Neural Vector Representations beyond Words: Sentence and Document Embeddings

Gerard de Melo

http://gerard.demelo.org

Rutgers University

Outline

- Word Representations
- Phrase Representations
- Sentence Representations
- Document Representations
- Applications and Outlook

Phrases/Multiword Expressions

Kind	Example
Compound Noun	dog park
Adjective Noun	fresh food
Verb Object	win money
Verb Particle Phrase	walk up the stairs
Named Entities	East London
...	...

Compositional Models

Models for Composition

Composition Function

$$\mathbf{v}_p = f(\mathbf{v}_{w_1}, \ldots, \mathbf{v}_{w_n})$$

Vector for phrase,
e.g. “mobile phone”

Vectors for parts of phrase
e.g. “mobile” and “phone”

Models for Composition: Elementwise Composition

$$f(\mathbf{v}_{w_1}, \ldots, \mathbf{v}_{w_n}) = \mathbf{v}_{w_1} + \cdots + \mathbf{v}_{w_n}$$

Vector 1		Vector 2		Result Vector
0.0		0.0		0.0
0.3		0.1		0.4
0.4	+	0.2	=	0.6
0.0		0.1		0.1
0.3		0.0		0.3
0.2		0.0		0.2

Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. ACL 2008

Models for Composition: Elementwise Composition

$$f(\mathbf{v}_{w_1}, \ldots, \mathbf{v}_{w_n}) = \mathbf{v}_{w_1} \odot \cdots \odot \mathbf{v}_{w_n}$$

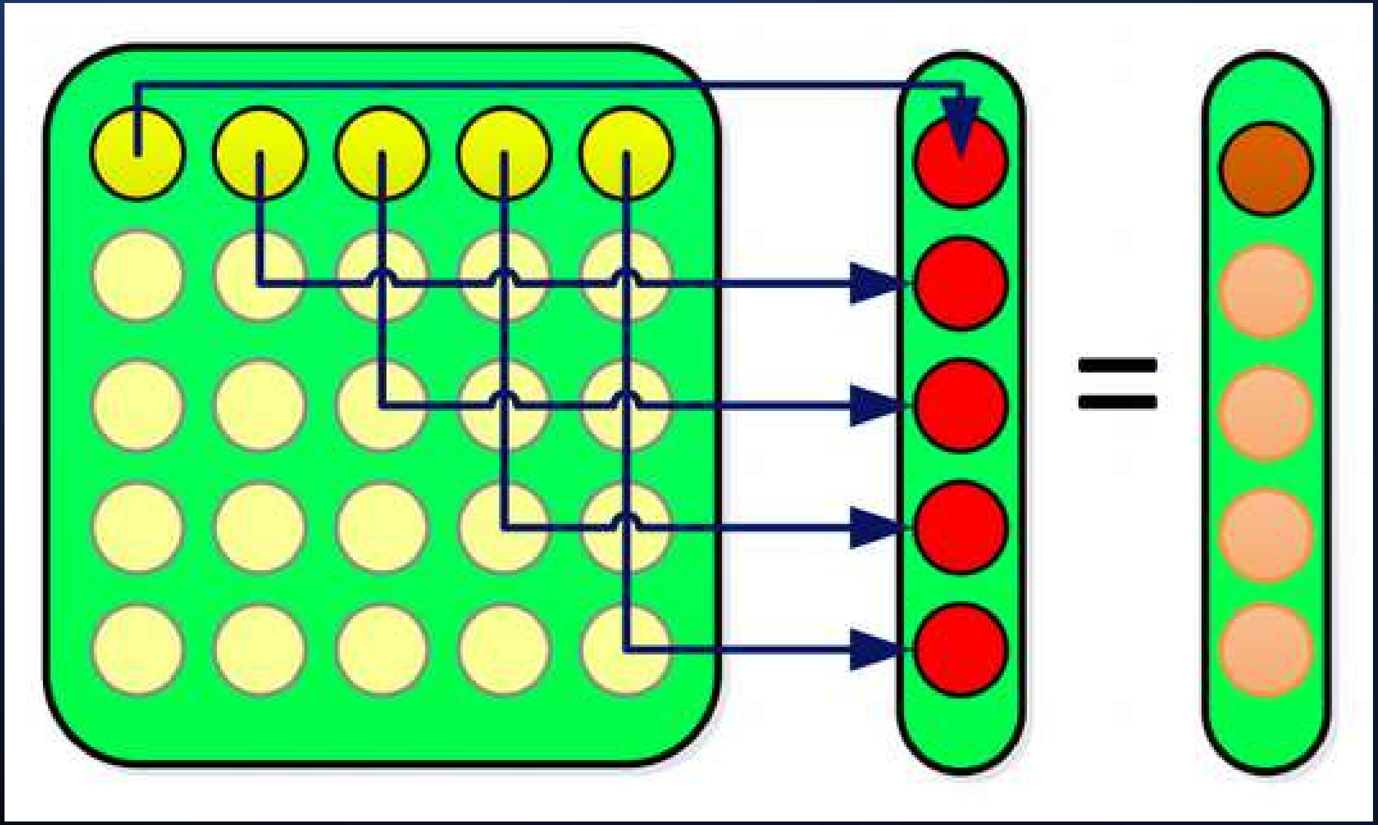
Elementwise multiplication

Vector 1		Vector 2		Result Vector
0.0		0.0		0.00
0.3		0.1		0.03
0.4	+	0.2	=	0.08
0.0		0.1		0.00
0.3		0.0		0.00
0.2		0.0		0.00

Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. ACL 2008

Models for Composition: Tensor Models

Tensors that can be applied to one or more vectors for arguments



Bob Coecke, Mehrnoosh Sadrzadeh, Stephen Clark (2010)

Phrases/Multiword Expressions

Kind	Example
Compound Noun	dog park
Adjective Noun	fresh food
Verb Object	win money
Verb Particle Phrase	walk up the stairs
Named Entities	East London
…	…

Phrases/Multiword Expressions

Kind	Compositional Example	Non-Compositional Example
Compound Noun	dog park	zebra crossing
Adjective Noun	fresh food	hot dog
Verb Object	win money	kick the bucket
Verb Particle Phrase	walk up the stairs	wrap up the session
Named Entities	East London	Los Angeles
…	…	…

Learning from Text

word2vec Implementation

1. **PMI-like scores to find frequent MWEs**

Bigram Count

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

Unigram Counts

2. **Select MWEs above threshold, then repeat to find even longer MWEs**

3. **Finally, simply treat MWEs as unigrams**

Tomas Mikolov, Ilya Sutskever, Kai Chen. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013

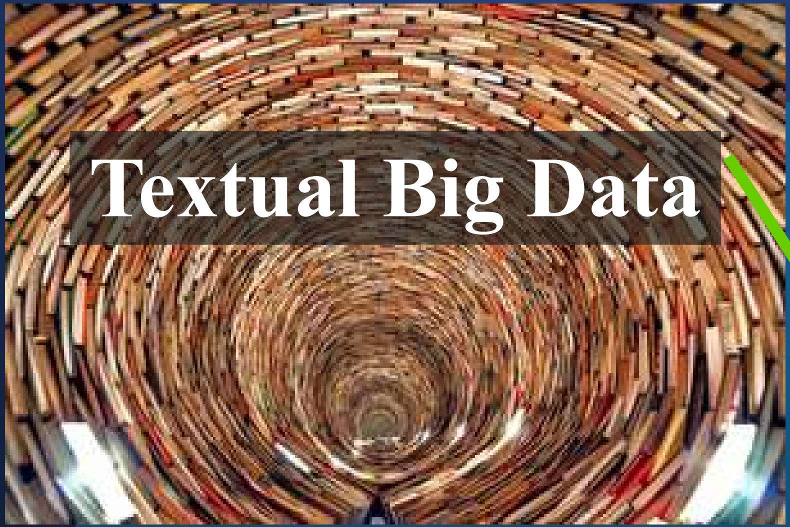
word2vec Implementation

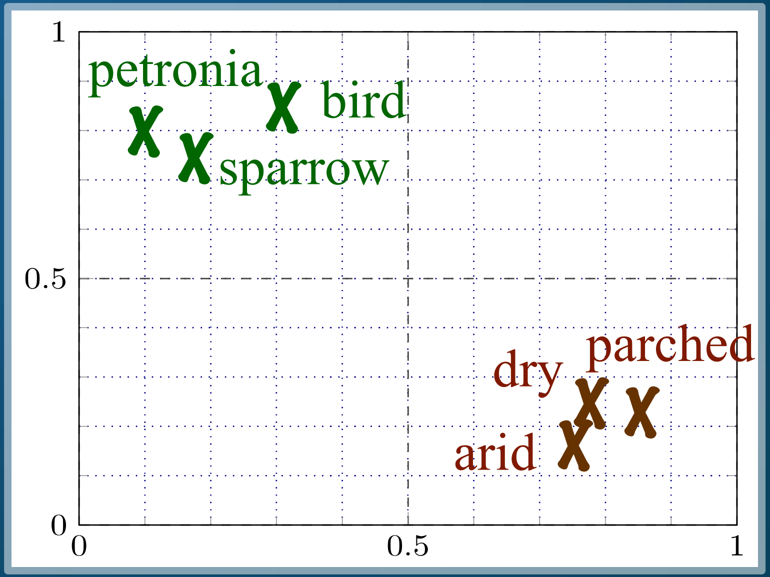
Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna Check crown Polish zolty CTK	Hanoi Ho Chi Minh City Viet Nam Vietnamese	airline Lufthansa carrier Lufthansa flag carrier Lufthansa Lufthansa	Moscow Volga River upriver Russia	Juliette Binoche Vanessa Paradis Charlotte Gainsbourg Cecile De

Nearest neighbours of Component-wise sums are meaningful

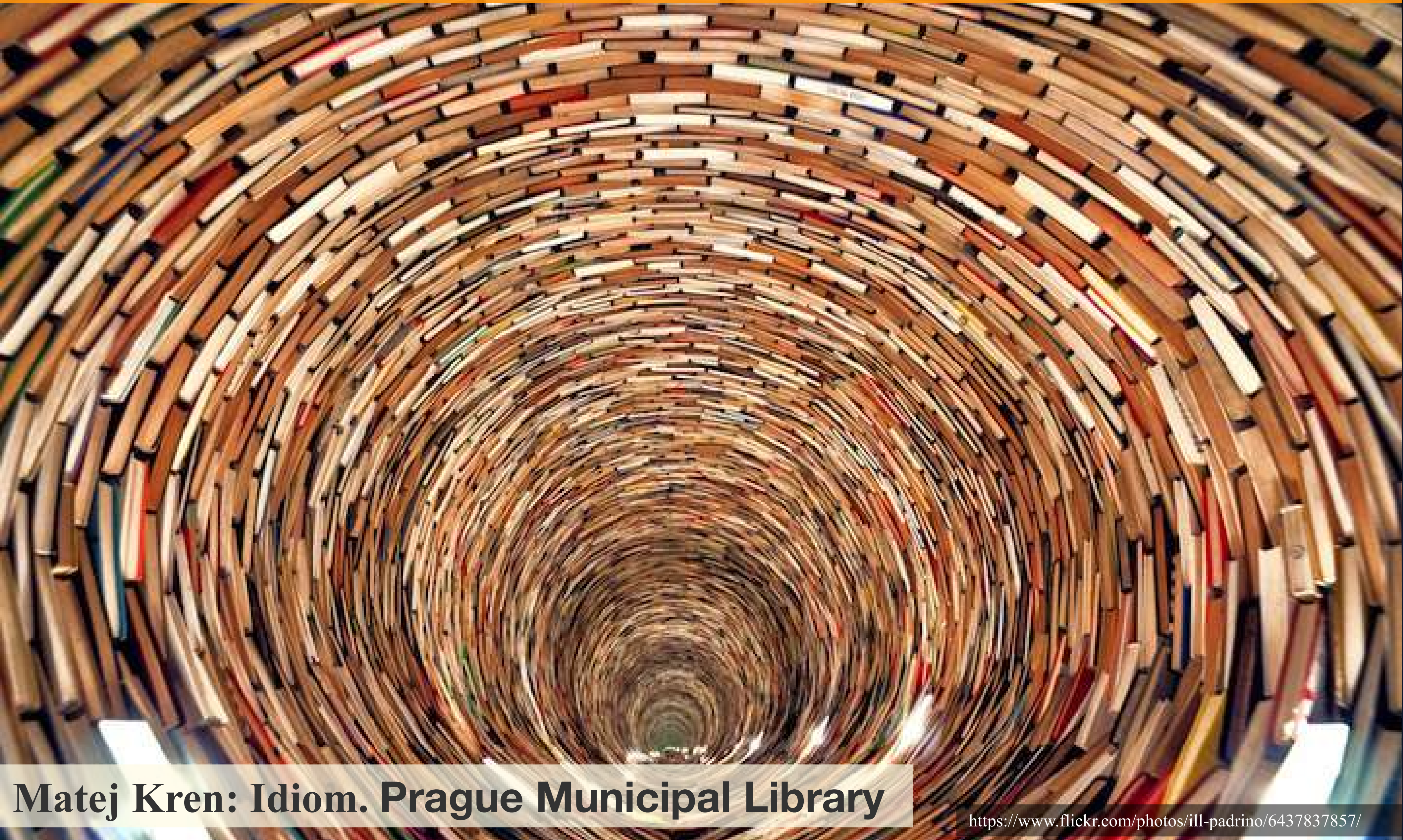
Tomas Mikolov, Ilya Sutskever, Kai Chen. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013

Learning from Heterogeneous Data



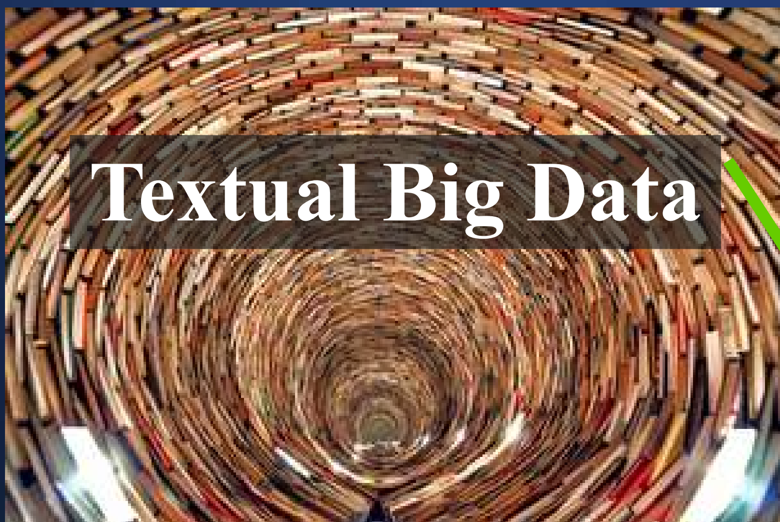


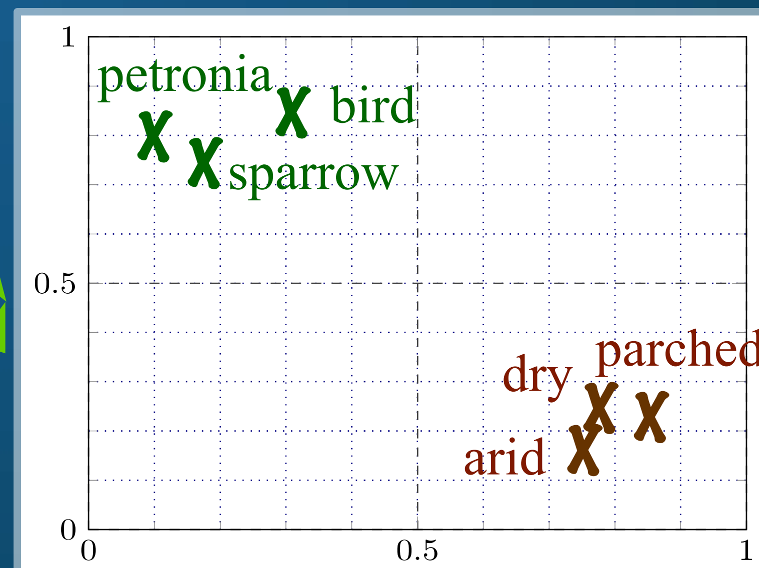
Textual Big Data



High-Quality Knowledge

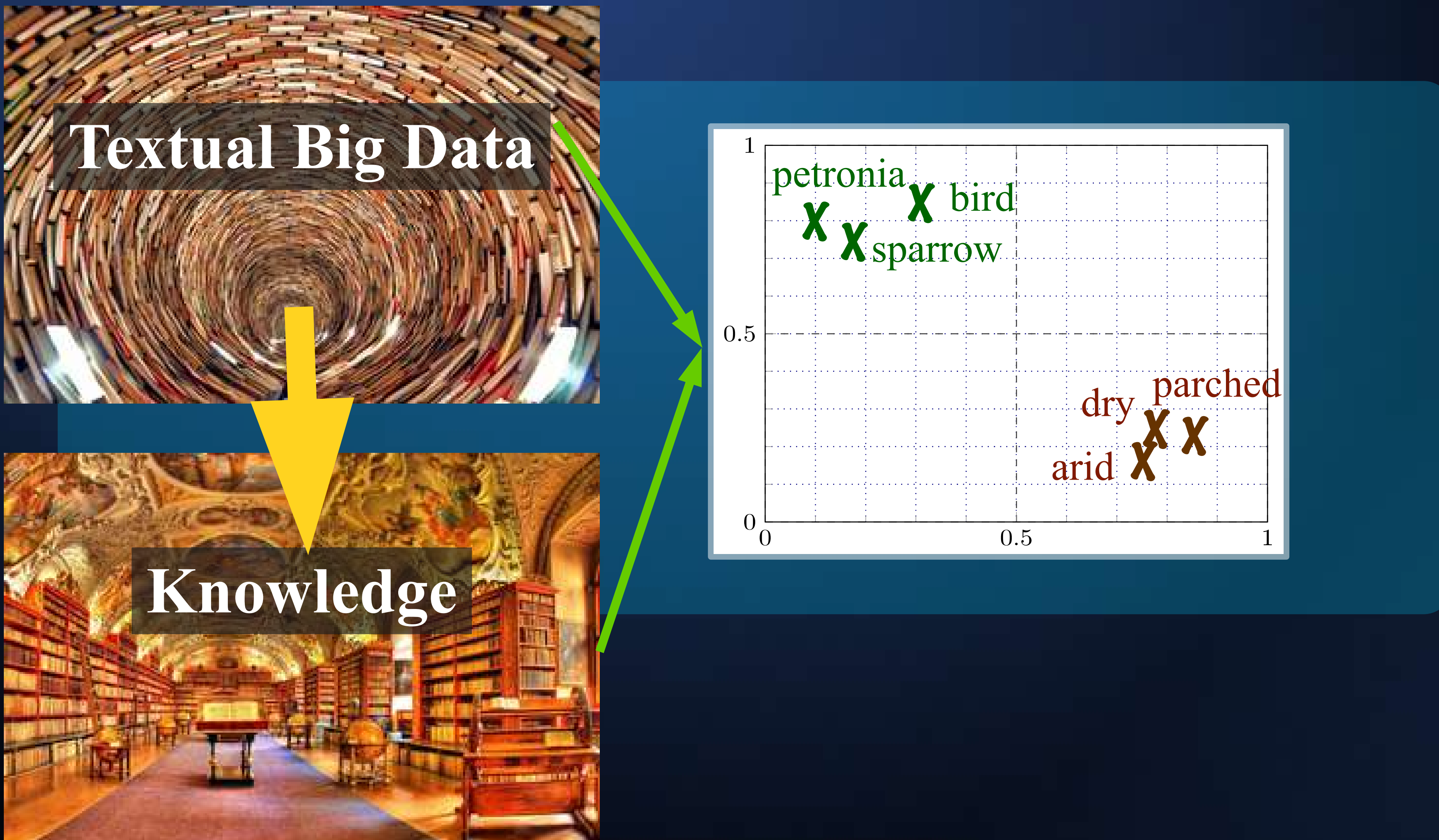
Learning from Heterogeneous Data





Gerard de Melo (2015). Wiktionary-Based Word Embeddings.

Information Extraction Approach



Information Extraction Approach

semantic!

…Greek and Roman mythology...

look for semantically salient contexts in text!

Jiaqiang Chen and Gerard de Melo (2015). Semantic Information Extraction for Improved Word Embeddings

Information Extraction Approach: Joint Training

Given an extracted pair of semantically related words, the intuition is that the embeddings for the two words should be pulled together

$$\max \frac{1}{T} \sum_{t=1}^{T} \sum_{w_r} \log p(w_r | w_t)$$

w_r: related words of w_t

Best Paper Award at NAACL Vector Space Modeling Workshop

Jiaqiang Chen and Gerard de Melo. Semantic Information Extraction for Improved Word Embeddings

Information Extraction Approach: Joint Training

$$\max \frac{1}{T} \sum_{t=1}^{T} \sum_{w_r} \log p(w_r | w_t)$$

softmax function $p(w_r | w_t) = \frac{\exp\left({V'_{w_r}}^T V_{w_t}\right)}{\sum_{w=1}^{W} \exp({V'_w}^T V_{w_t})}$

To compute its gradient is impractical. It is proportional to T, which is often large($10^5 - 10^7$ terms)

Best Paper Award at NAACL Vector Space Modeling Workshop

Case Study: Information Extraction

List Extraction

- Look for **repeated** occurrences of **commas**
- Short units of roughly equal length
- noun phrases, adjectives

Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Case Study: Information Extraction

List Extraction

- Look for **repeated** occurrences of **commas**
- Short units of roughly equal length
- noun phrases, adjectives
- Also: Hearst patterns, e.g.
  "cities **such as** <u>New York</u>, <u>London</u>, ..."

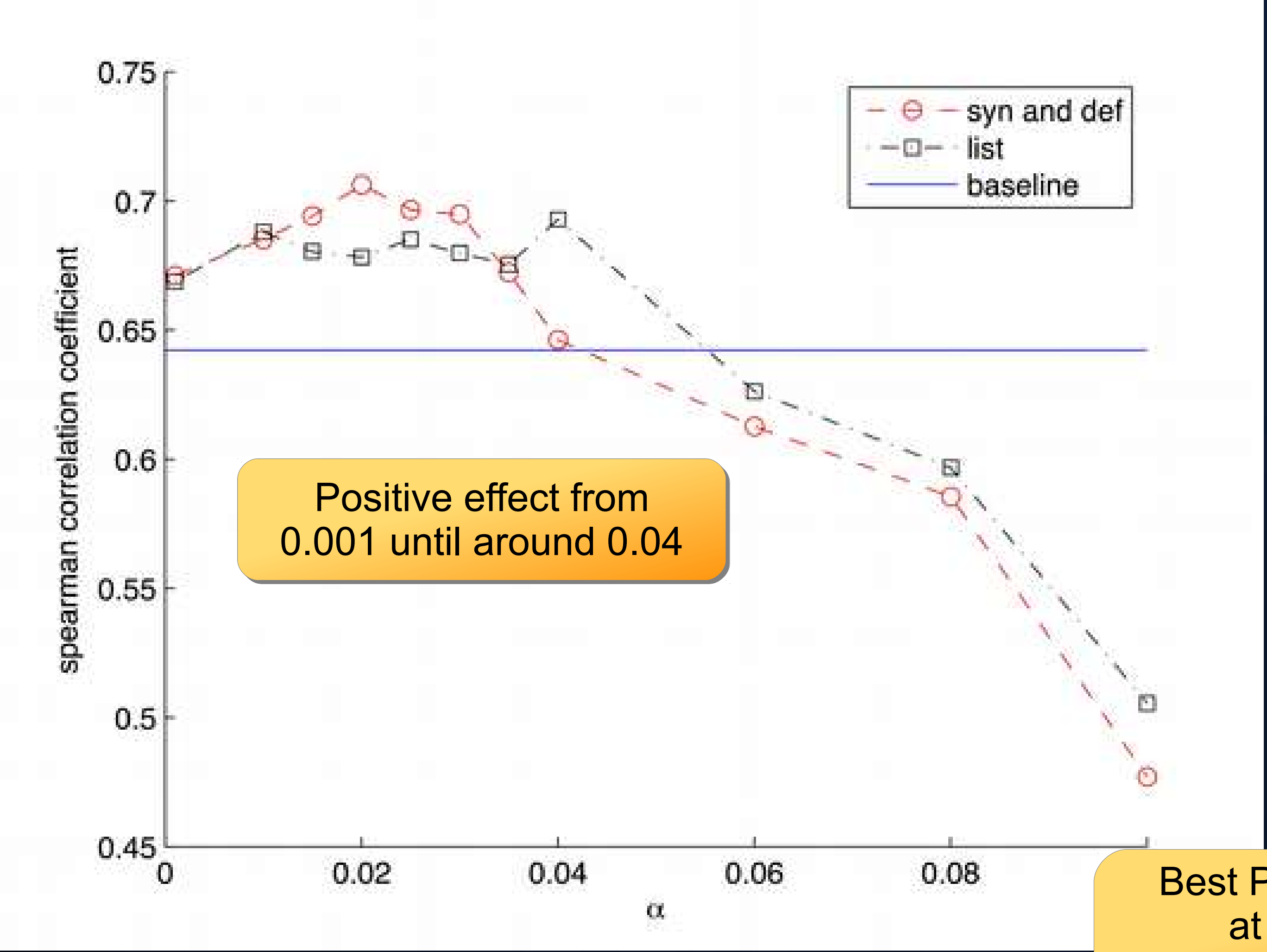
Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Jiaqiang Chen and Gerard de Melo. Semantic Information Extraction for Improved Word Embeddings

Case Study: Information Extraction

Extracted Lists
player captain manager director vice-chairman
group race culture religion organisation person person
Italian Mexican Chinese Creole French
Self-Portraits Portraits iris Still-Lives with Sunflowers view from the Asylum Works after Millet Vineyards
ballscrews leadscrews worm gear screwjacks linear actuator
Cleveland Essex Lincolnshire Northamptonshire Nottinghamshire Thames Valley South Wales
ant.py dimdriver.py dimdriverdatafile.py dimdriverdatasetdef.py dimexception.py dimmaker.py dimoperators.py dimparser.py dimrex.py dimension.py

Case Study: Results on WS353

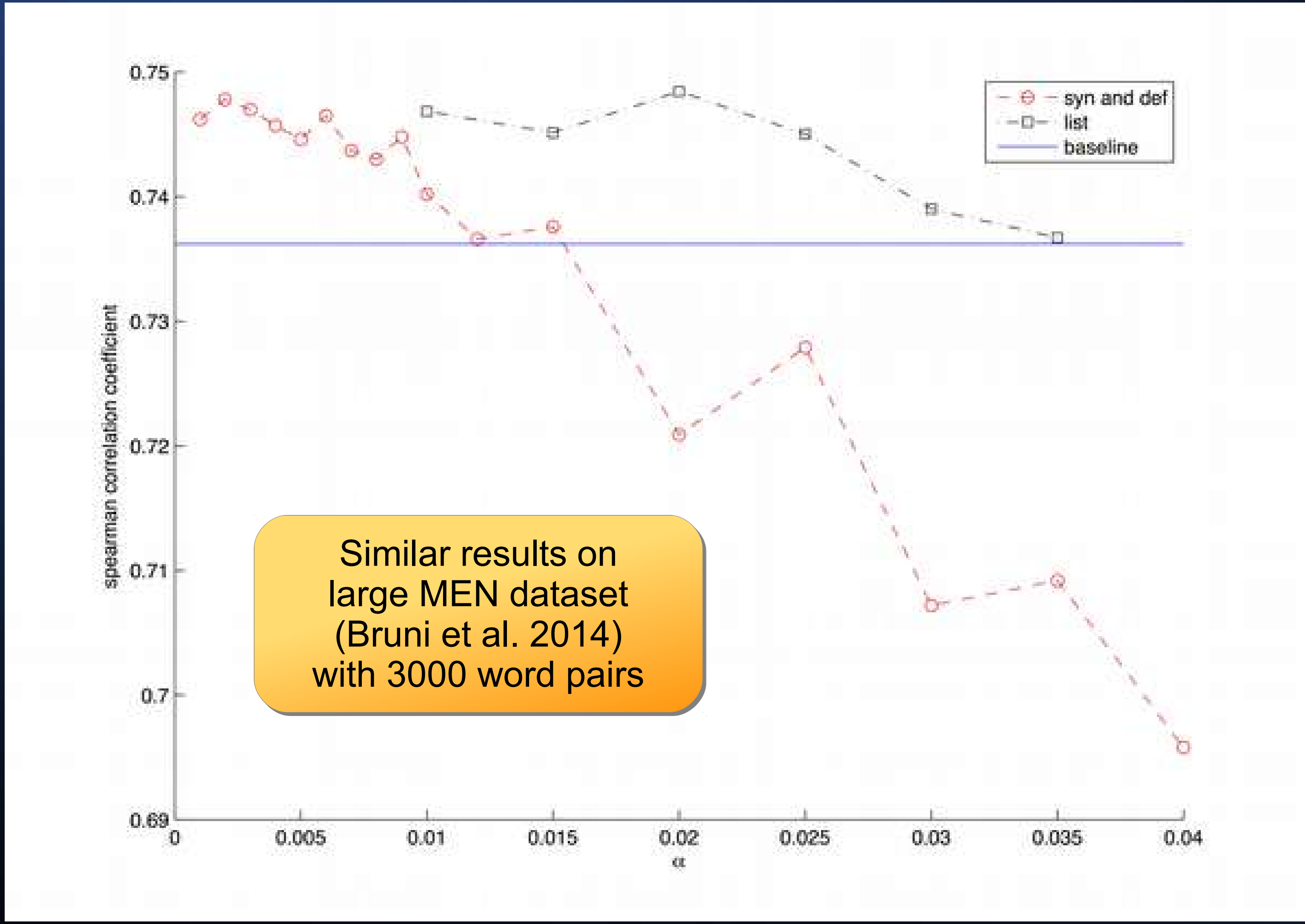


Positive effect from 0.001 until around 0.04

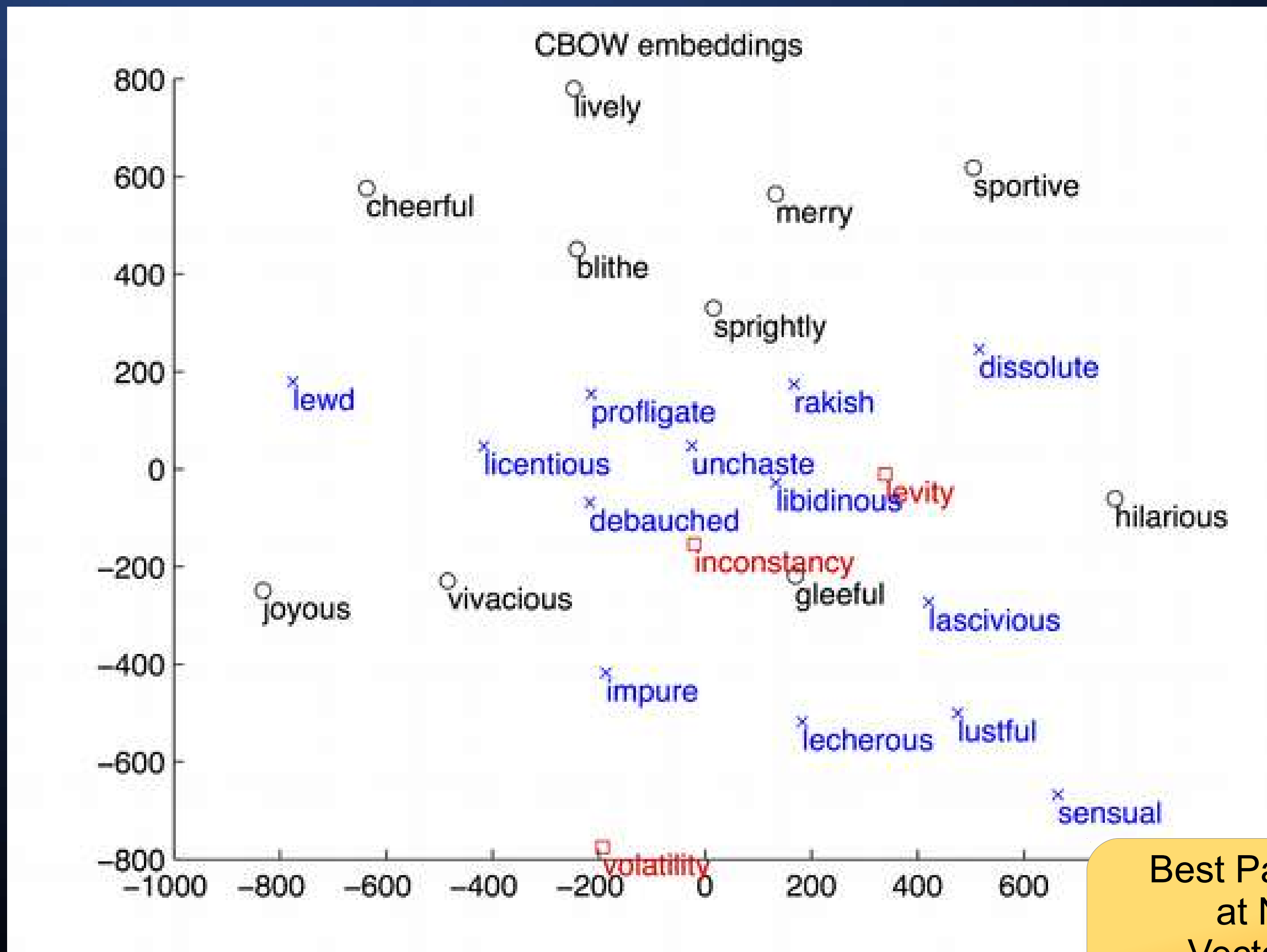
Best Paper Award at NAACL Vector Space Modeling Workshop

Jiaqiang Chen and Gerard de Melo. Semantic Information Extraction for Improved Word Embeddings

Case Study: Results on MEN



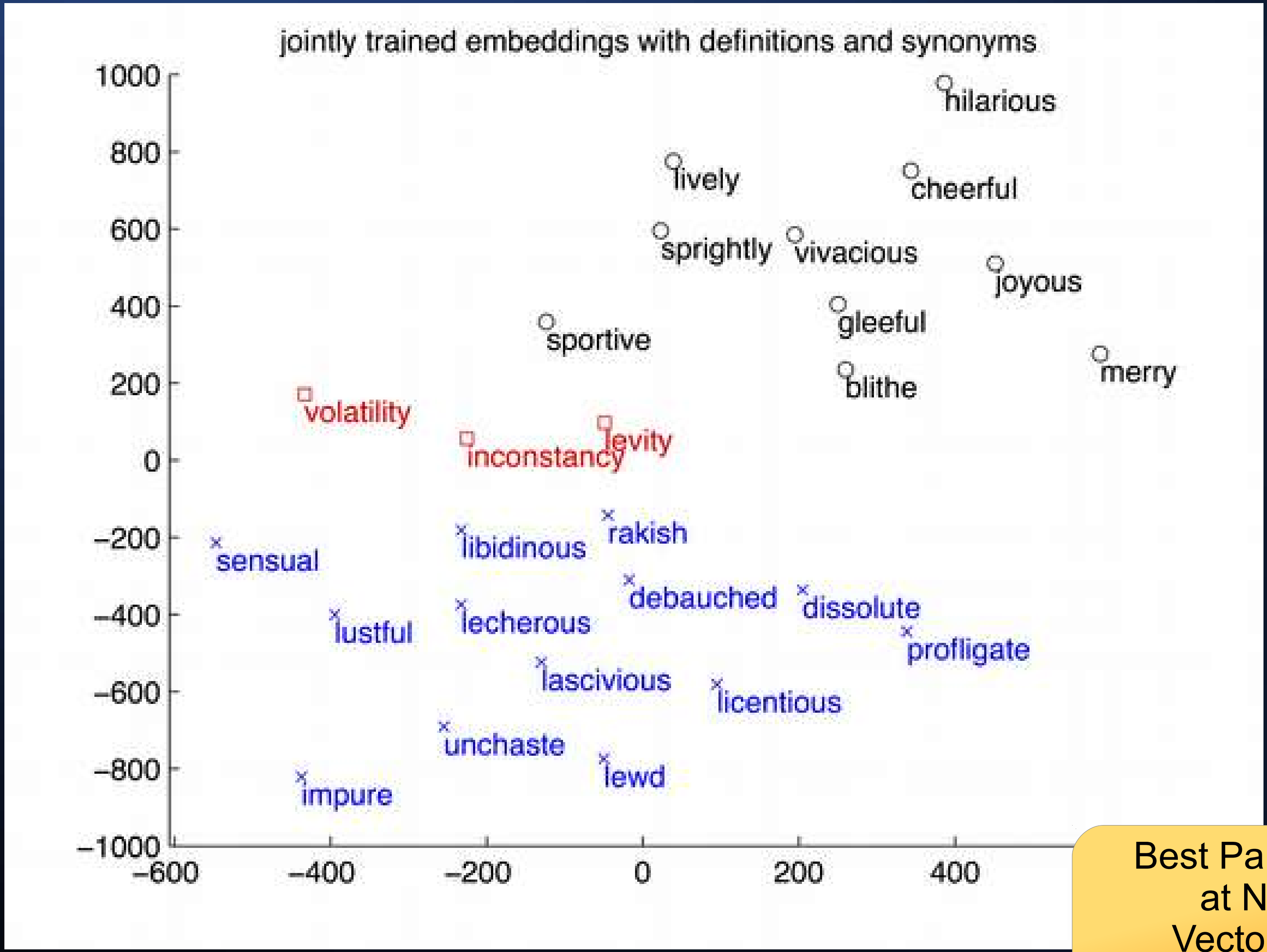
Case Study: Example



Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Jiaqiang Chen and Gerard de Melo. Semantic Information Extraction for Improved Word Embeddings

Case Study: Example

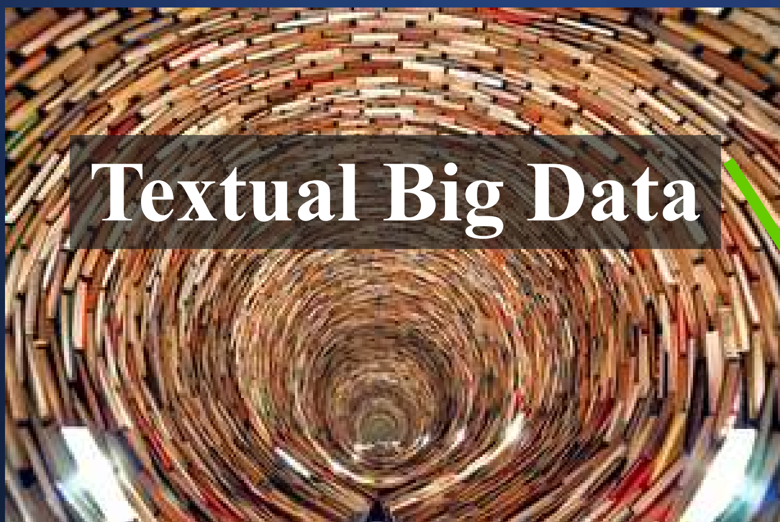


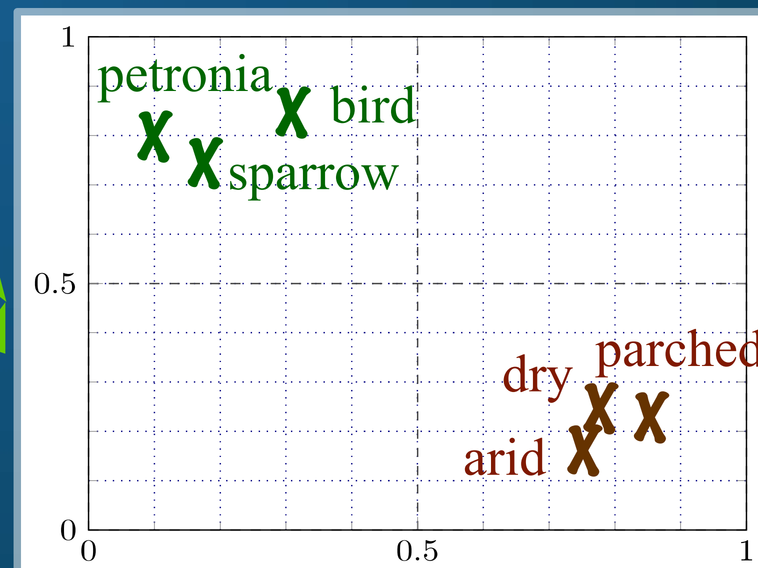
Best Paper Award at NAACL Vector Space Modeling Workshop

Jiaqiang Chen and Gerard de Melo. Semantic Information Extraction for Improved Word Embeddings

Learning from Text and Structured Data

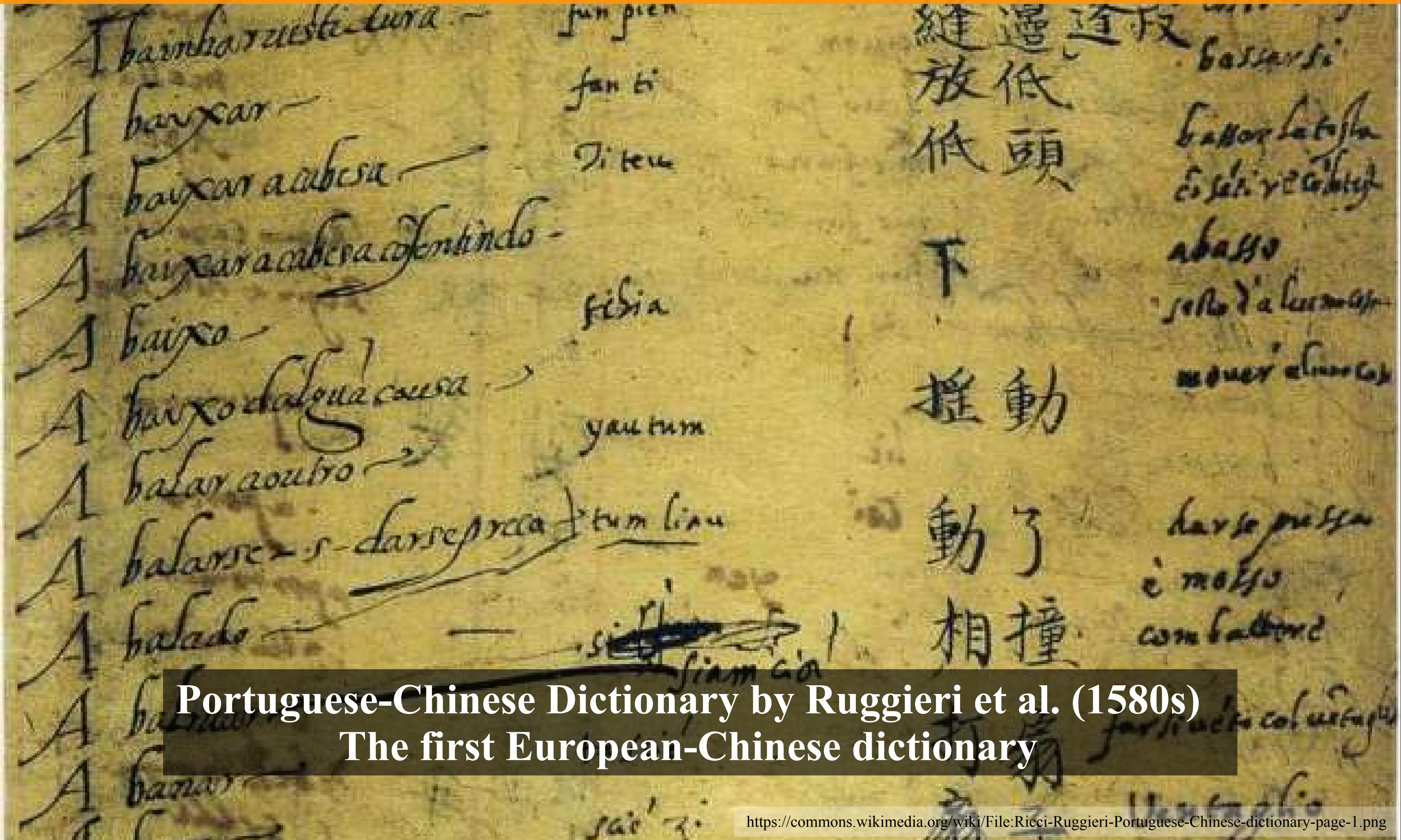
Learning from Heterogeneous Data





Gerard de Melo (2015). Wiktionary-Based Word Embeddings.

Lexical Knowledge



Portuguese-Chinese Dictionary by Ruggieri et al. (1580s)
The first European-Chinese dictionary

https://commons.wikimedia.org/wiki/File:Ricci-Ruggieri-Portuguese-Chinese-dictionary-page-1.png

Lexical Knowledge: Wiktionary

Provides translations, antonyms, etc.

Lexical Knowledge: Wiktionary

Lexical Knowledge: Wiktionary

Etymology 1 [edit]

From Latin *pulsus* ("beat"), from *pellere* ("to drive"), from Proto-Indo-European **pel* ("to drive, strike, thrust").

For spelling, the *-e* (on *-lse*) is so the end is pronounced /ls/, rather than /lz/ as in *pulls*, and does not change the vowel ('u'). Compare *else*, *false*, *convulse*.

Pronunciation [edit]

- IPA(key): /pʌls/
- Audio (US) 0:00 MENU

Noun [edit]

pulse (*plural* **pulses**)

1. (*physiology*) A normally regular beat felt when arteries are depressed, caused by the pumping action of the heart.
2. A beat or throb. [quotations ▼]
3. (*music*) The beat or tactus of a piece of music.
4. An autosoliton.

Related terms [edit]

- impulse
- repulse

Translations [edit]

regular beat caused by the heart [hide ▲]

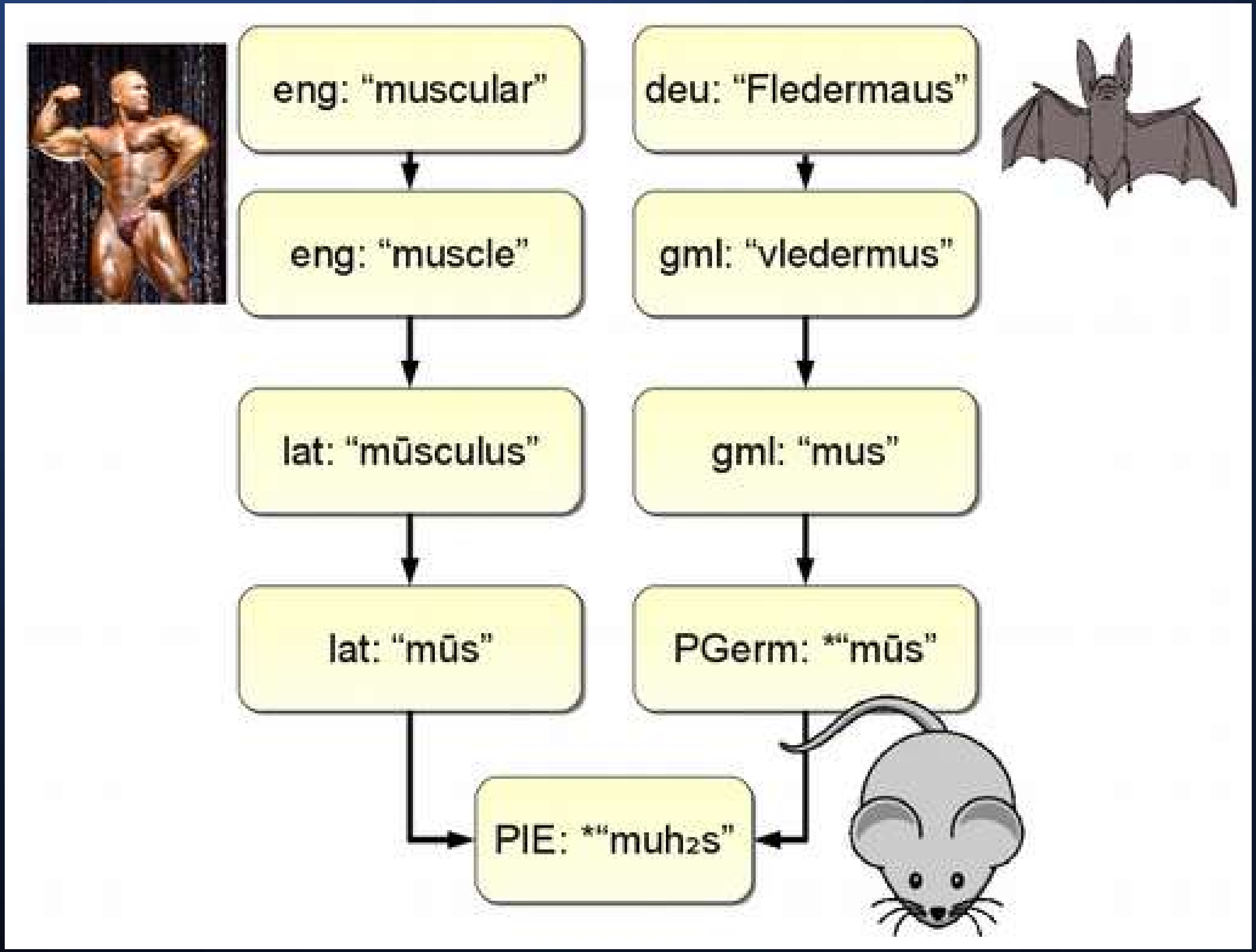
Select targeted languages

- Arabic: نبضة *f* (nábDa)
- Chinese:
  - Mandarin: 脈搏, 脉搏 (zh) (màibó), 脈 (zh), 脉 (zh) (mài)
- Czech: tep (cs) *m*, puls *m*
- Dutch: pols (nl) *m*
- Esperanto: pulso (eo)
- Faroese: æðrasláttur *m*
- Finnish: pulssi
- French: pouls (fr) *m*
  - Old French: poulz *m*
  - Middle French: pouls *m*
- Irish: cuisle *f*
- Italian: polso *m*, battito (it) *m*
- Japanese: 脈拍 (みゃくはく, myakuhaku), 脈 (ja) (みゃく, myaku)
- Norman: pouls *m*
- Korean: 맥박 (ko) (maekbak) (脈搏)
- Malay: nadi
- Norwegian: puls
- Persian: نبض (fa) (nabz)
- Portuguese: pulso (pt) *m*
- Russian: пульс (ru) *m* (pul's)
- Slovak: pulz *m*

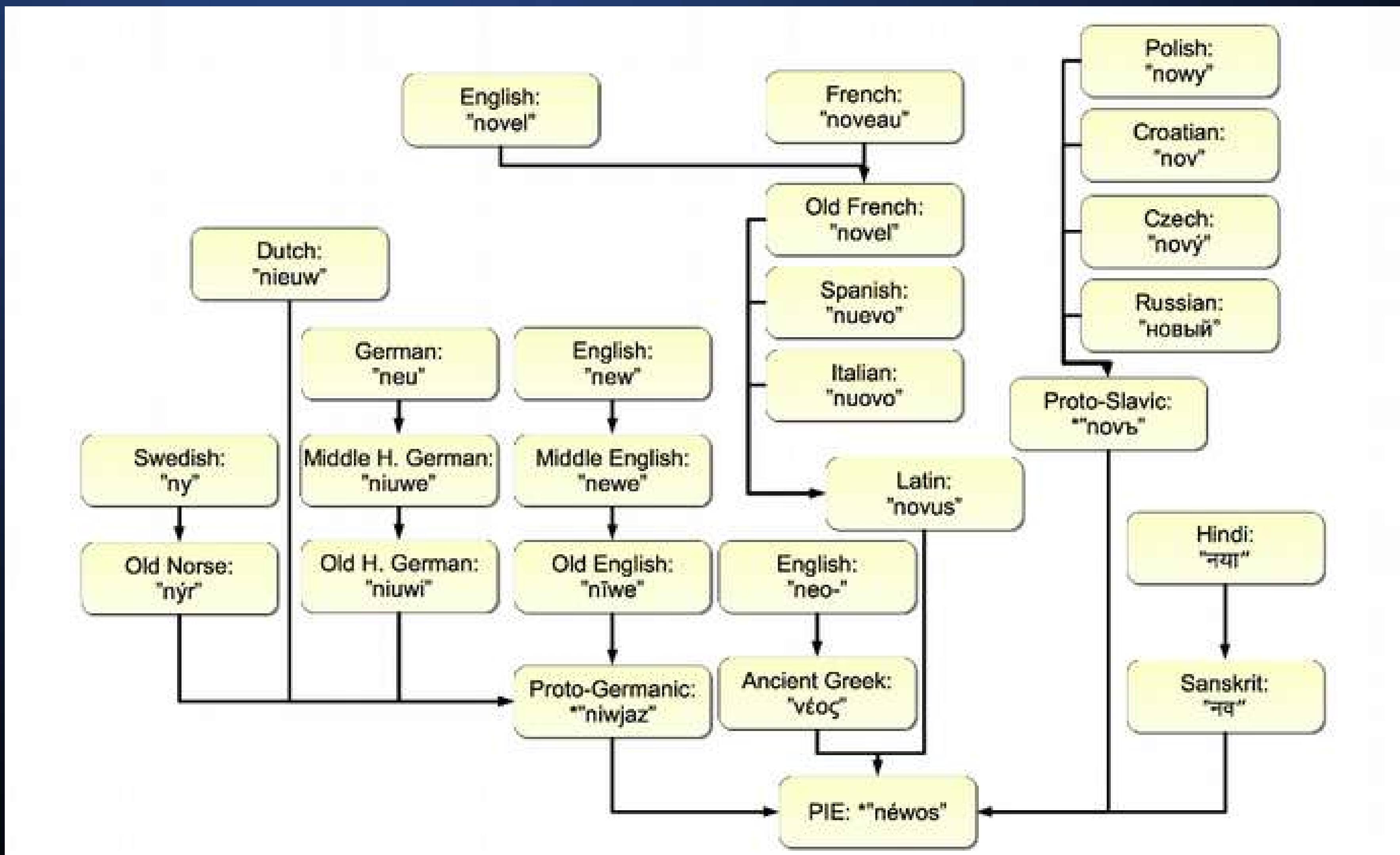
Etymological Wordnet

e.g. “salary” < Lat. “salarius” < Lat. “sal” (salt)

Etymological Wordnet



Etymological Wordnet



LREC 2014

Lexical Knowledge: Wiktionary

Etymology 1 [edit]

From Latin *pulsus* ("beat"), from *pellere* ("to drive"), from Proto-Indo-European **pel* ("to drive, strike, thrust").

For spelling, the *-e* (on *-lse*) is so the end is pronounced /ls/, rather than /lz/ as in *pulls*, and does not change the vowel ('u'). Compare *else*, *false*, *convulse*.

Pronunciation [edit]

- IPA(key): /pʌls/
- Audio (US) 0:00 MENU

Noun [edit]

pulse (*plural* **pulses**)

1. (*physiology*) A normally regular beat felt when arteries are depressed, caused by the pumping action of the heart.
2. A beat or throb. [quotations ▼]
3. (*music*) The beat or tactus of a piece of music.

Related terms [edit]

- impulse
- repulse

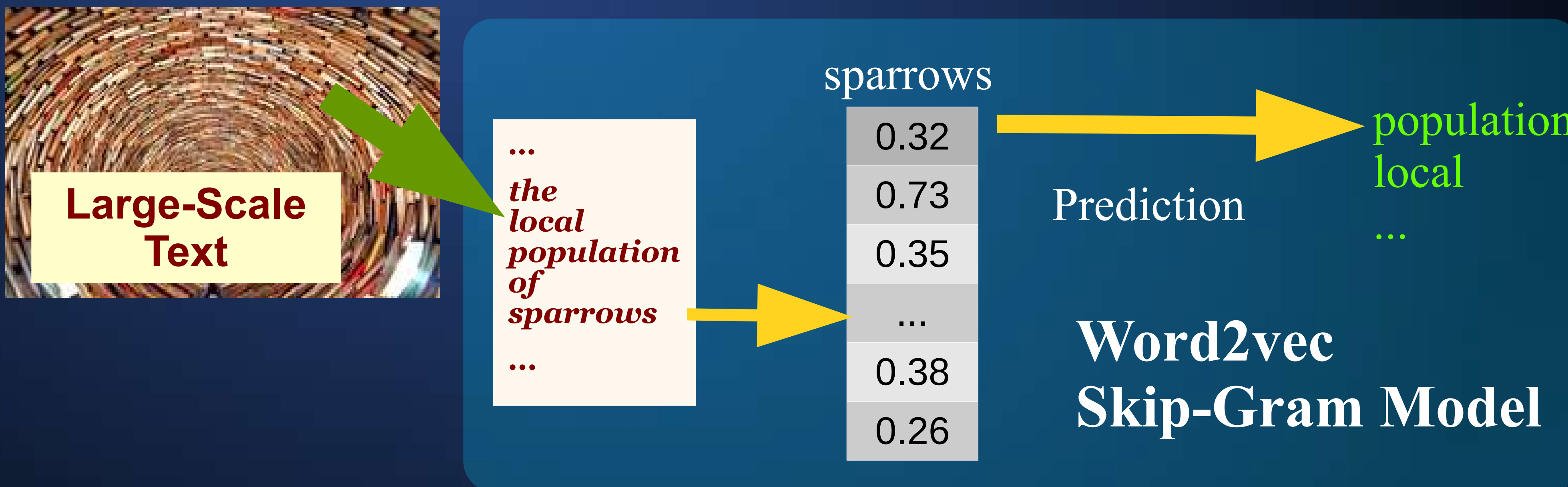
Translations [edit]

regular beat caused by the heart [hide ▲]

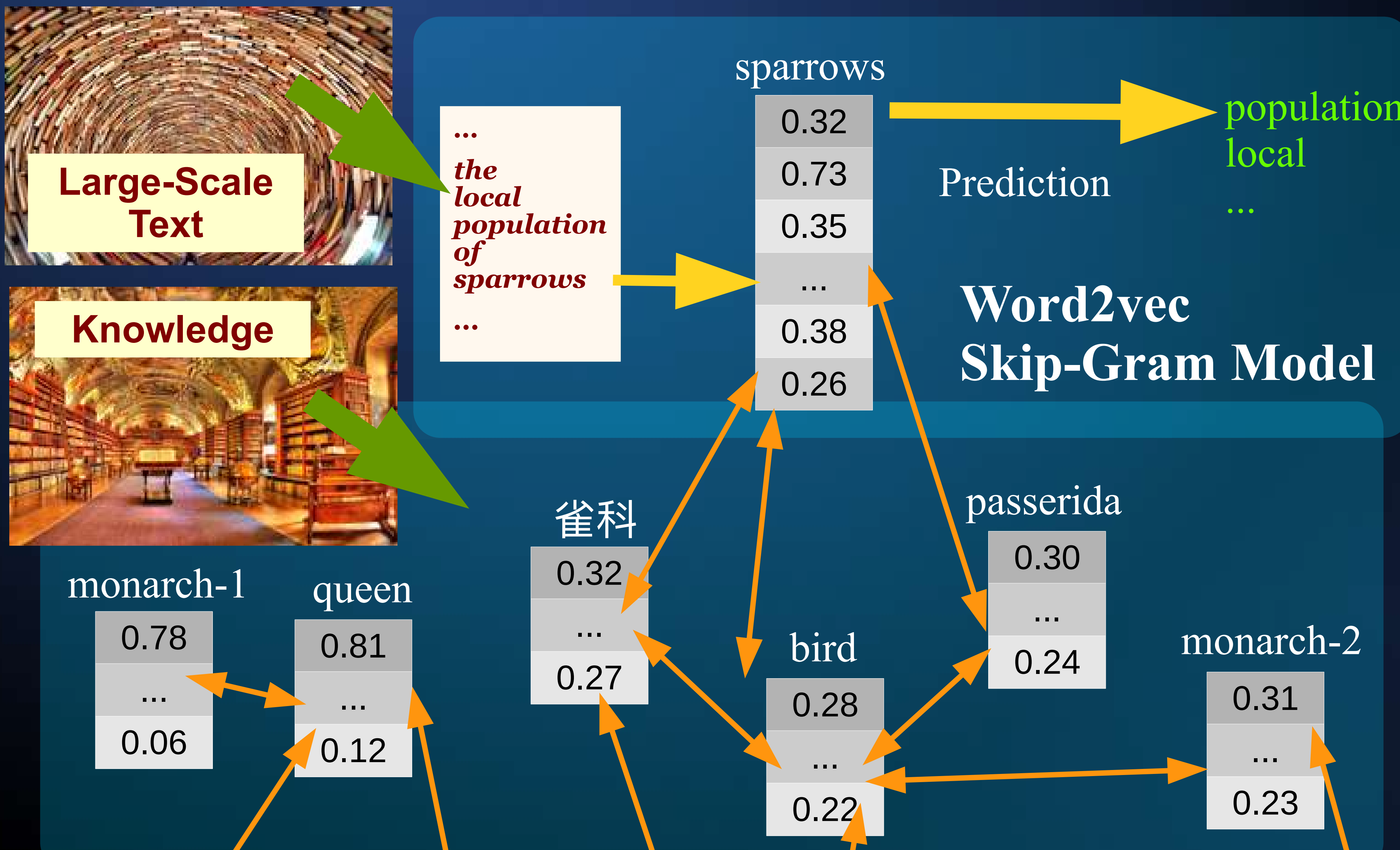
Select targeted languages

- Arabic: نبضة *f* (nābDa)
- Chinese:
  - Mandarin: 脈搏, 脉搏 (zh) (màibó), 脈 (zh), 脉 (zh) (mài)
- Czech: tep (cs) *m*, puls *m*
- Dutch: pols (nl) *m*
- Esperanto: pulso (eo)
- Faroese: æðrasláttur *m*
- Finnish: pulssi
- French: pouls (fr) *m*
  - Old French: poulz *m*
  - Middle French: pouls *m*
- Irish: cuisle *f*
- Italian: polso *m*, battito (it) *m*
- Japanese: 脈搏 (みゃくはく, myakuhaku), 脈 (ja) (みゃく, myaku)
- Norman: pouls *m*
- Korean: 맥박 (ko) (maekbak) (脈搏)
- Malay: nadi
- Norwegian: puls
- Persian: نبض (fa) (nabz)
- Portuguese: pulso (pt) *m*
- Russian: пульс (ru) *m* (pul's)
- Slovak: pulz *m*

Word Vector Representations: word2vec

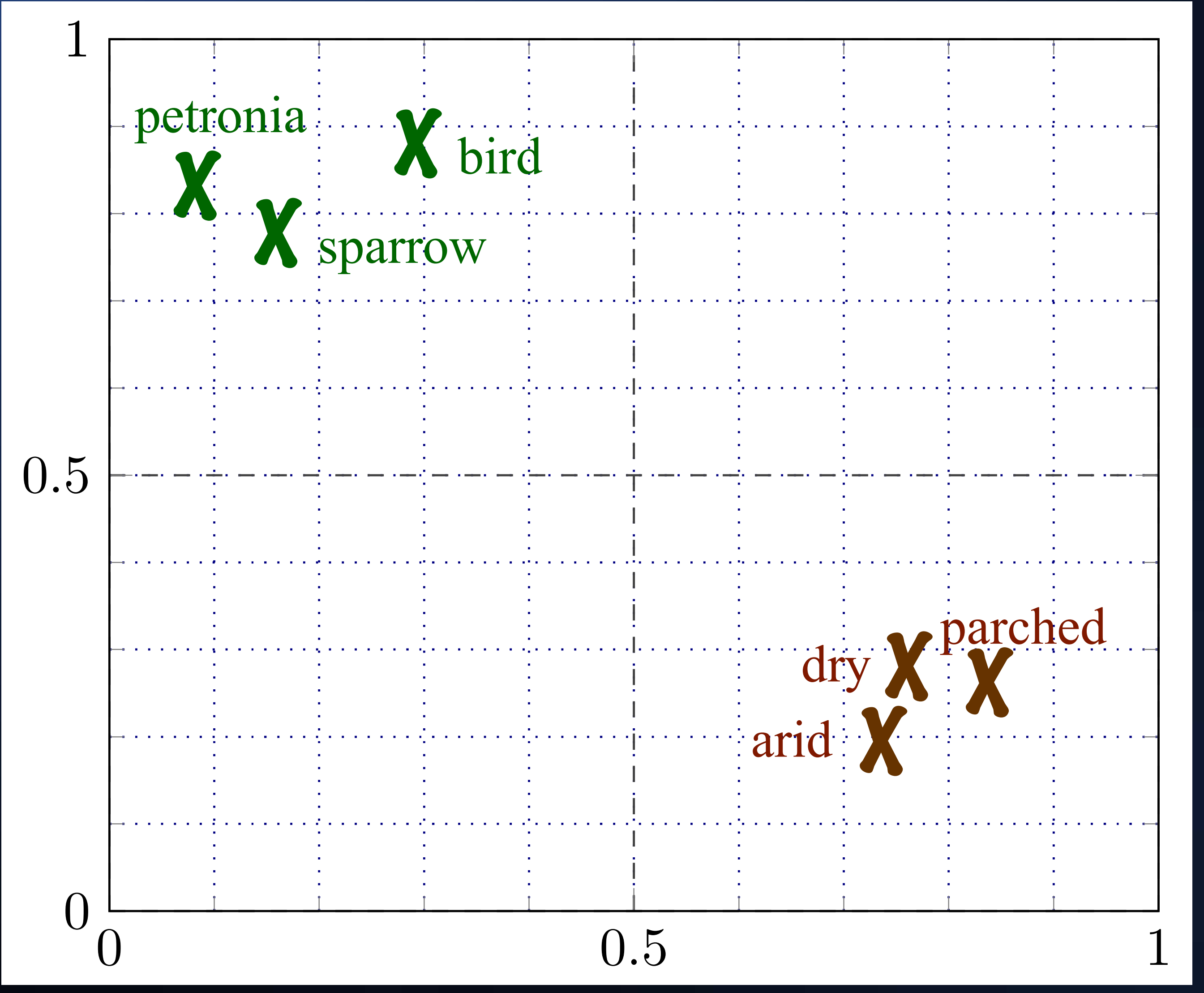


Word Vector Representations: word2vec



Gerard de Melo. Wiktionary-Based Word Embeddings.

Multilingual Word Vectors

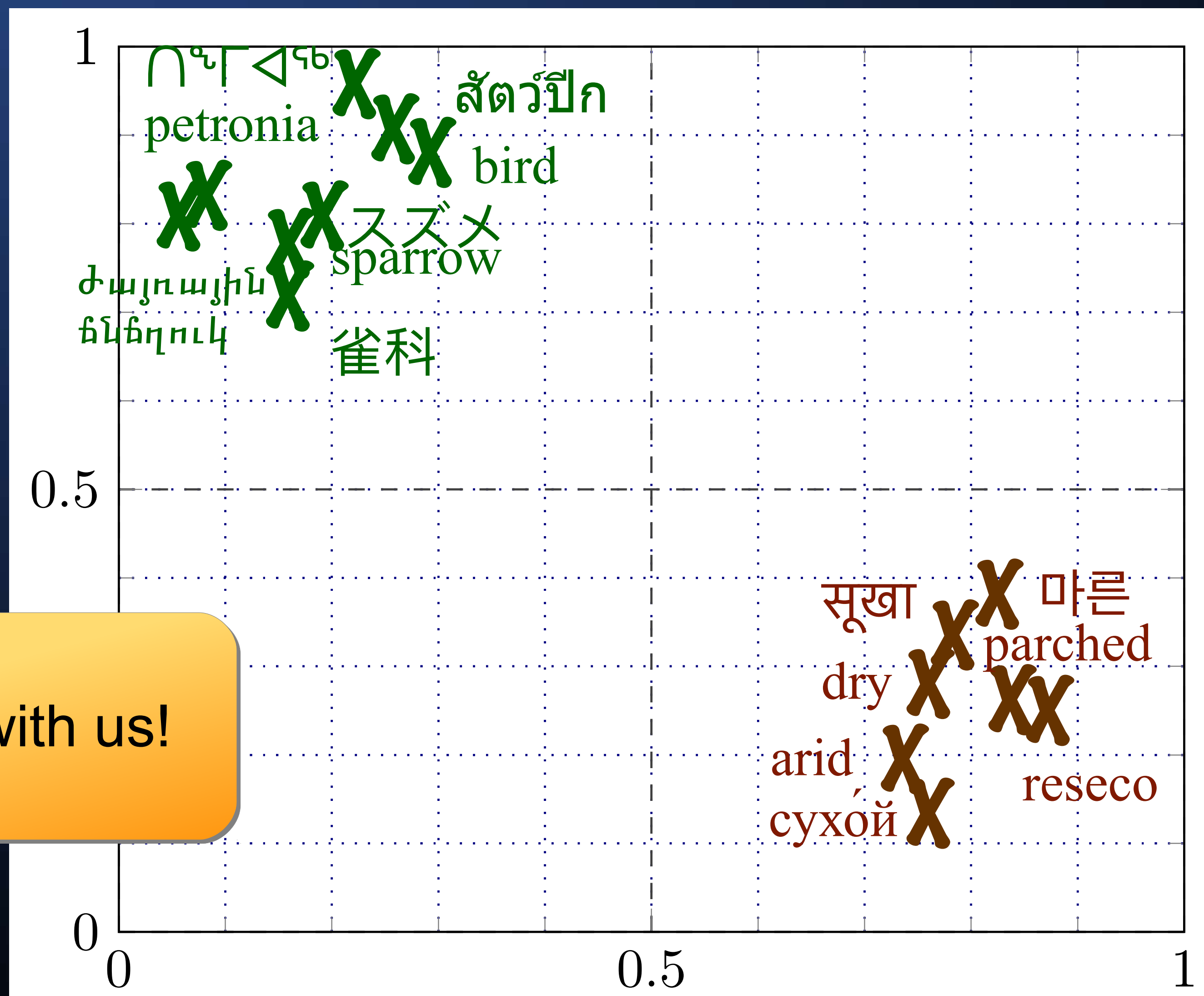


Gerard de Melo. Wiktionary-Based Word Embeddings.

Multilingual Word Vectors

5 million words, over 300 languages

Get in touch with us!



Gerard de Melo. Wiktionary-Based Word Embeddings.

Multilingual Word Vectors

Combine word2vec/GloVe Vectors
with multilingual lexical data from Wiktionary

Example: German language datasets

RG65	Chandar A P et al. (2014) En-Es Vectors	0.629 @ 55.4%
	Ours (word2vec)	0.834 @ 100.0%
	Ours (Glove)	0.809 @ 100.0%
MC30	Chandar A P et al. (2014) En-Es Vectors	0.430 @ 60.0%
	Faruqui et al. (2015)	0.591 @ N/A
	Ours (word2vec)	0.811 @ 76.7%
	Ours (Glove)	0.848 @ 76.7%
WS353	Chandar A P et al. (2014) En-Es Vectors	0.256 @ 65.1%
	Ours (word2vec)	0.548 @ 65.6%
	Ours (Glove)	0.591 @ 65.6%

Gerard de Melo. Wiktionary-Based Word Embeddings.

Multilingual Word Vectors

English–German RG65	Chandar A P et al. (2014) En-De Vectors	0.441 @ 38.4%
	Ours (word2vec)	0.812 @ 97.6%
	Ours (GloVe)	0.828 @ 97.6%
English–Spanish RG65	Chandar A P et al. (2014) En-Es Vectors	0.588 @ 59.5%
	Ours (word2vec)	0.869 @ 100.0%
	Ours (GloVe)	0.863 @ 100.0%
English–French RG65	Chandar A P et al. (2014) En-Fr Vectors	0.598 @ 52.0%
	Ours (word2vec)	0.864 @ 100.0%
	Ours (GloVe)	0.855 @ 100.0%
English–Spanish MC30	Chandar A P et al. (2014) En-Es Vectors	0.351 @ 70.0%
	Ours (word2vec)	0.745 @ 90.0%
	Ours (GloVe)	0.797 @ 90.0%
Spanish–English MC30	Chandar A P et al. (2014) En-Es Vectors	0.645 @ 56.7%
	Ours (word2vec)	0.713 @ 83.3%
	Ours (GloVe)	0.721 @ 83.3%
English–Spanish WS353	Chandar A P et al. (2014) En-Es Vectors	0.303 @ 75.9%
	Ours (word2vec)	0.582 @ 79.8%
	Ours (GloVe)	0.641 @ 79.8%
Spanish–English WS353	Chandar A P et al. (2014) En-Es Vectors	0.299 @ 73.3%
	Ours (word2vec)	0.550 @ 78.7%
	Ours (GloVe)	0.612 @ 78.7%

Gerard de Melo. Wiktionary-Based Word Embeddings.

Multilingual Word Choice Quizzes

gourmet	dale	brace
a) enjoys cooking	a) plain	a) to scream
b) has indigestion	b) retreat	b) prepare for danger
c) has an expert appreciation of food	c) shelter	c) hold your breath
d) is hungry	d) valley	d) close your eyes

Word Meanings

https://en.wikipedia.org/wiki/File:Zoiseau_la_vierge1.JPG

Word Meanings

Monarch

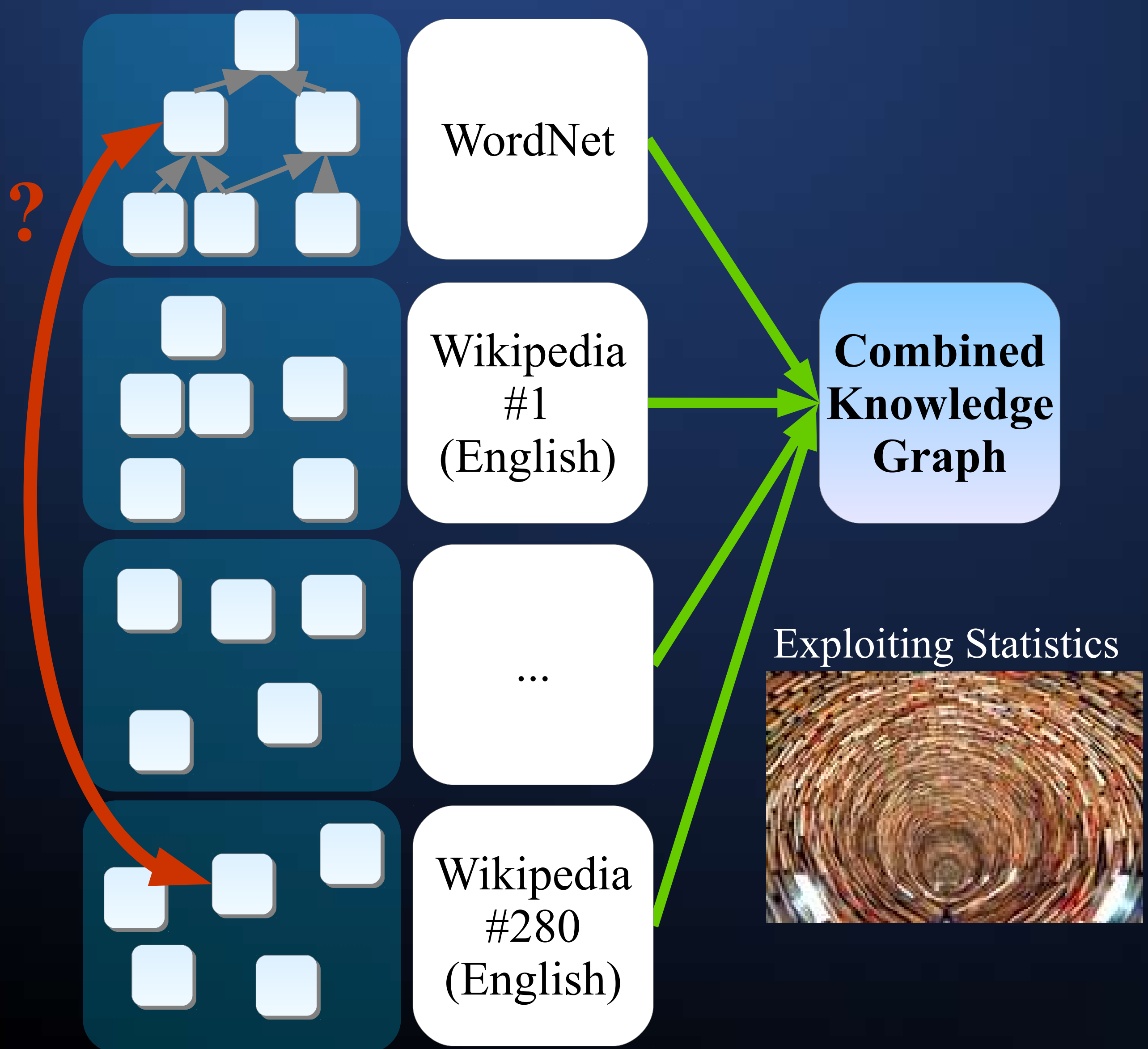
Image from:
Elizabeth: The Golden Age (2007)

Word Meanings

Kiwi

Word Meanings

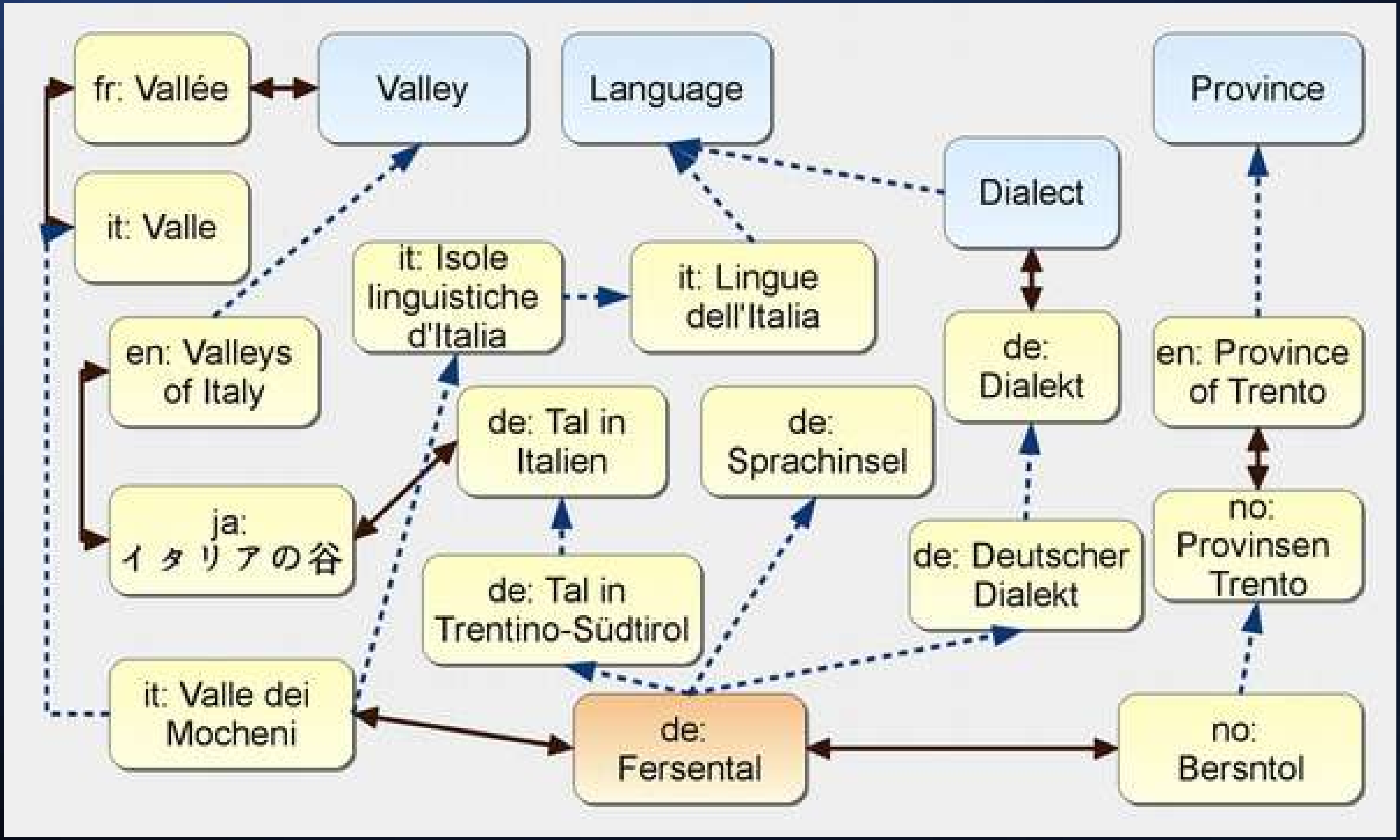
Connecting Multiple Sources



Gerard de Melo, Gerhard Weikum. Taxonomic Data Integration from Multilingual Wikipedia Editions

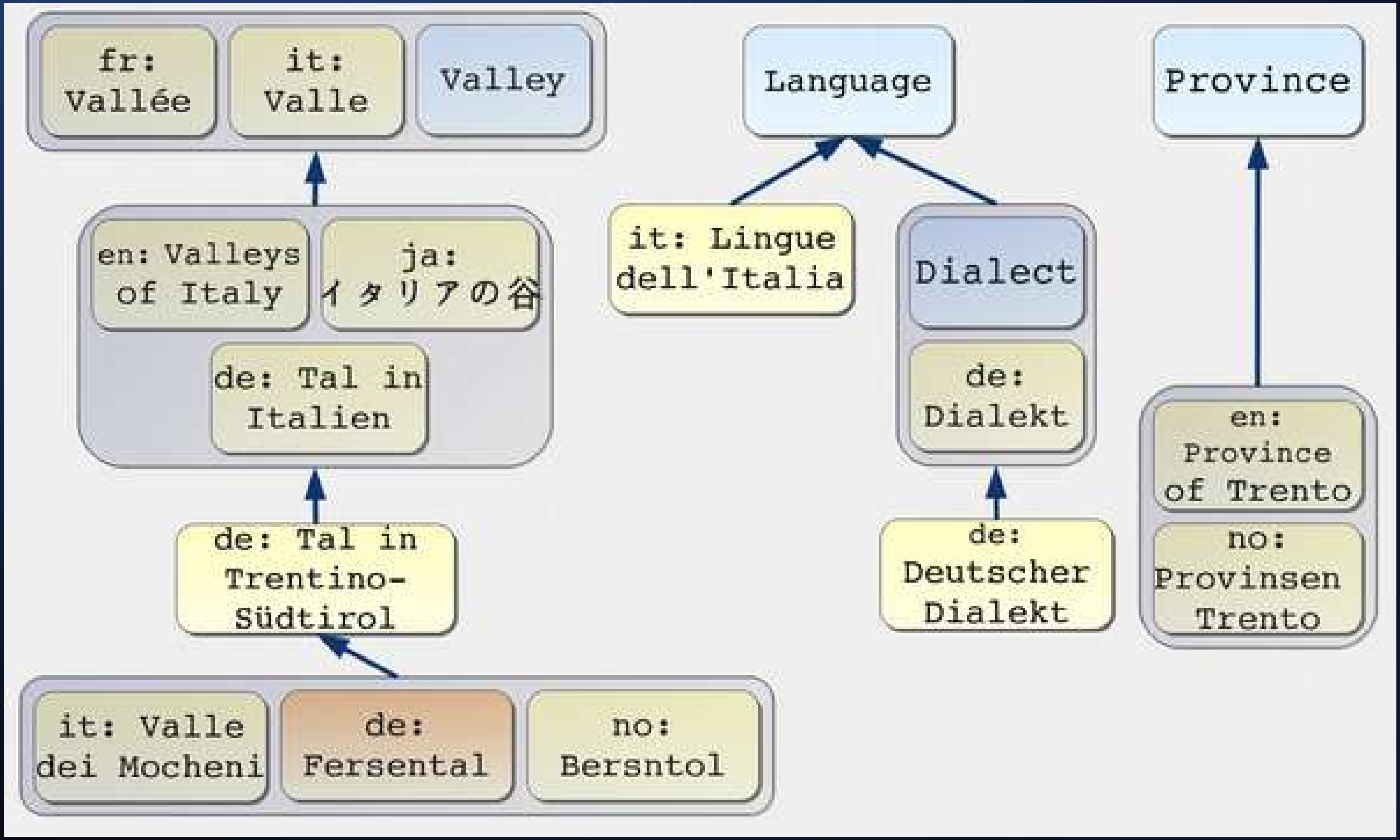
Extraction from Wikipedia

MENTA
Integration Approach



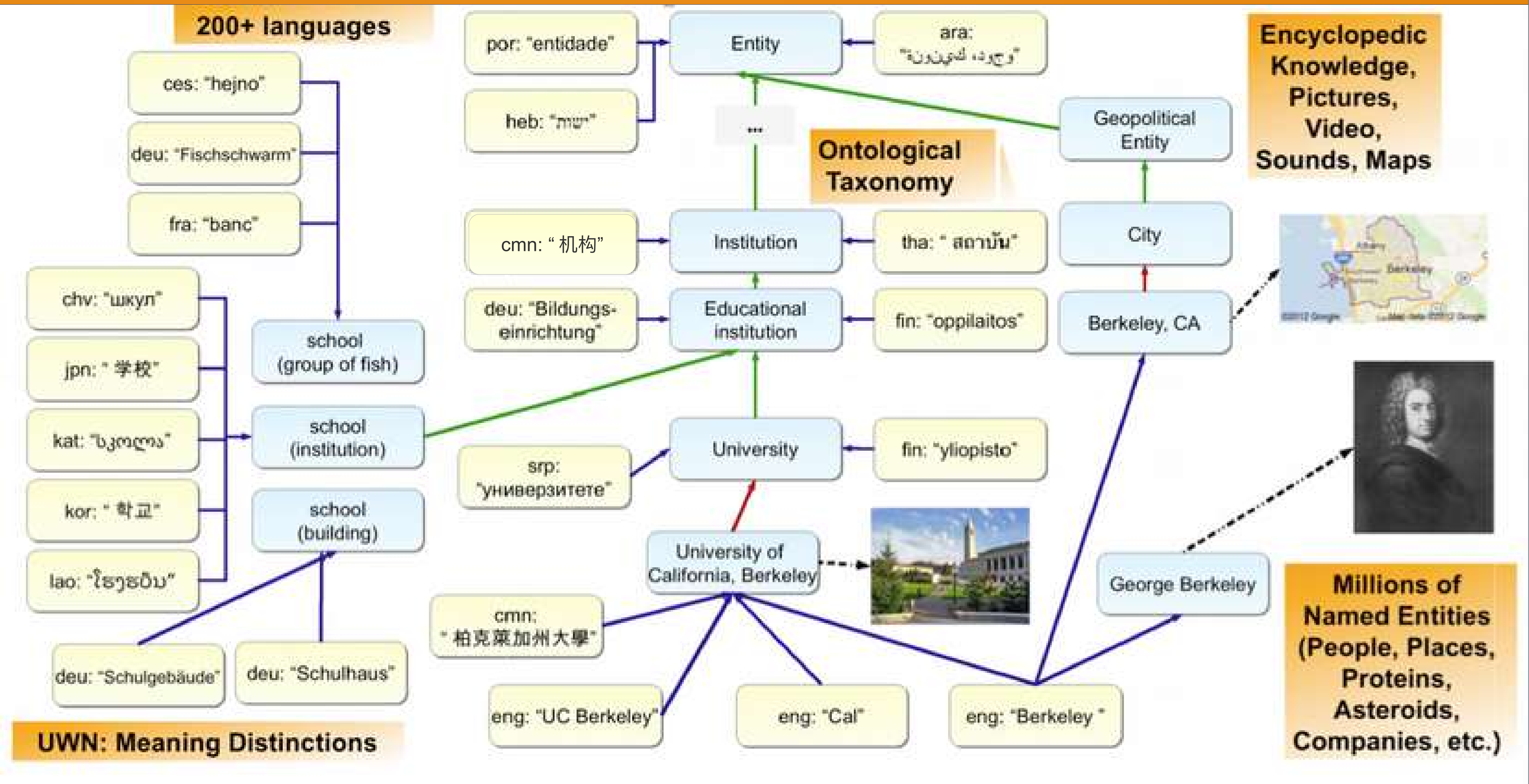
Gerard de Melo, Gerhard Weikum. Taxonomic Data Integration from Multilingual Wikipedia Editions

MENTA Integration Approach



Gerard de Melo, Gerhard Weikum. Taxonomic Data Integration from Multilingual Wikipedia Editions

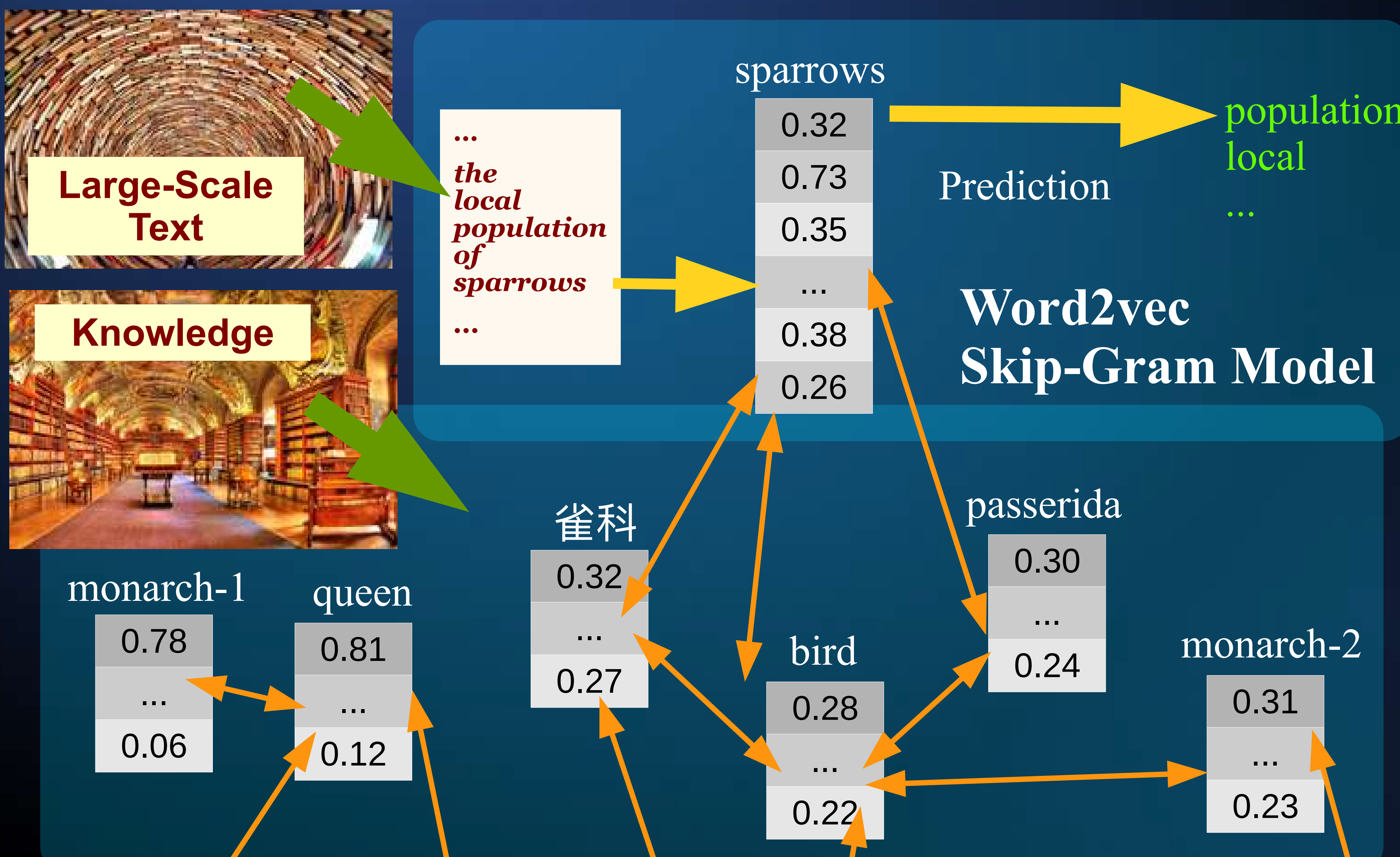
Universal Wordnet



http://lexvo.org/uwn/

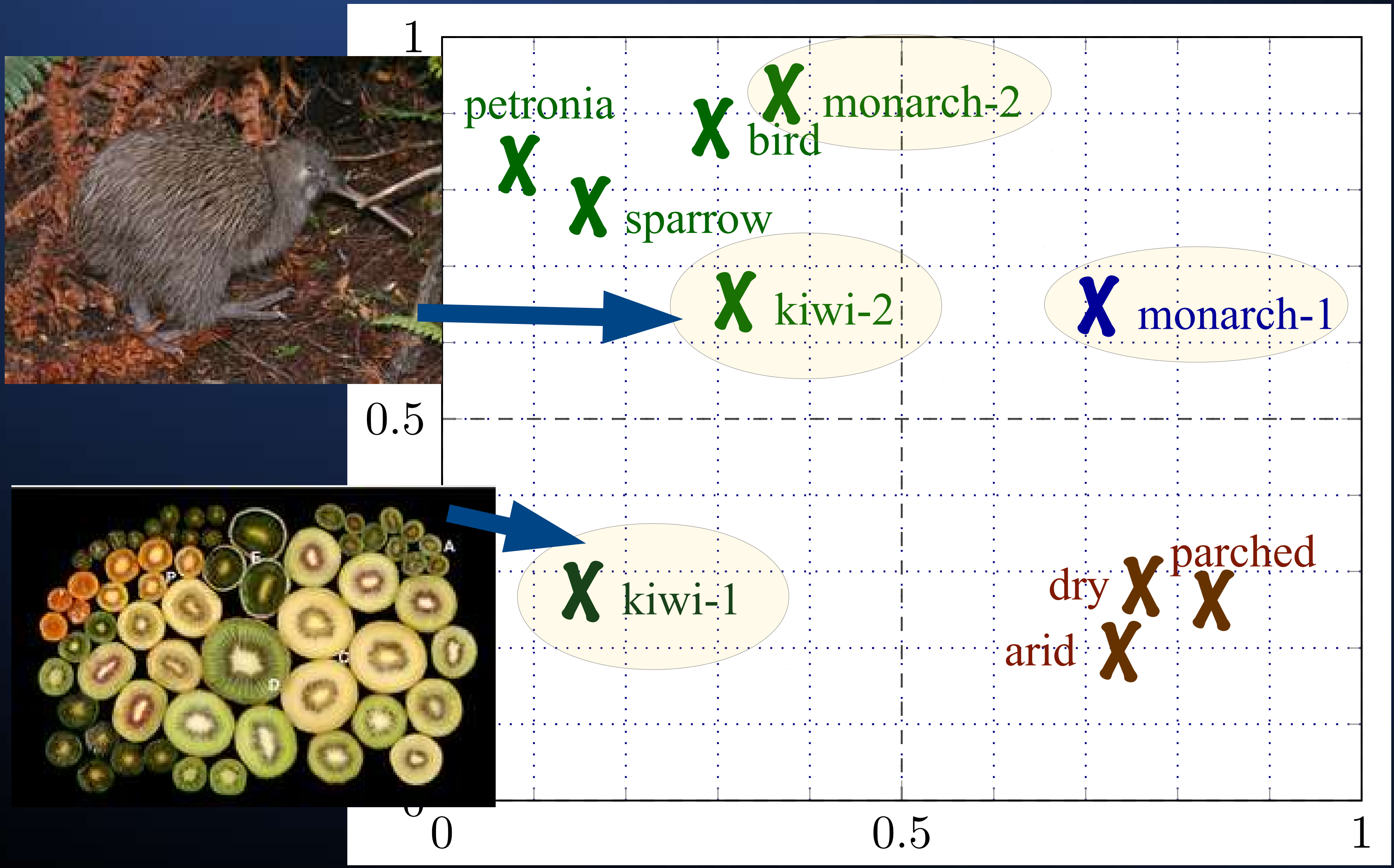
Gerard de Melo. Towards a Universal Wordnet by Learning from Combined Evidence.

Word Vector Representations: word2vec



Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Sense Representations



Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Application: Verbal Questions in IQ Tests

Identify two words (one from each set of brackets) that form a connection (analogy) when paired with the words in capitals.

CHEMISTRY (laboratory, reaction, substances)
FAUNA (plants, animals, countryside)

Answer: substances, animals

Multilingual Analogical Reasoning

Coffee is to Starbucks
as ... is to Lipton?

System: Tea

Source vectors: GLOVE

Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Ongoing Research: Multilingual Analogical Reasoning

Apple is to Macbook
as ... is to Thinkpad?

System: 1. IBM 2. Compaq 3. Lenovo

Source vectors: GLOVE

Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Ongoing Research: Multilingual Analogical Reasoning

m is to meters
as ... is to kilograms?

System: 1. kilos 2. kg 3. kilogrammes

Source vectors: GLOVE

Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Ongoing Research: Multilingual Analogical Reasoning

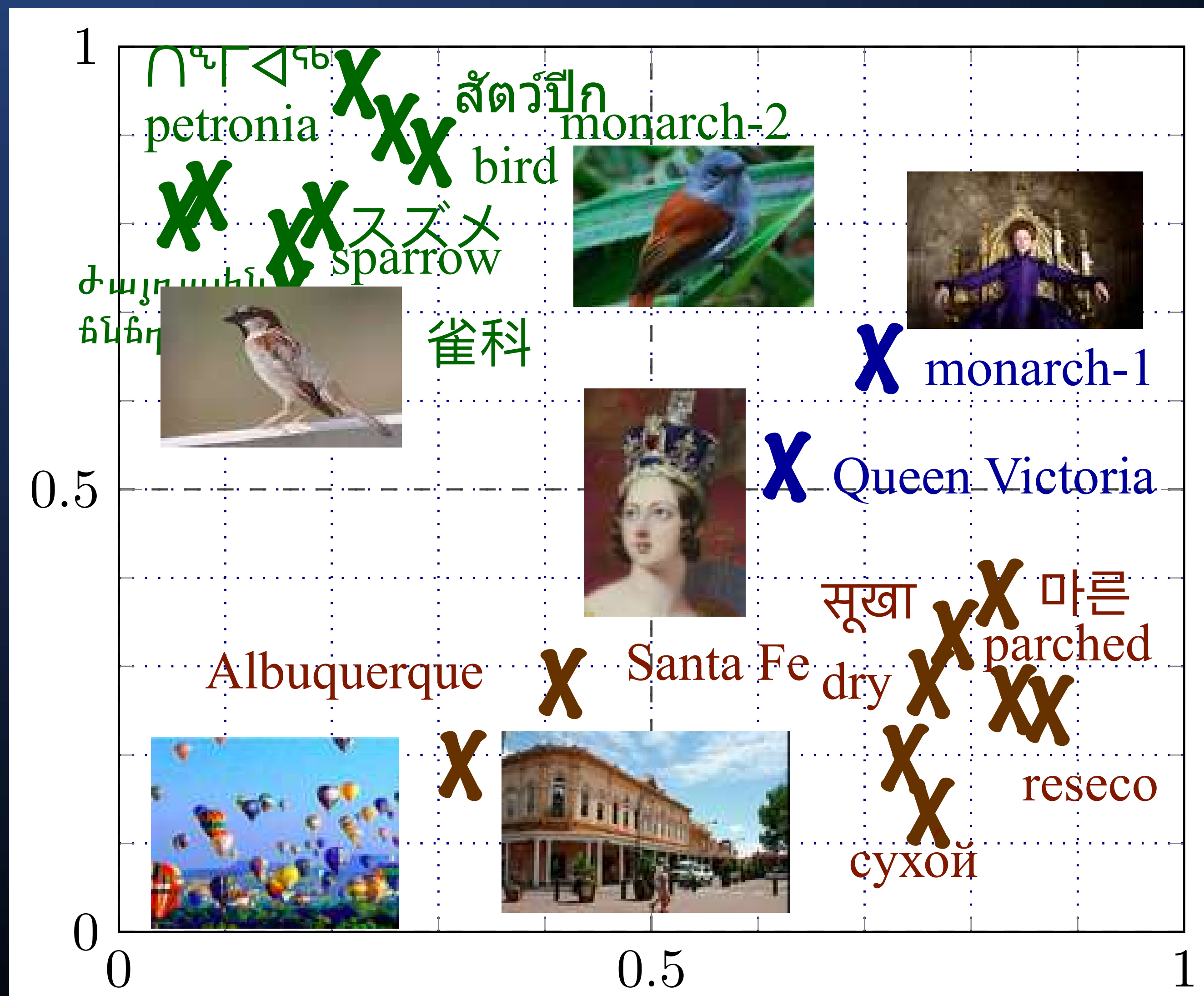
Germany is to Merkel
as ... is to Putin?

System: 1. Russia 2. Moscow 3. Europe

Source vectors: GLOVE

Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Multilingual and Multimodal Word Vectors



Gerard de Melo (2017). Inducing Conceptual Embedding Spaces from Wikipedia

Images: https://commons.wikimedia.org/wiki/File:Albuquerque_Infobox_Photo.png https://commons.wikimedia.org/wiki/File:Santa_Fe,_New_Mexico_Montage_1.png https://commons.wikimedia.org/wiki/File:House_Sparrow_mar08.jpg https://en.wikipedia.org/wiki/Queen_Victoria#/media/File:Dronning_victoria.jpg etc.

Questions?

Image: https://www.flickr.com/photos/opensourceway/5556249000