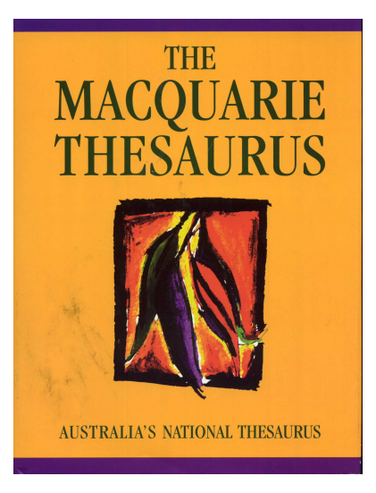
TOR, TORMD: Distributional Profiles of Concepts for Unsupervised Word Sense Disambiguation

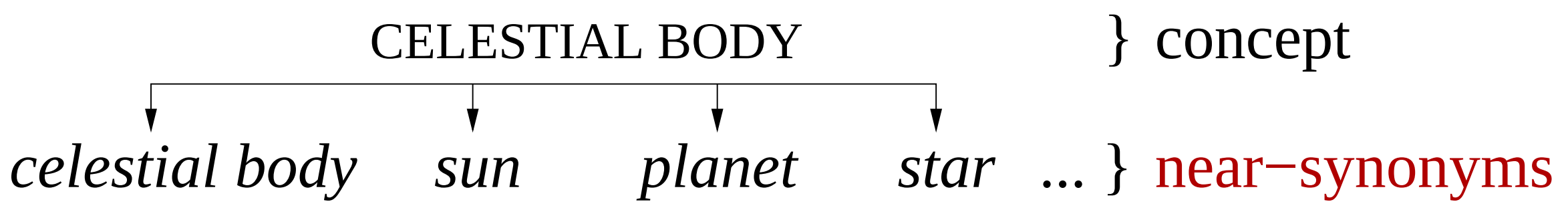
Saif Mohammad, Graeme Hirst, and Philip Resnik
University of Toronto and University of Maryland
{smm,gh}@cs.toronto.edu, resnik@umiacs.umd.edu

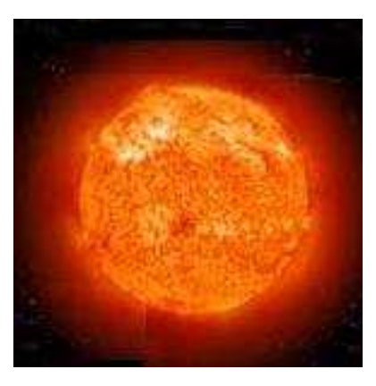
How to represent a concept?



As a category from a thesaurus

A set of near−synonymous words:

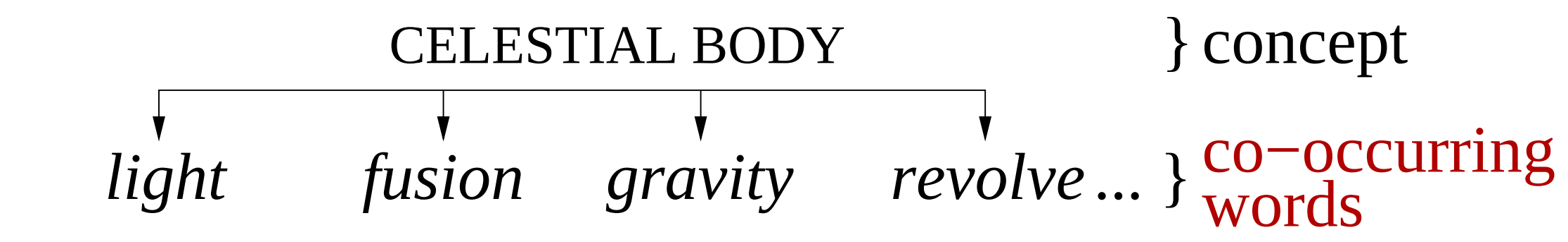




CELESTIAL BODY

By its usage in text

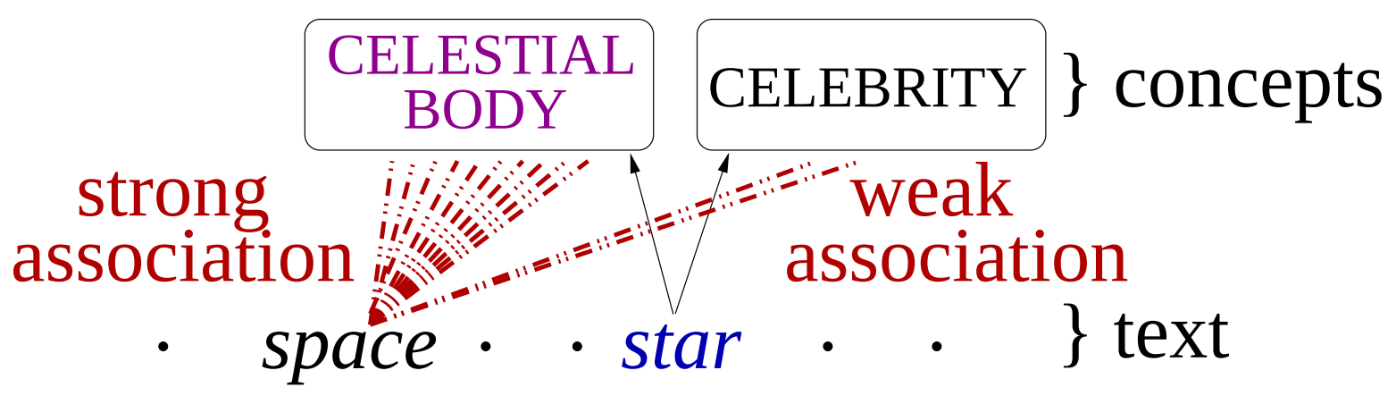
Words that co−occur:



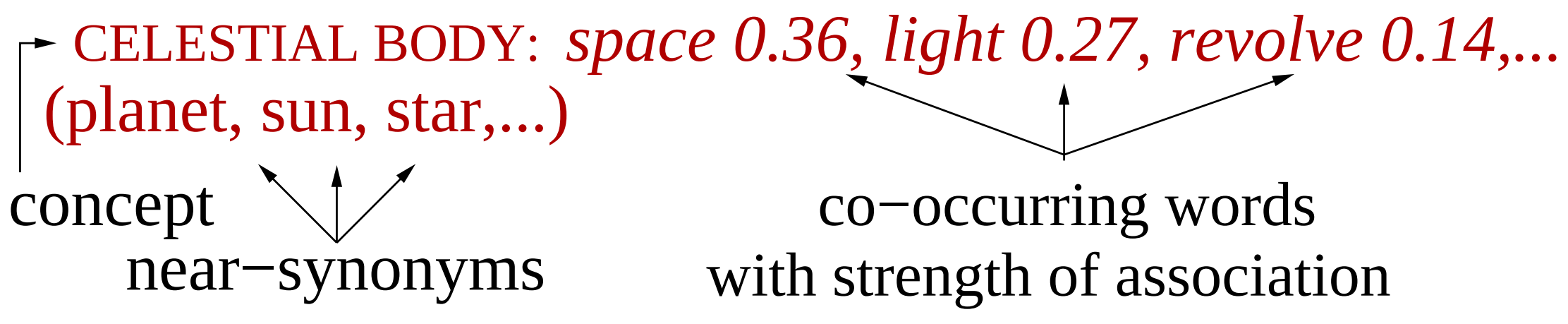
Combining the two

Central Idea: DISTRIBUTIONAL PROFILES OF CONCEPTS

Capturing co-occurrence associations between words and concepts



Distributional Profile of a concept (DPC)


CELESTIAL BODY: space 0.36, light 0.27, revolve 0.14,...
(planet, sun, star,...)
concept
near−synonyms
co−occurring words
with strength of association

How to create these DPCs?

You know someone by the company they keep.

Create Word–Category Co-occurrence Matrix (WCCM)

Unsupervised naïve Bayes word sense classifier

$$\text{desired concept} = \underset{c_j \in C}{\operatorname{argmax}}\; P(c_j) \prod_{w_i \in W} P(w_i|c_j)$$

categories/concepts →

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

← words

The WCCM can be used to estimate probabilities (in an unsupervised manner), that are traditionally calculated using sense−annotated data.

$$\frac{\sum_i m_{ij}}{\sum_{i,j} m_{ij}} \qquad \frac{m_{ij}}{\sum_i m_{ij}}$$

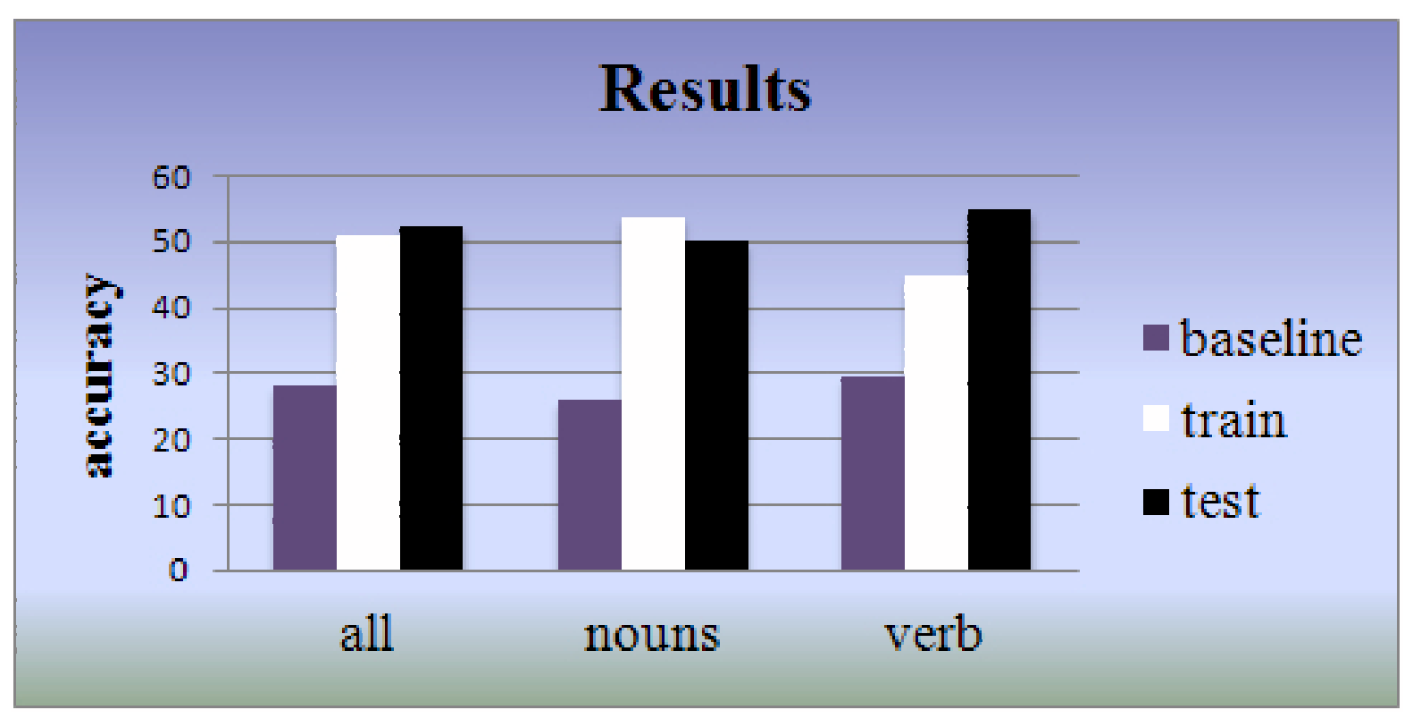
Base WCCM: m_{ij} is the number of times word w_i co-occurs with any word *that has* c_j *as a sense*.
Bootstrapped WCCM: m_{ij} is the number of times word w_i co-occurs with any word *used in sense* c_j.

Apply this monolingually

Apply this cross−lingually

English Lexical Sample Task

Accuracies were markedly better than the random baseline—an increase of more than twenty percentage points.

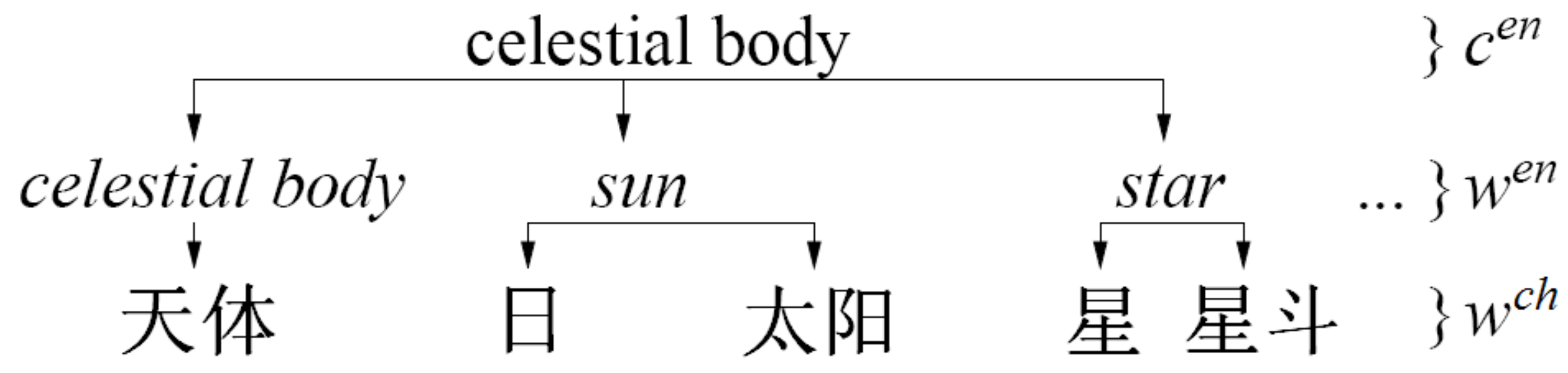


Conclusions

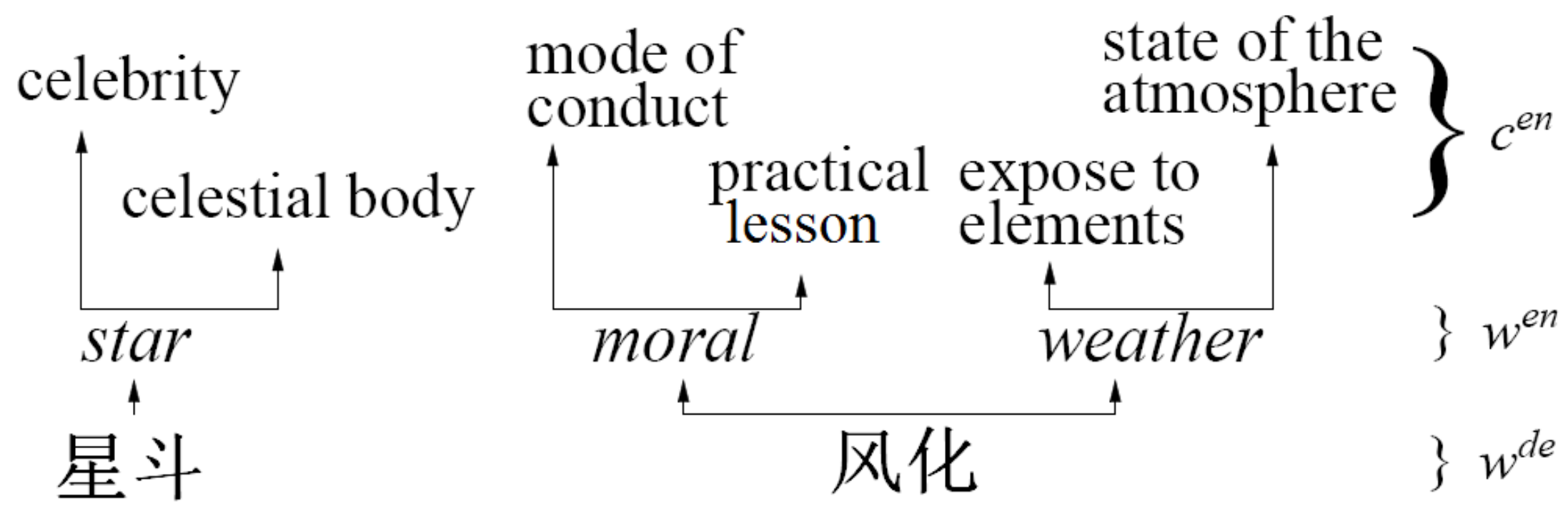
- Placed first among unsupervised systems in the Chinese–English Task.
- Only about 1 percentage point behind the best in the English Lexical Task.
- Cross-lingual DPCs can help automatic machine translation.
- DPCs create simple yet powerful baselines for WSD.

Multilingual Chinese–English Lexical Sample Task

Words having ‘celestial body’ as cross-lingual candidate sense

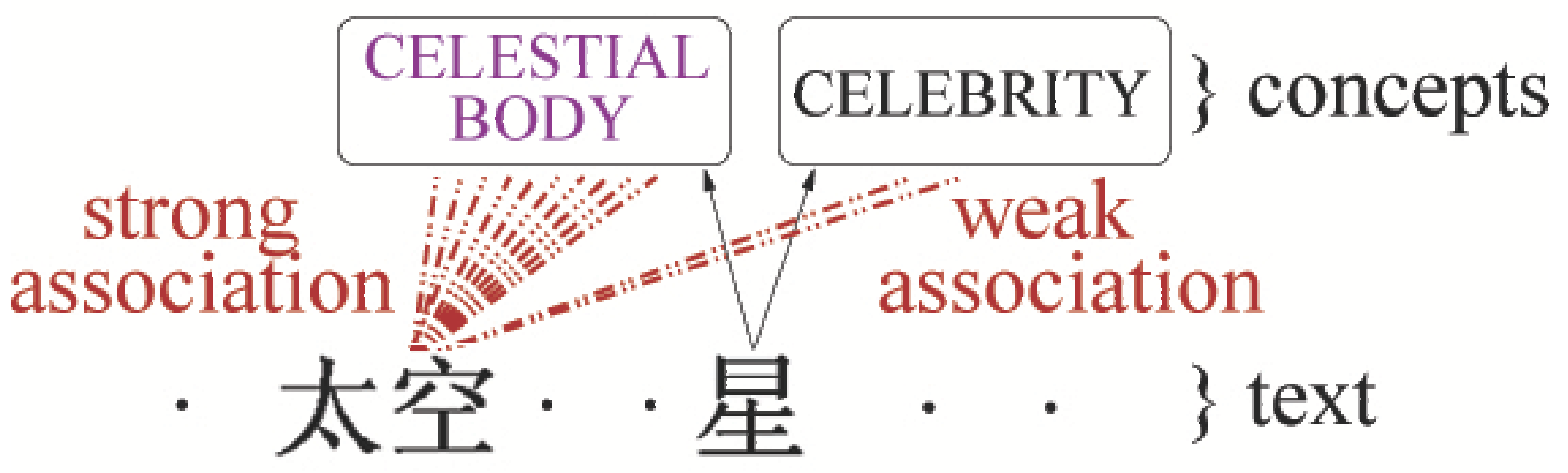


Cross-lingual candidate senses of Chinese words 星斗 and 风化

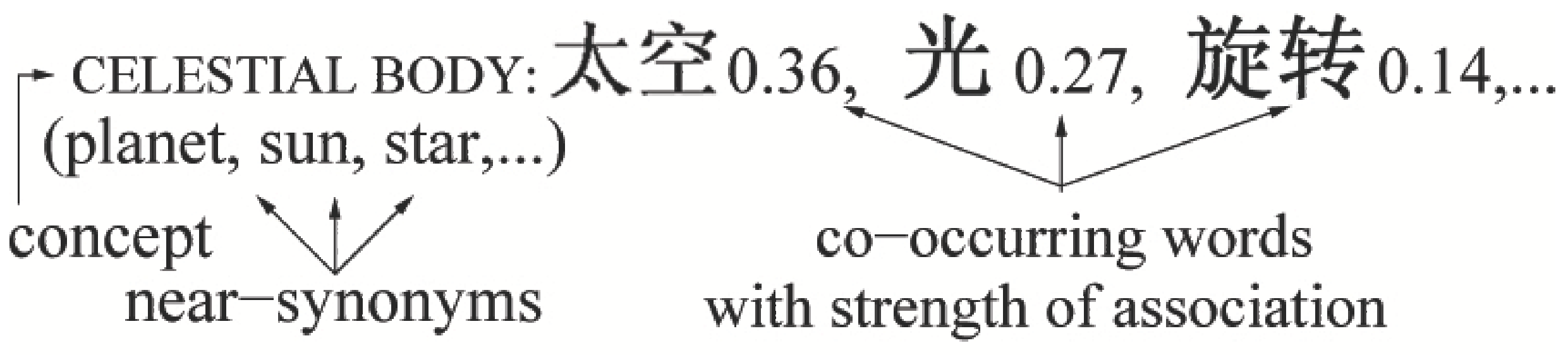


So effectively we have Chinese words with English senses

Capturing co-occurrence associations between Chinese words and English concepts



Cross-lingual Distributional Profiles of Concepts

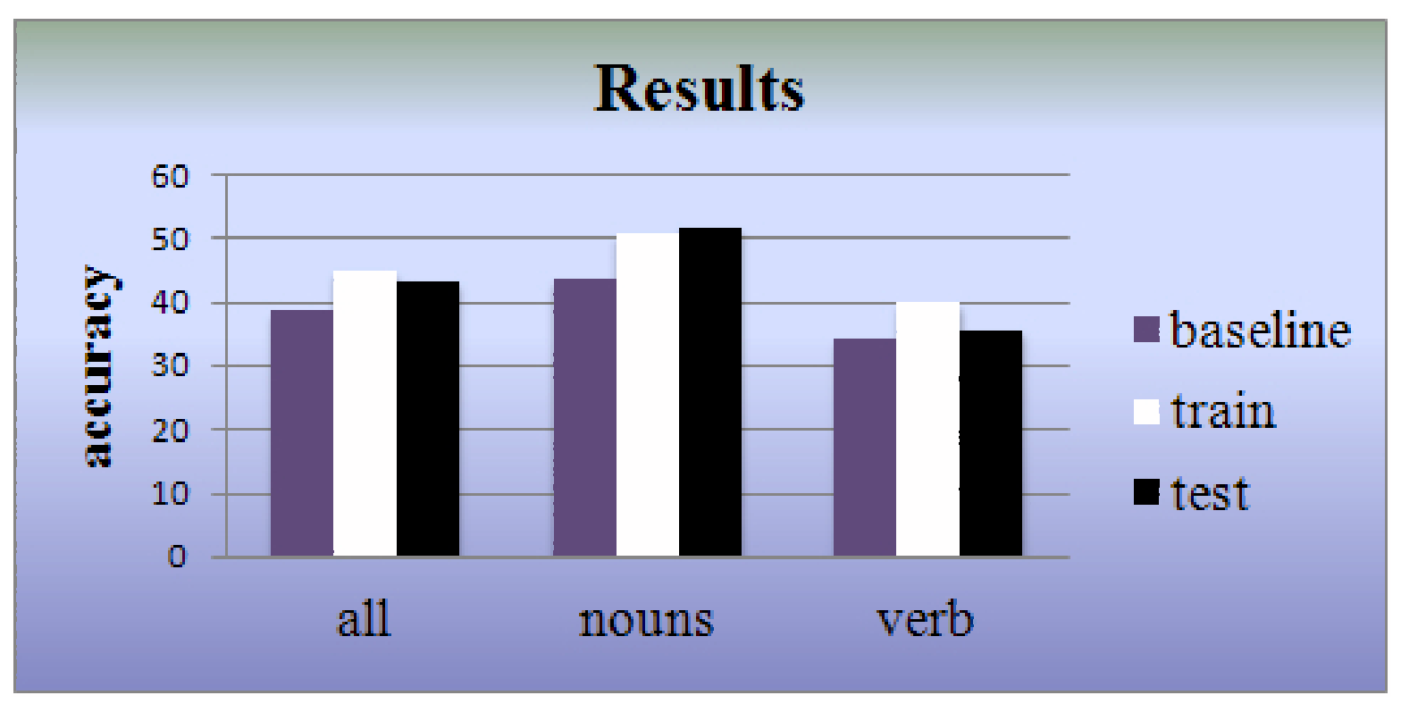

CELESTIAL BODY: 太空 0.36, 光 0.27, 旋转 0.14,...
(planet, sun, star,...)
concept
near−synonyms
co−occurring words
with strength of association

Chinese word–English category co-occurrence matrix

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

Base WCCM: m_{ij} is the number of times Chinese word w_i^{ch} co-occurs with a word *that has* c_j *as English sense*.

Bootstrapped WCCM: m_{ij} is the number of times Chinese word w_i^{ch} co-occurs with a word *used in English sense* c_j.



See how cross-lingual DPCs can be used to obtain state-of-the-art semantic distance accuracies in a resource-poor language using a knowledge source from a resource-rich one.

Come to EMNLP’s Friday morning session