstellation* 0.11, . . .

Cross-lingual DP of 'celebrity'

'celebrity' (*celebrity*, *hero*, . . .): *berühmt* 0.24, *Film* 0.14, *reich* 0.14, . . .

Measures we used

Cross-lingual and hybrid

- Distributional measures
  - α-skew divergence
  - Cosine
  - Jensen-Shannon divergence
  - Lin's distributional measure

Comparison measures

Monolingual and GermaNet-based

- Lesk-like measures (Gurevych, 2005):
  - Hypernym pseudo-gloss
  - Radial pseudo-gloss
- Information content measures (Budanitsky and Hirst, 2006):
  - Jiang and Conrath's WordNet measure
  - Lin's WordNet measure
  - Resnik's WordNet measure

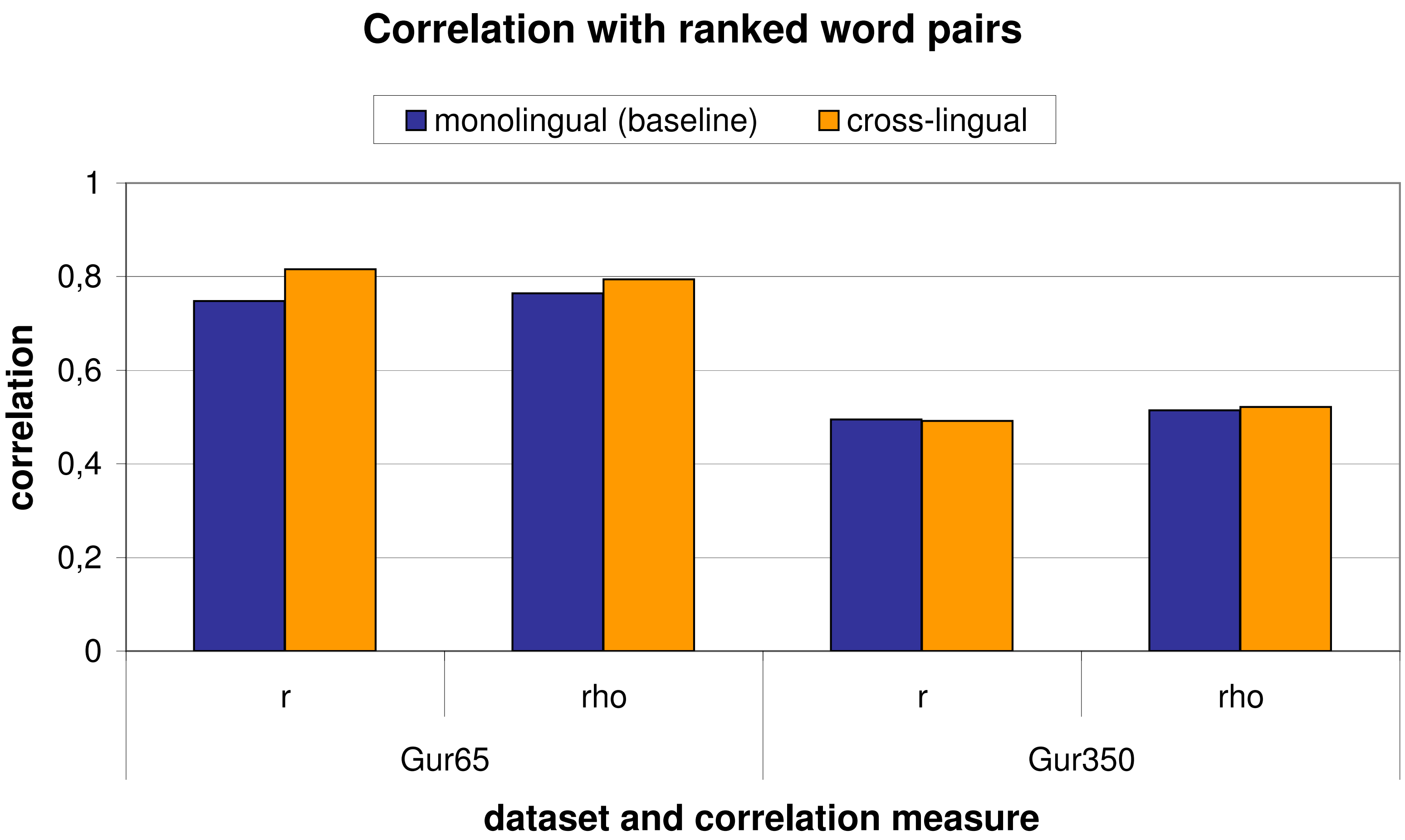
Evaluation

1. Rank closeness of word pairs

Dataset	# pairs	PoS	Relations	Scores	# subjects	Correlation
Gur65	65	N	classical	{0,1,2,3,4}	24	.810
Gur350	350	N, V, A	both	{0,1,2,3,4}	8	.690

- Automatic measures rank word pairs
  - From near-synonyms to unrelated
- Correlation with human ranking
  - Spearman's rank order correlation (ρ)
  - Pearson's correlation coefficient (r)

Evaluation



Correlation with ranked word pairs
monolingual (baseline)
cross-lingual
1
0,8
0,6
0,4
0,2
0
correlation
r
rho
r
rho
Gur65
Gur350
dataset and correlation measure

Evaluation

2. Solve word choice problems

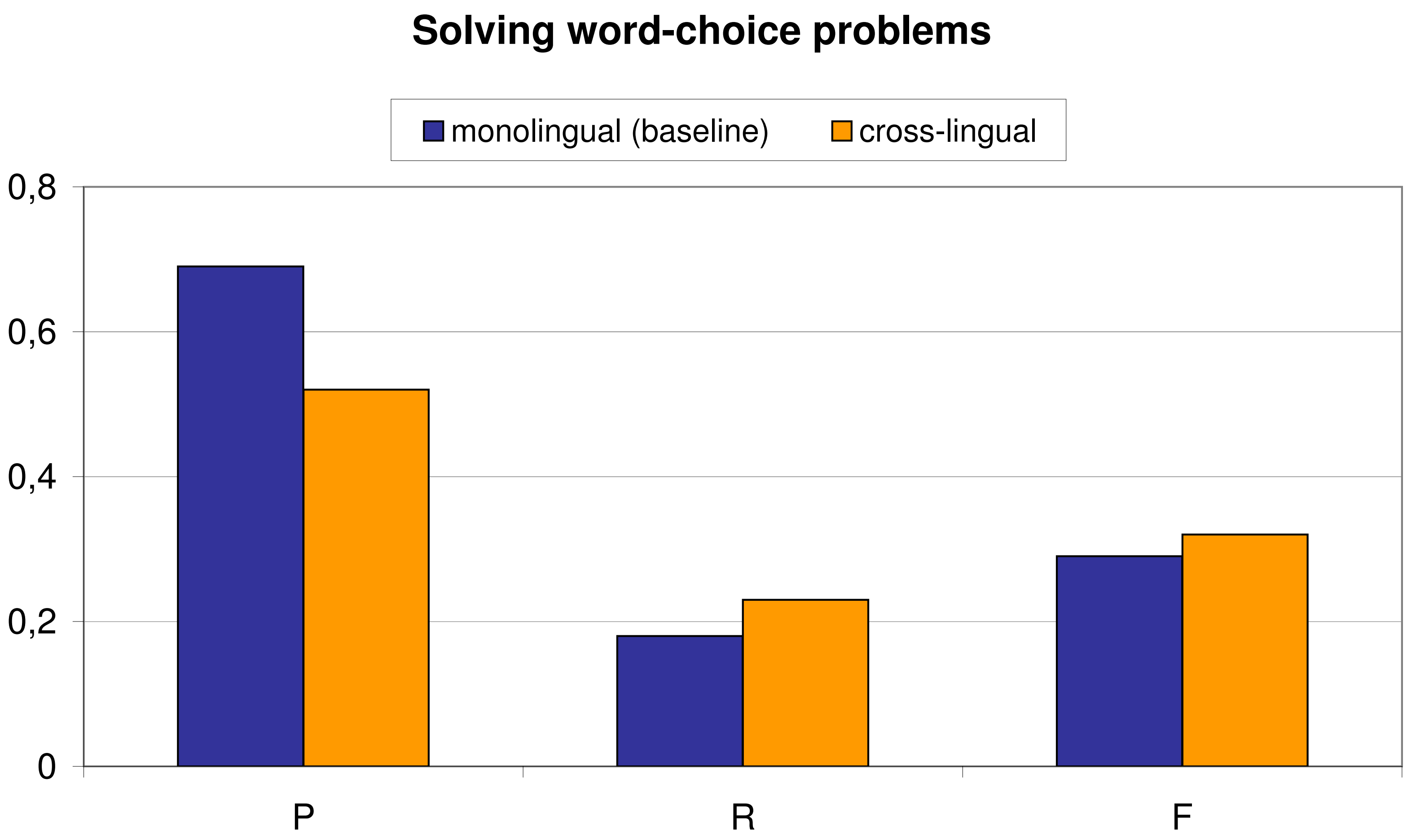
1008 *Reader's Digest* questions:

Duplikat (duplicate)

a. *Einzelstück* (single copy) b. *Doppelkinn* (double chin)

c. *Nachbildung* (replica) d. *Zweitschrift* (copy)

Evaluation


Solving word-choice problems
monolingual (baseline)
cross-lingual
0,8
0,6
0,4
0,2
0
P
R
F

Unsupervised Naïve Bayes word sense classifier

- Estimated probabilities from the cross-lingual DPCs
- Took part in SemEval-07's:
  - Multilingual Chinese–English Lexical Sample Task
    - Placed clear first among unsupervised systems

Summary

- Algorithm to determine semantic distance in resource-poor languages
  - Combine its text with a thesaurus in another language
    - Bilingual lexicon and a bootstrapping algorithm
    - NO sense-annotated data or parallel corpora

- Evaluated on word pair ranking and word choice problems
  - Compared with best monolingual approaches

Conclusions

- State-of-the-art accuracies can be achieved even for languages poor in linguistic resources.
  - Improvement even over established resources
  - Superior coverage (despite the bilingual lexicon step)

- Cross-lingual DPCs allow for a seamless and largely loss-free transition from words in one language to a concepts in another.
  - Machine translation, multi-lingual document clustering, multilingual information retrieval,. . .

Future work

- Using Wikipedia instead of a published thesaurus
- Adding cross-lingual semantic distance as a feature to an MT system
- Determining cognates using semantic distance between words in different languages
- Cross-lingual document clustering
- Cross-lingual information retrieval
- Cross-lingual document summarization

Capturing DPCs

- Method
  - Direct: sense-annotated data
  - Alternative: Mohammad and Hirst (EACL-2006)
    - Combining raw text and a knowledge source

- Sense inventory
  - Published thesaurus

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 are closely related.
  - A category can be thought of as a very coarse-grained concept (Yarowsky, 1992).
    - Represents senses of the words in it

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

Precomputing Distances

Distributional word–word distance matrix

$\approx 100{,}000 \times 100{,}000$

$$
\begin{array}{c|cccc}
 & w_1 & \ldots & w_j & \ldots \\
\hline
w_1 & m_{11} & \ldots & m_{1j} & \ldots \\
\vdots & \vdots & \ddots & \vdots & \ldots \\
w_i & m_{i1} & \ldots & m_{ij} & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{array}
$$

WordNet-based concept-concept distance matrix

$\approx 75{,}000 \times 75{,}000$

$$
\begin{array}{c|cccc}
 & c_1 & \ldots & c_j & \ldots \\
\hline
c_1 & m_{11} & \ldots & m_{1j} & \ldots \\
\vdots & \vdots & \ddots & \vdots & \ldots \\
c_i & m_{i1} & \ldots & m_{ij} & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{array}
$$

Why a Thesaurus?

- Computational ease: concept–concept distance matrix is much smaller (roughly .01%).

- Coarse senses: WordNet is much too fine grained.

- Availability: Thesauri are available in many languages.

- Words for a sense: Each sense can be represented unambiguously with a set of (possibly ambiguous) words.

Concept-Distance Approach

$$
\begin{tikzcd}
\text{CELEBRITY} \arrow[rr, dotted, no head] \arrow[ddrr, dotted, no head] & & \text{MOTION PICTURE} \\
\textit{star} \arrow[u] \arrow[d] & & \textit{film} \arrow[u] \arrow[d] \\
\text{CELESTIAL BODY} \arrow[rr, dotted, no head] \arrow[uurr, dotted, no head] & & \text{THIN MEMBRANE}
\end{tikzcd}
$$

$$
\begin{aligned}
\mathit{distance}(\mathit{star}, \mathit{film}) = \\
& \min \Big(\mathit{distance}(\text{CELEBRITY, MOTION PICTURE}), \\
& \quad \mathit{distance}(\text{CELEBRITY, THIN MEMBRANE}), \\
& \quad \mathit{distance}(\text{CELESTIAL BODY, MOTION PICTURE}), \\
& \quad \mathit{distance}(\text{CELESTIAL BODY, THIN MEMBRANE}) \Big)
\end{aligned}
$$