Semantic Representations of Word Senses and Concepts

Sapienza Università di Roma

José Camacho Collados
Ignacio Iacobacci
Roberto Navigli

University of Cambridge

Mohammad Taher Pilehvar

54 ACL 2016
AUGUST 7 – 12 | BERLIN, GERMANY

Outline

- **Foundations**
- **Sense representations**
    - Introduction
    - **Knowledge-based techniques**
        - WordNet
        - Large knowledge resources
            - Wikipedia
            - BabelNet
            - FreeBase - WikiData

Coffee Break (30 mins)

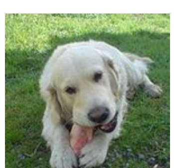
- **Unsupervised techniques**
- Advantages and limitations
- **Applications**
- **Open problems and future work**

Key points

- **What** do we want to **represent**?
- What does "**semantic representation**" **mean**?
- **Why** semantic representations?
- What **problems** affect mainstream representations?
- How to **address** these problems?
- What comes **next**?

What do we want to represent?

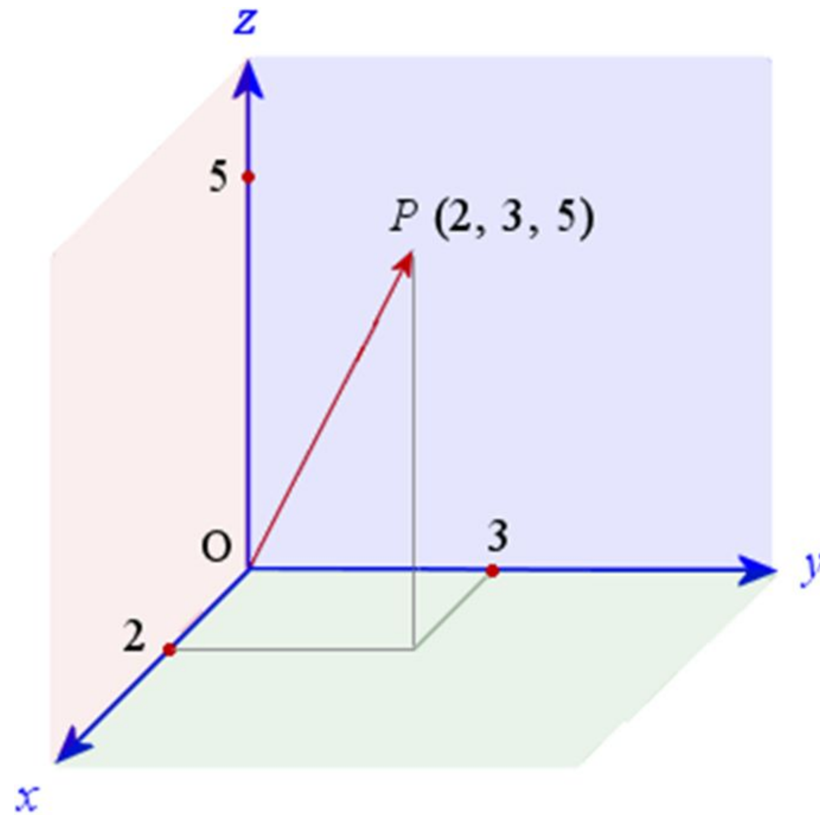
Linguistic items of different kinds:

- **Documents**: the Wikipedia page for "On the Internet, nobody knows you're a dog"
- **Sentences**: On the Internet, nobody knows you're a dog
- **Phrases**: on the Internet
- **Words**: dog
- **Senses**: 

"On the Internet, nobody knows you're a dog."

What kinds of representation can we provide?

Vector representations (see Turney and Pantel, 2010 for a survey)



Vector space models

Words are represented as vectors

- Semantically similar words are close in the space

Term-document matrix

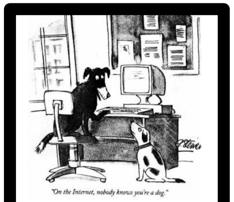
- Useful if you have a set collection of documents
- Rows are words, columns are documents
- For instance, Wikipedia documents and the terms occurring in it:

$$
\begin{array}{l}
\texttt{dog} \\
\texttt{Internet} \\
\texttt{cartoon} \\
\texttt{cartoonist}
\end{array}
\begin{pmatrix}
3 & 10 & 0 & 0 & 0 \\
4 & 0 & 1 & 0 & 7 \\
1 & 0 & 5 & 0 & 0 \\
1 & 0 & 10 & 0 & 0
\end{pmatrix}
$$

Term-document matrix

- Useful if you have a set collection of documents
- Rows are words, columns are documents
- For instance, Wikipedia documents and the terms occurring in it:





dog	3	10	0	0	0
Internet	4	0	1	0	7
cartoon	1	0	5	0	0
cartoonist	1	0	10	0	0

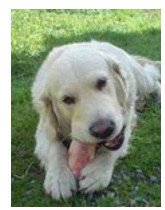
term

count of term occurrences within the document

document

Term-document matrix

- Useful if you have a set collection of documents
- Rows are words, columns are documents
- For instance, Wikipedia documents and the terms occurring in it:

dog	3	10	0	0	0
Internet	4	0	1	0	7
cartoon	1	0	5	0	0
cartoonist	1	0	10	0	0

A word is defined by a **vector of counts over documents**

Generalization: the word-context matrix

A document might not be the best item for measuring word similarity

What is the optimal granularity of "context" to measure the similarity between words?

– n-gram, sentence, grammatical relations, paragraph, document, etc.

Distributional hypothesis (Harris, 1954): words occurring in similar context tend to have similar meanings

A word is defined by a vector of counts over documents

Extract and **count** the cooccurrences in a corpus

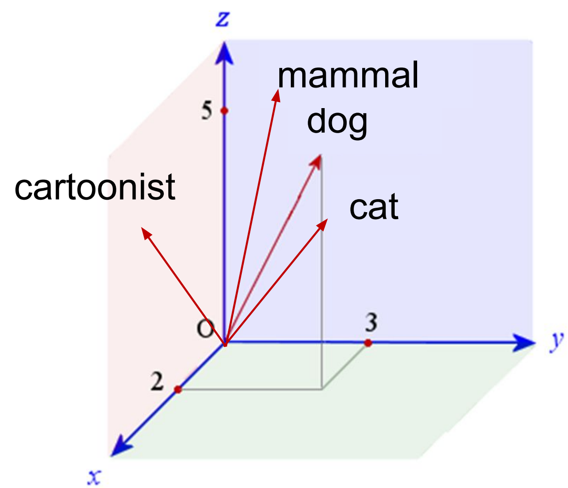
brown fox jumps over the lazy	dog	". Changing the numbers will
near his feet, is a sleeping pet	dog	. This effigy seems from the bearings
feet is an animal, probably a	dog	, and the hands are joined in the
again! The car is regarded as	dog	's property, Sue gets her bottom
only superstar who looks like a	dog	. Oh, and a final thought about
you have to do to bring your pet	dog	, cat or ferret into (or back
burying a living child, a calf, a	dog	, goat, or lambthe lamb slain
fly tipping / litter enforcement	dog	warden service dog fouling gypsies
enforcement dog warden service	dog	fouling gypsies and travellers
really coach a man like you would a	dog	? Or is Katie about to learn that

Distributional semantics

Resulting in a **cooccurrence vector**, e.g.:

	animal,	bark,	hair,	cat,	eat,	feed,	…,	train	
dog = (	10 ,	25 ,	3 ,	5 ,	7 ,	8 ,	…,	5	)

Dog is expected to be more similar to other mammals than to, e.g., cartoonist



A word is defined by a **vector of counts over contexts**

What values should we use for non-zero components?

Beyond raw counts, we can calculate functions of term frequency, cooccurrence, frequency in topics, etc.

For term-document matrices, we can use TF-IDF:

$$TF-IDF(t,d) = \frac{f_{t,d}}{|d|} * log \frac{|D|}{|\{d: t \in d\}|}$$

or lexical specificity (less sensitive to document lengths)

What values should we use for non-zero components?

For term-context matrices, we can use

- Dice:

$$Dice(w, w') = \frac{2c(w, w')}{c(w) + c(w')}$$

- Pointwise Mutual Information: $PMI(w, w') = log\frac{P(w, w')}{P(w)P(w')}$

- Positive Pointwise Mutual Information:

$$PPMI(w, w') = \begin{cases} PMI(w, w') & if\ PMI(w, w') > 0 \\ 0 & else \end{cases}$$

(However, biased towards infrequent words)

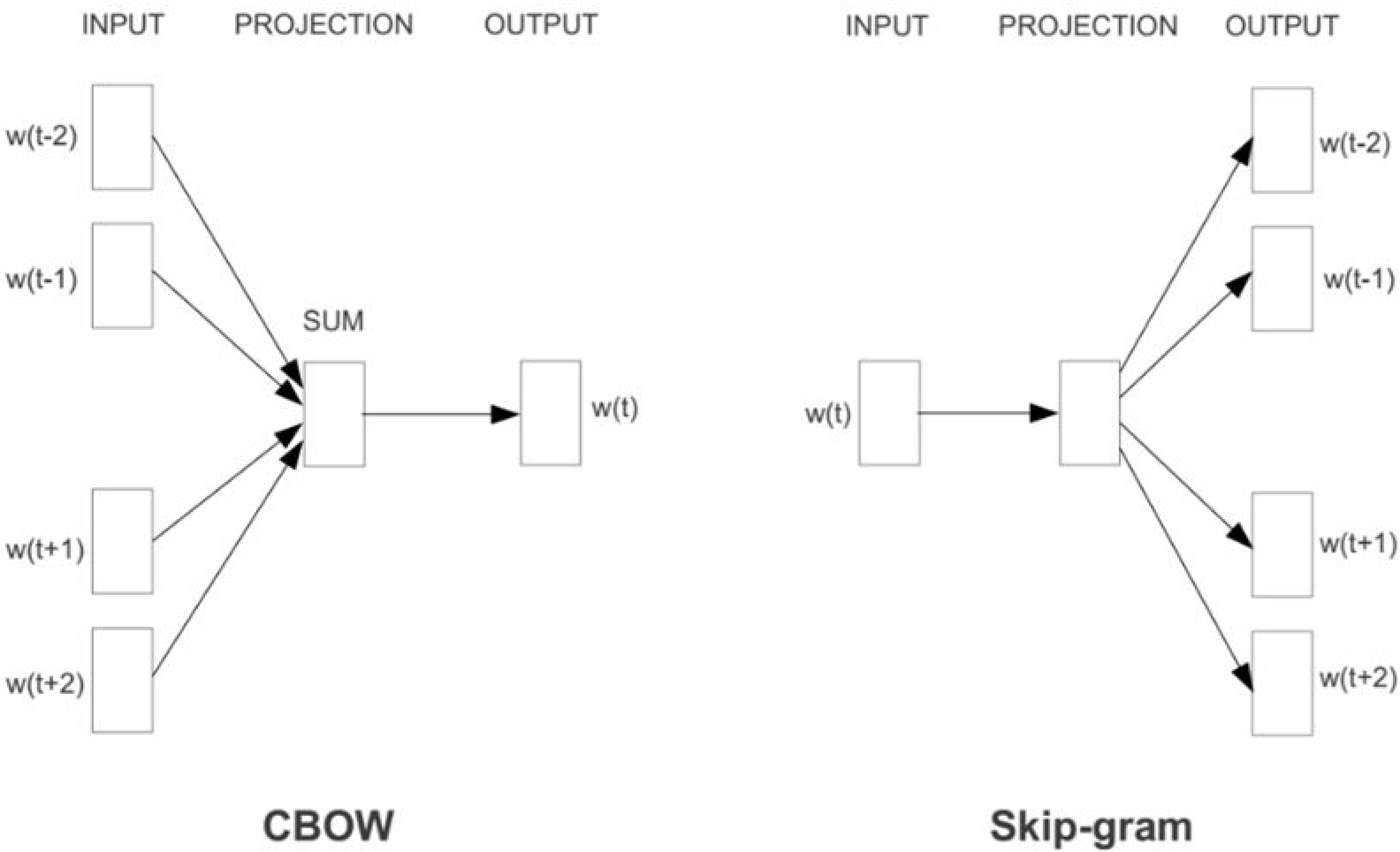
Comparing word representations

- Parametric
  - Cosine $$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$
  - Tanimoto similarity $$f(A,B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$
  - Kullback–Leibler (KL) divergence $$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$
  - Jensen–Shannon (JS) divergence $$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M) \quad \text{where } M = \frac{1}{2}(P+Q)$$
- Non-parametric
  - Rank-Biased Overlap $$RBO\ (\mathcal{S}_1, \mathcal{S}_2) = (1-p) \sum_{d=1}^{|H|} p^{d-1} \frac{|H_d|}{d}$$
  - Weighted Overlap $$WO(v_1, v_2) = \frac{\sum_{q \in O} \left(rank(q, v_1) + rank(q, v_2) \right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}}$$

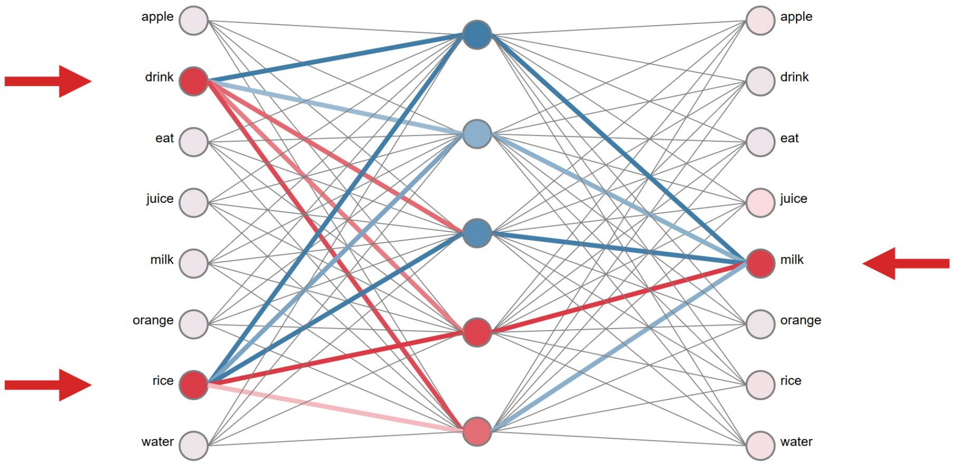
Small is good

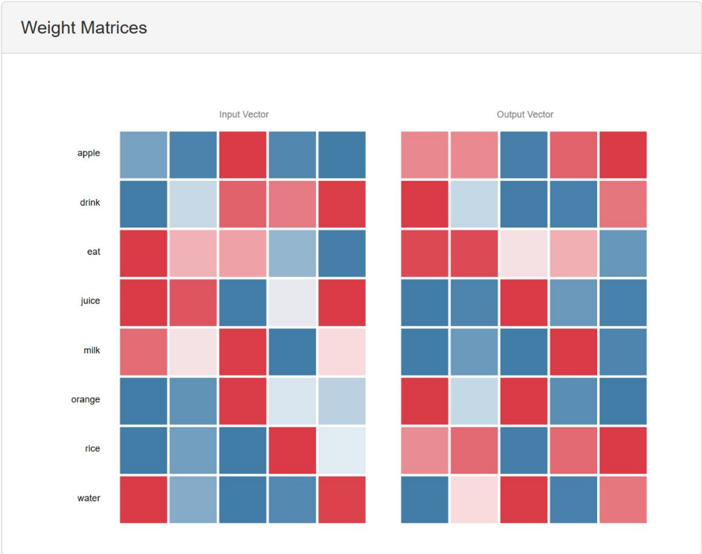
- Vectors often have thousands to millions of **dimensions**
- The dimensionality of these vectors can be **reduced** in many different ways:
  - Random indexing
  - Non-negative matrix factorization
  - Singular Value Decomposition
  - Latent Dirichlet Allocation
  - Neural Network Embeddings

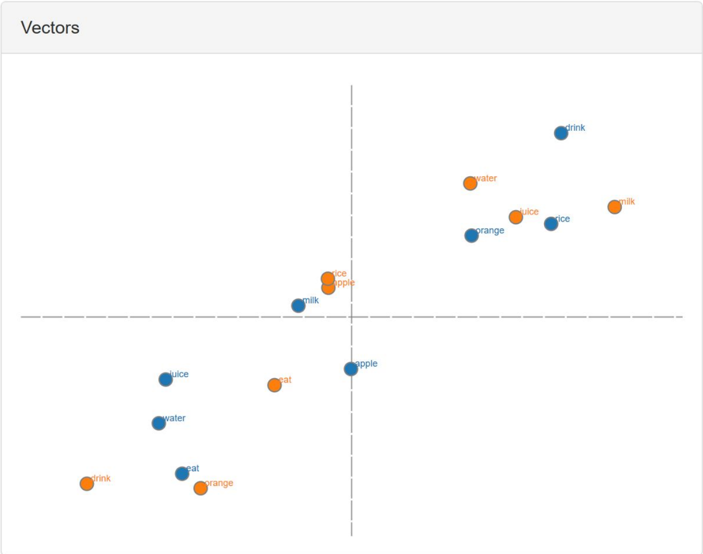
The word2vec architectures (Mikolov et al., 2013)



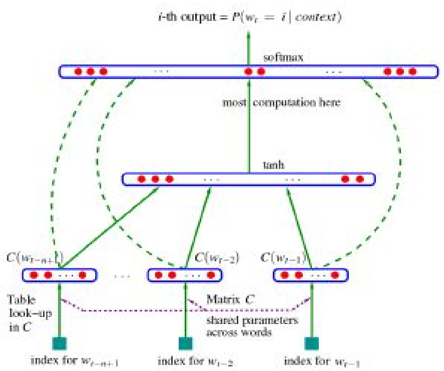
wevi: a tool for understanding word2vec [Rong, 2014]



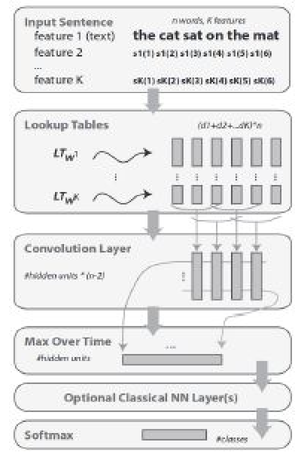




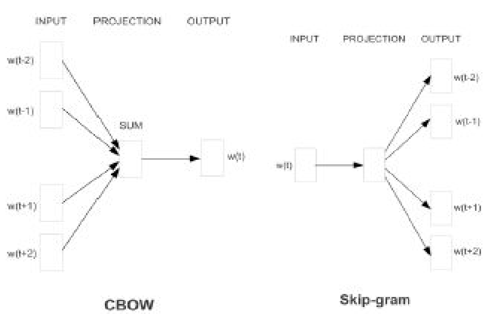
Much work on vector representations of meaning



Bengio et al. (2003)



Collobert & Weston (2008)



Mikolov et al. (2013)

Probability and Ratio	$k = solid$	$k = gas$	$k = water$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36

Pennington et al. (2014)

Why?

Embedded vector representations:

- are compact and fast to compute
- preserve important relational information between words (actually, meanings):

$$king - man + woman \approx queen$$

- are geared towards general use (word2vec, GloVe)
- are a successful example of unsupervised learning

Applications for word representations

- Semantic similarity
- Word clustering
- Word Sense Induction
- Word Sense Disambiguation and Entity Linking
- Semantic role labeling
- Plagiarism detection
- Automated essay marking
- (Open) Information extraction

The dream: machine reading

The word level is not enough

Word representations alone are **not enough** to perform a number of tasks

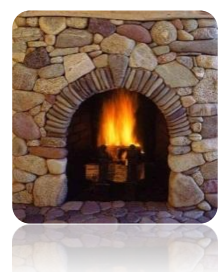
– at the sentence, paragraph and document level
– at the sense level

Let's see for example what we could do with **semantic similarity**

Semantic Similarity at different levels

Sentence Level

Word Level

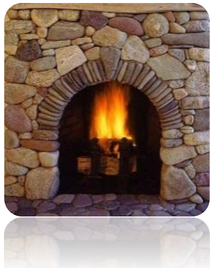


Sense Level

Semantic Similarity at different levels

Sentence Level

Word Level



Sense Level

Semantic Similarity at different levels

Word level

heater

fireplace

➢ Applications

- Lexical simplification

  (Biran et al., 2011)

  Locuacious ⟶ *Talkative*

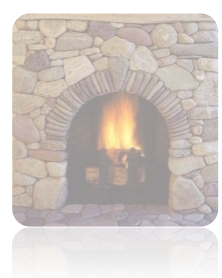
- Lexical substitution

  (McCarthy and Navigli, 2009)

Semantic Similarity at different levels

Sentence Level

Word Level



Sense Level

Semantic Similarity at different levels

Sentence level

The worker was terminated

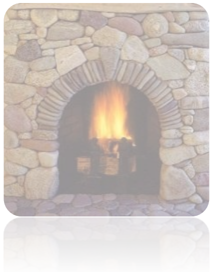
The boss fired him

- ➢ Applications
  - ➢ Paraphrase recognition (Tsatsaronis et al., 2010)
  - ➢ MT evaluation (Kauchak and Barzilay, 2006)
  - • Question Answering (Surdeanu et al., 2011)
  - • Textual Entailment (Dagan et al., 2006)

Semantic Similarity at different levels

Sentence Level

Word Level



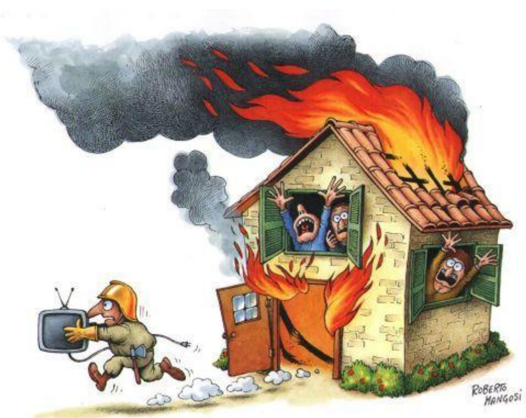
Sense Level

Semantic Similarity at different levels

Sense level

fire sense #1

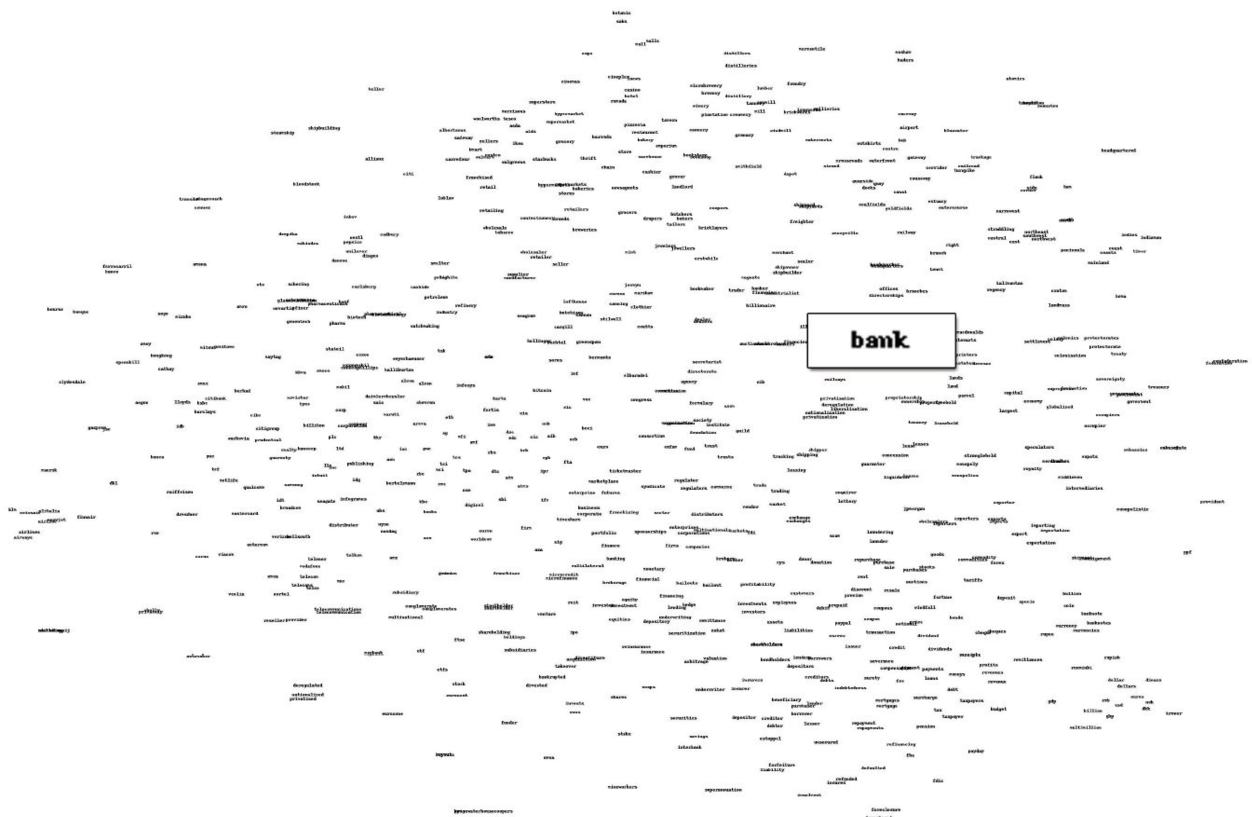
fire sense #8



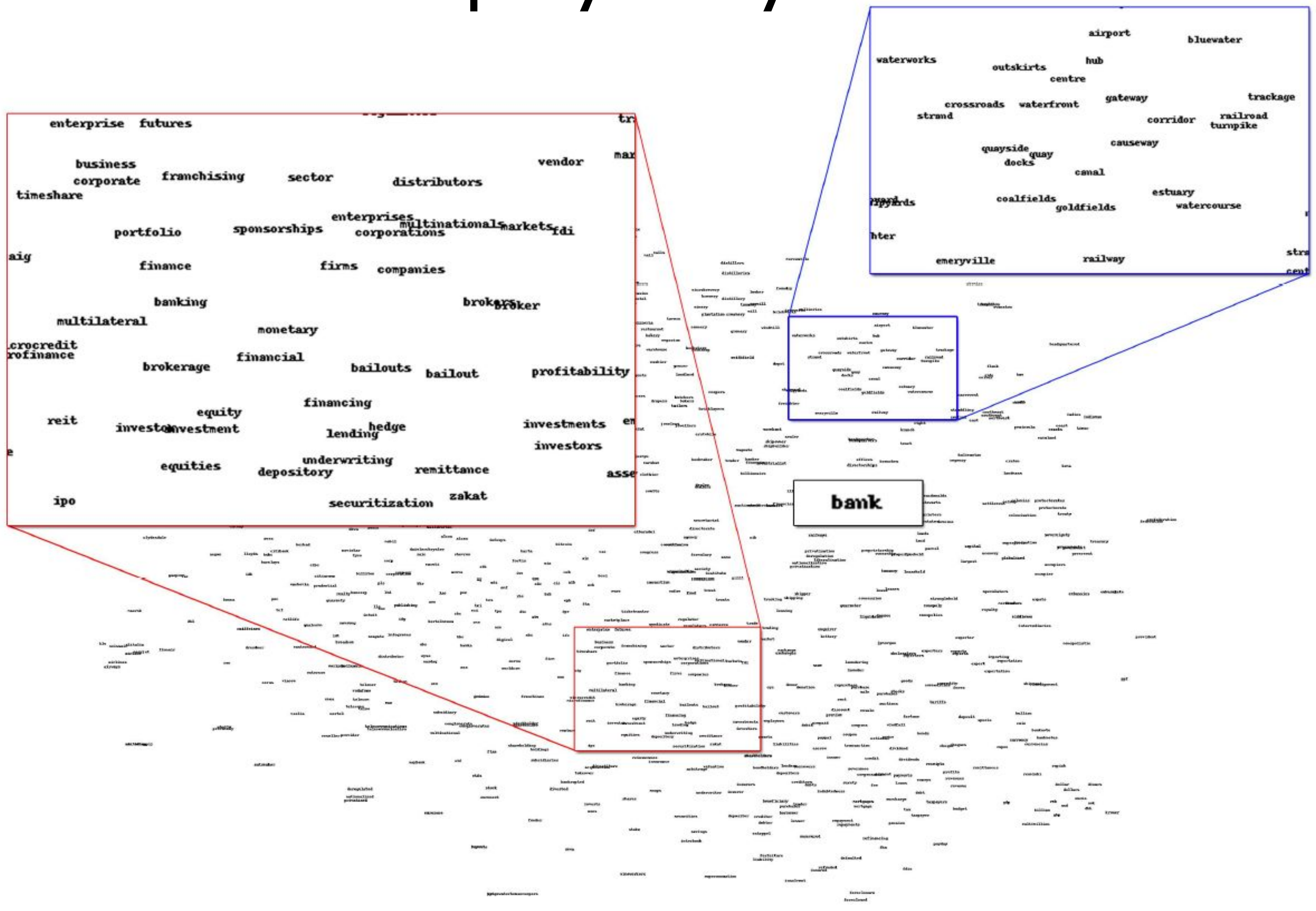
➢ Applications

- Coarsening sense inventories (Navigli, 2006; Snow et al., 2007)
- Semantic priming (Neely et al., 1989)
- Word Sense Disambiguation (Navigli, 2009)

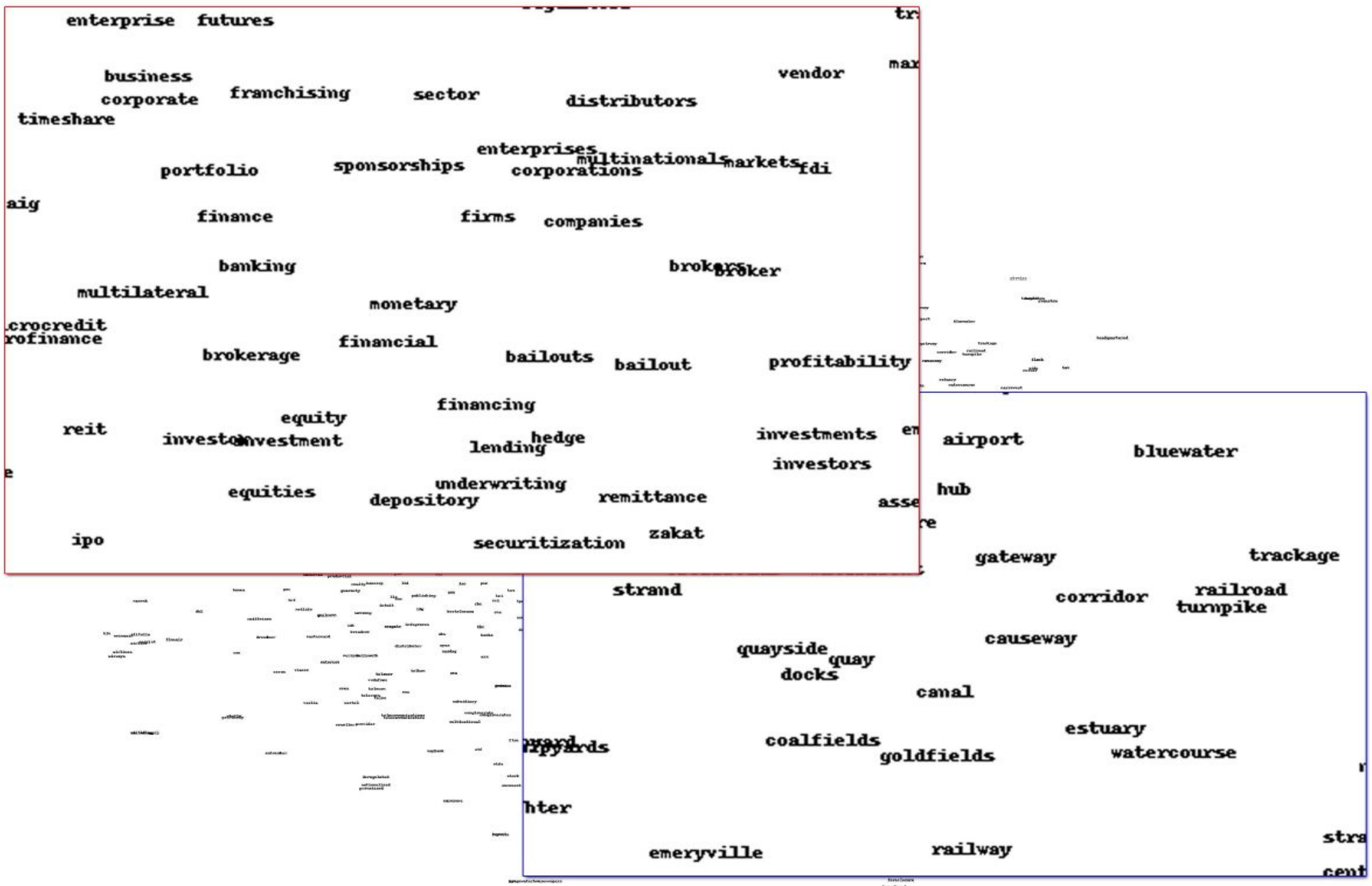
Problem 1: word representations cannot capture polysemy



Problem 1:
word representations cannot capture polysemy



Problem 1: word representations cannot capture polysemy



Word representations and the triangular inequality

Example from Neelakantan et al (2014)

For distance d, $d(a, c) \leq d(a, b) + d(b, c)$.

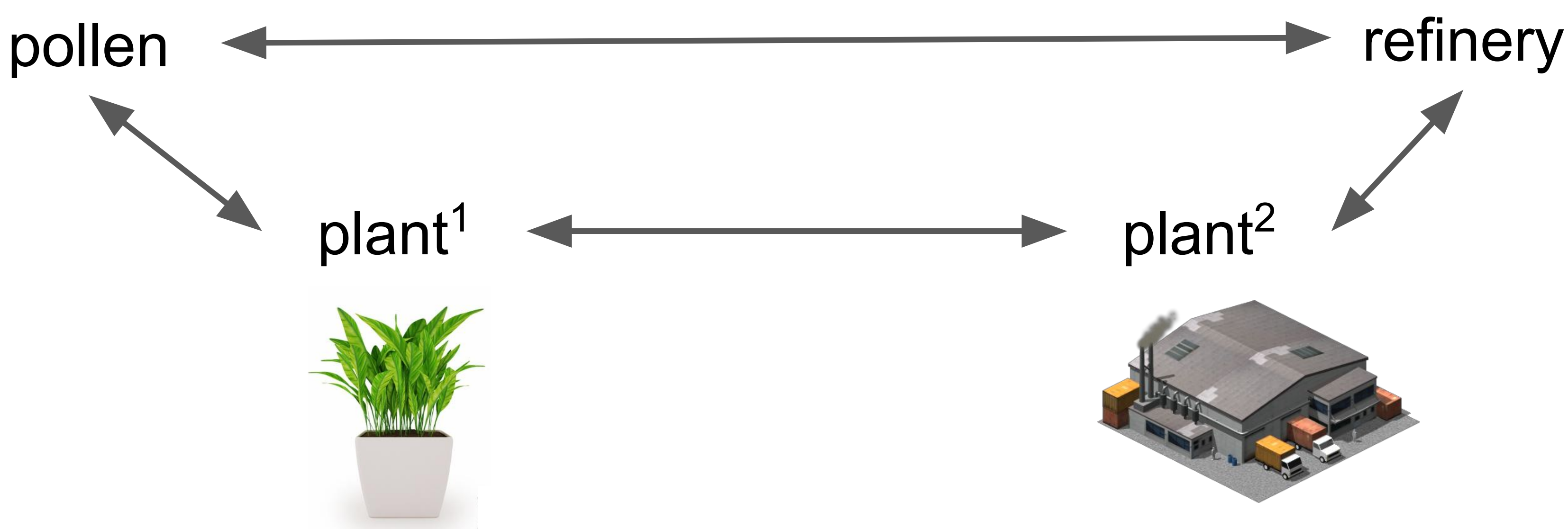
pollen ⟷ refinery

plant

Word representations and the triangular inequality

Example from Neelakantan et al (2014)

For distance d, $d(a, c) \leq d(a, b) + d(b, c)$.



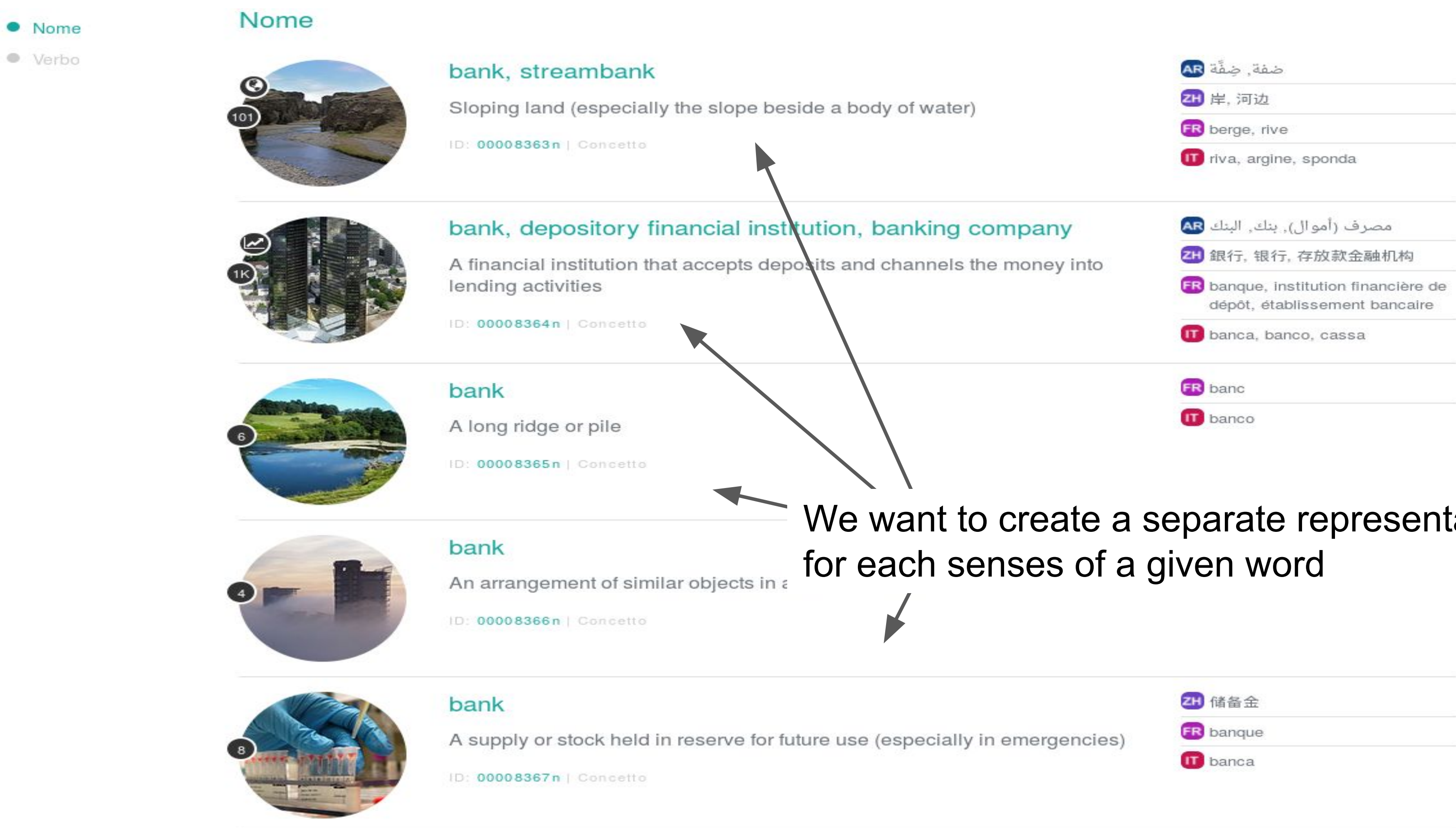
Problem 2: word representations do not take advantage of existing semantic resources

WIKIDATA

BabelNet

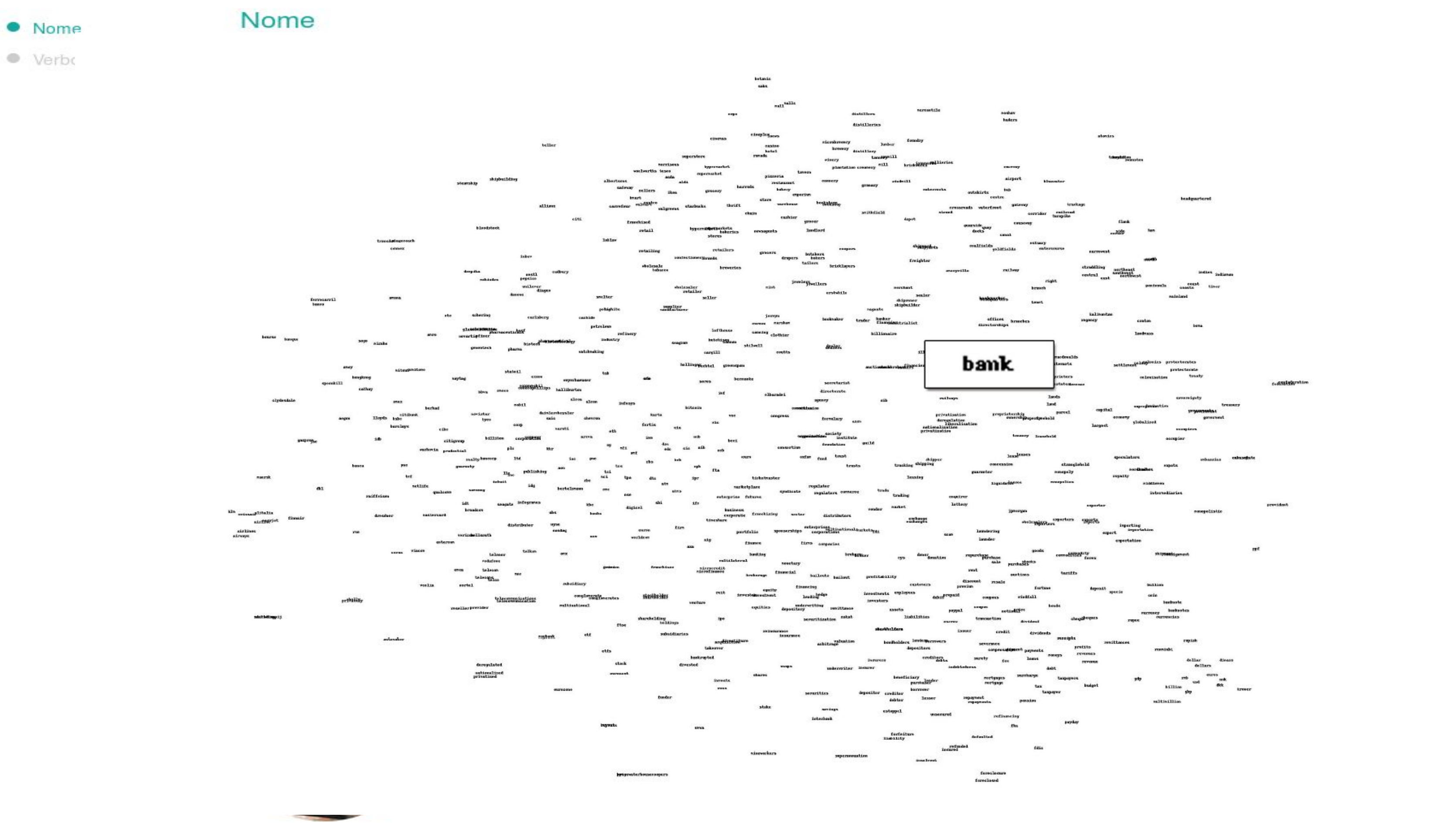
WIKIPEDIA

Example: the sense inventory of "bank" in BabelNet

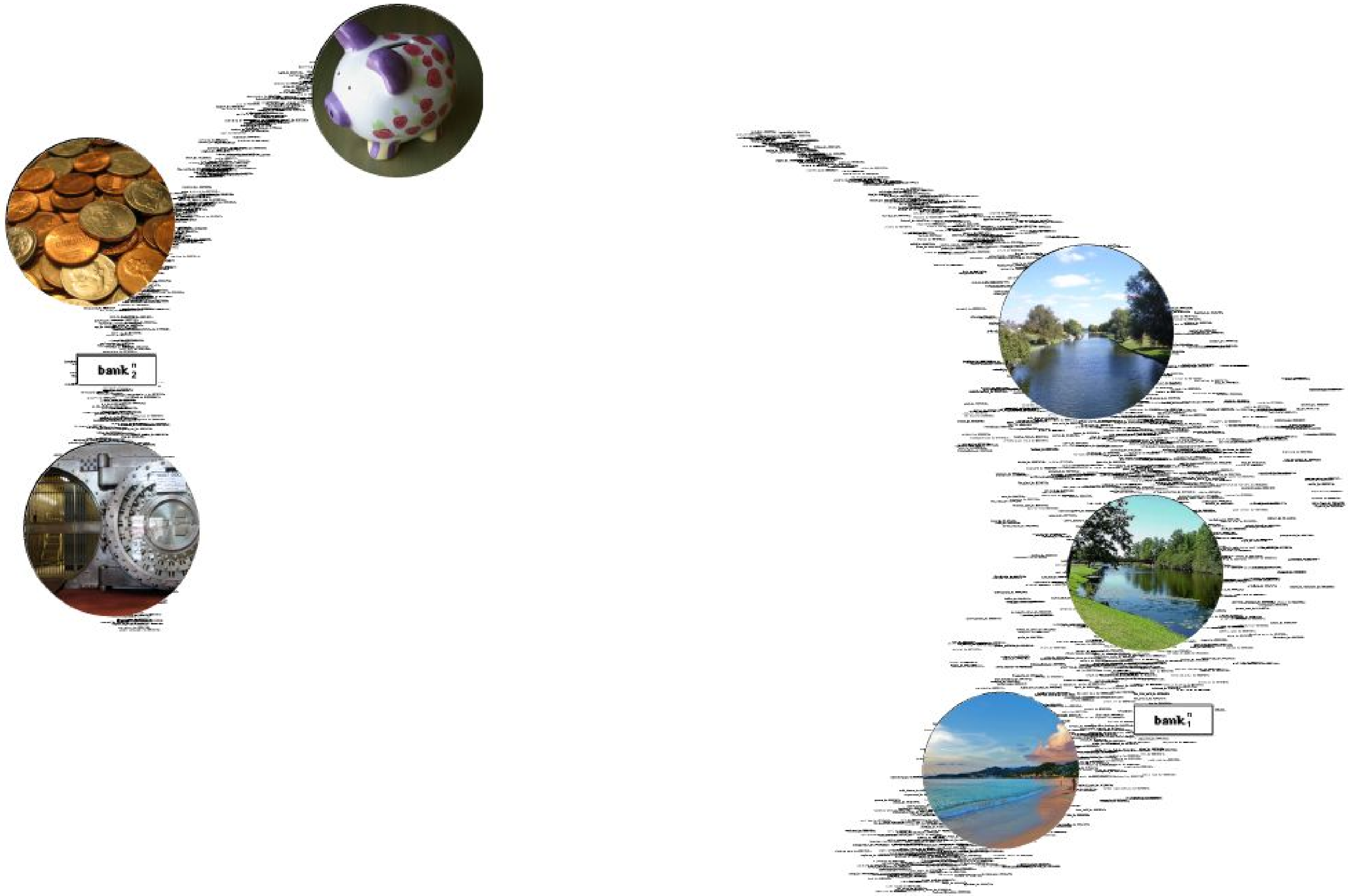


We want to create a separate representation for each senses of a given word

Key goal: obtain sense representations



Key goal: obtain sense representations



Sense Representation Techniques

Introduction

Two types of sense representation techniques

Linked to sense inventories

Knowledge-based

Wikipedia
The Free Encyclopedia

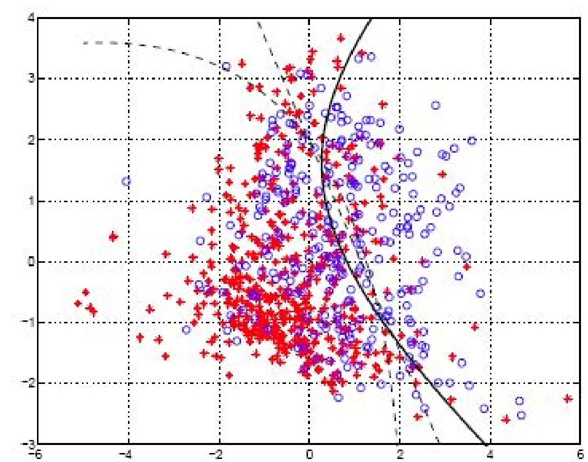
BabelNet

Freebase

Not linked

Unsupervised

(Multi prototype)



Unsupervised Sense Representations

Induce senses, then learn representations for the induced senses

Usually coupled with **clustering**

Unsupervised Sense Representations

Induce senses, then learn representations for the induced senses

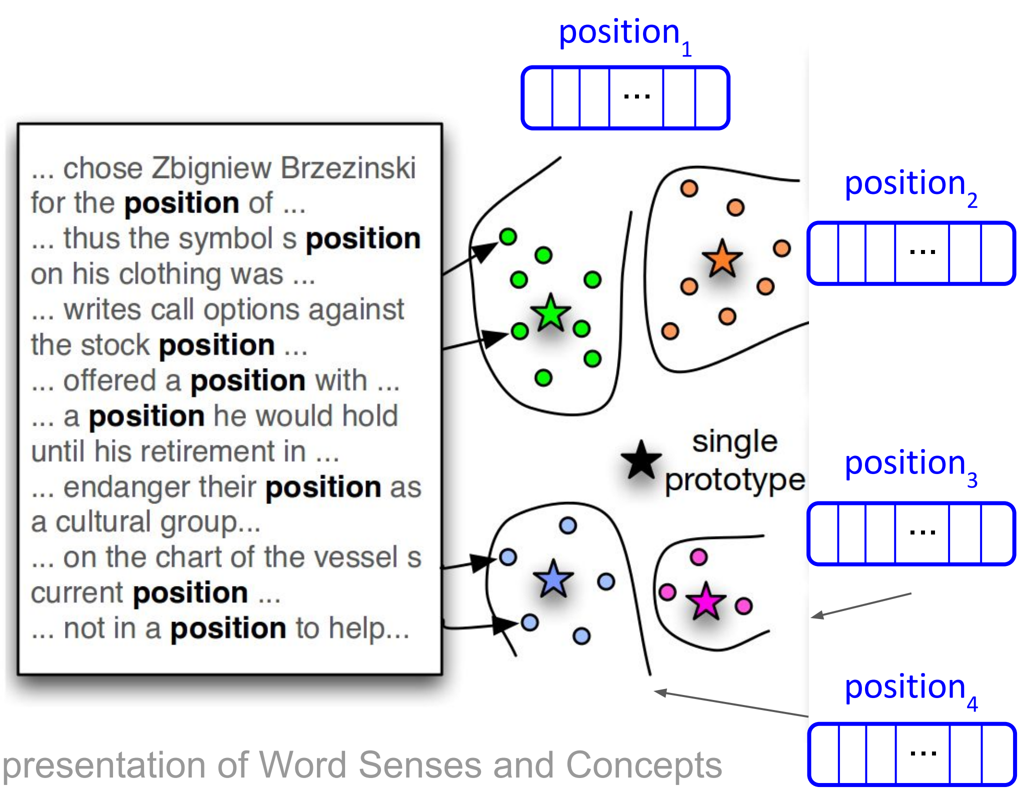


Image courtesy of Reisinger and Mooney (2010)

Unsupervised Sense Representations

Features:

- **Do not rely** on external **sense inventories**
- **Clustering** algorithms are generally used for distinguishing senses from each other
- Resulting sense representations are **not linked** to any inventory

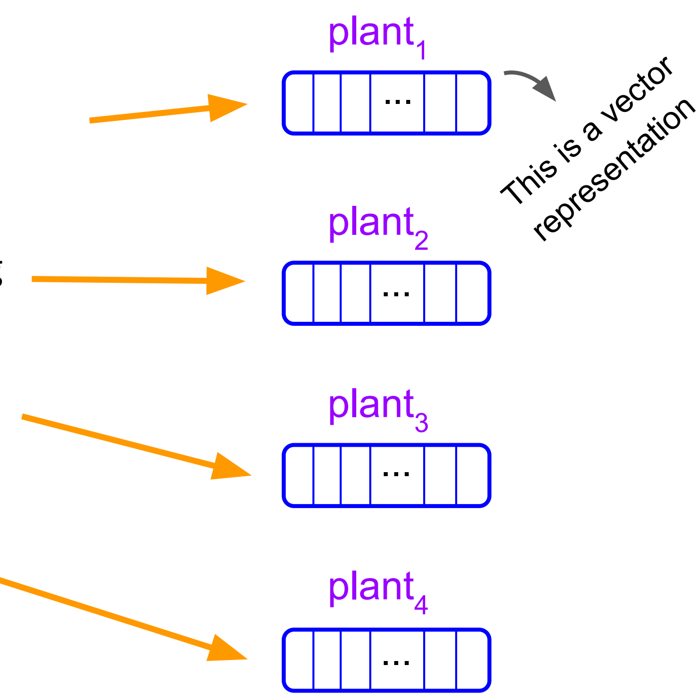
Knowledge-based Sense Representations

Knowledge-based Sense Representations

Represent word senses as defined by sense inventories

plant

- ***plant, works, industrial plant*** (buildings for carrying on industrial labor)
- ***plant, flora, plant life*** ((botany) a living organism lacking the power of locomotion)
- ***plant*** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- ***plant*** (something planted secretly for discovery by another)



Knowledge-based Sense Representations

Represent word senses as defined by sense inventories

Exploit various types of knowledge encoded in these resources:

sense definitions, synonymy, polysemy, semantic relations, structure, etc.

Knowledge-based Sense Representations

Features:

- Use **knowledge from lexical-semantic resources** for distinguishing senses from each other
- The resulting sense representations are **linked to the inventory**, hence useful for applications such as WSD

Knowledge-based Sense Representations

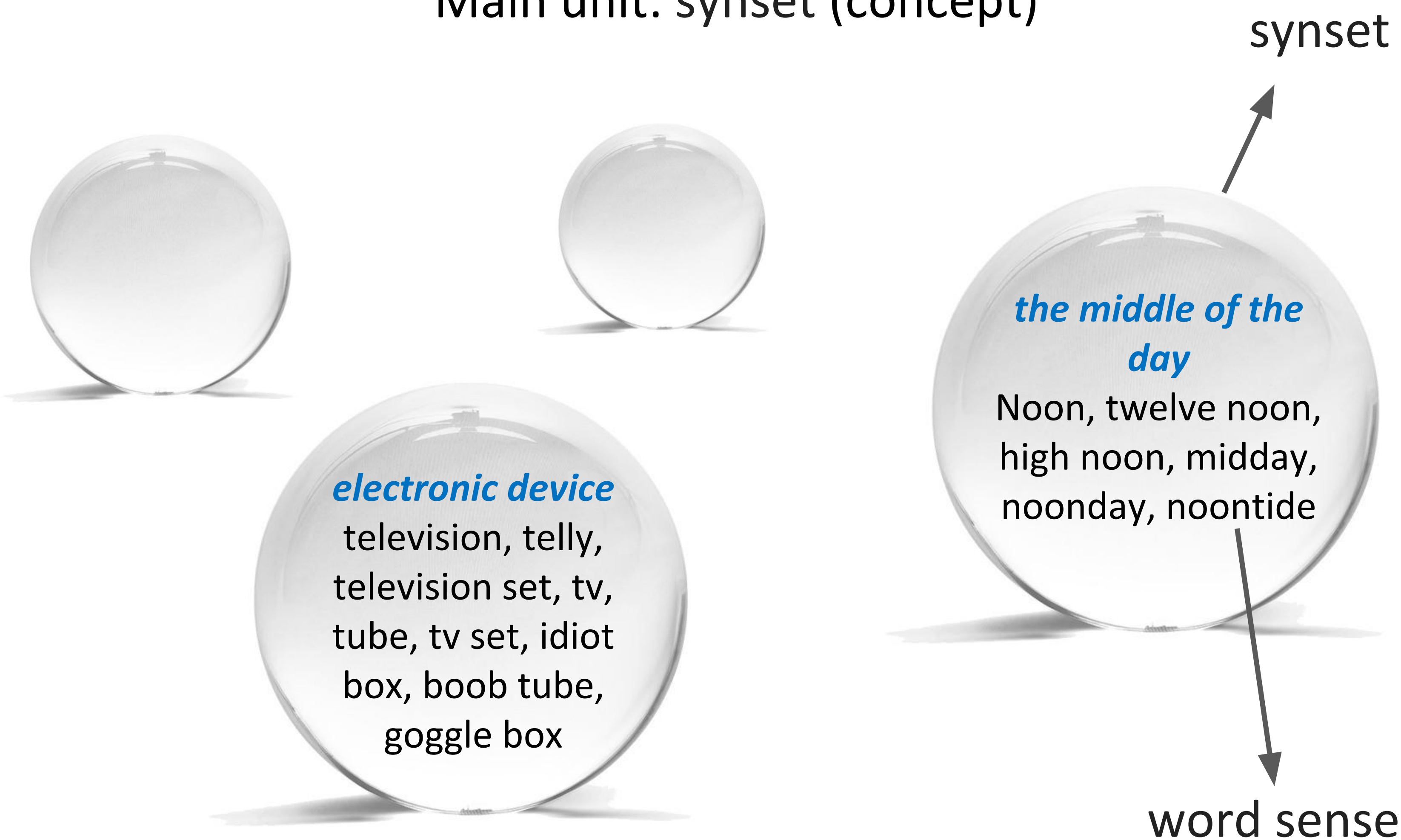
Represent word senses as defined by sense inventories

WordNet: the most commonly used

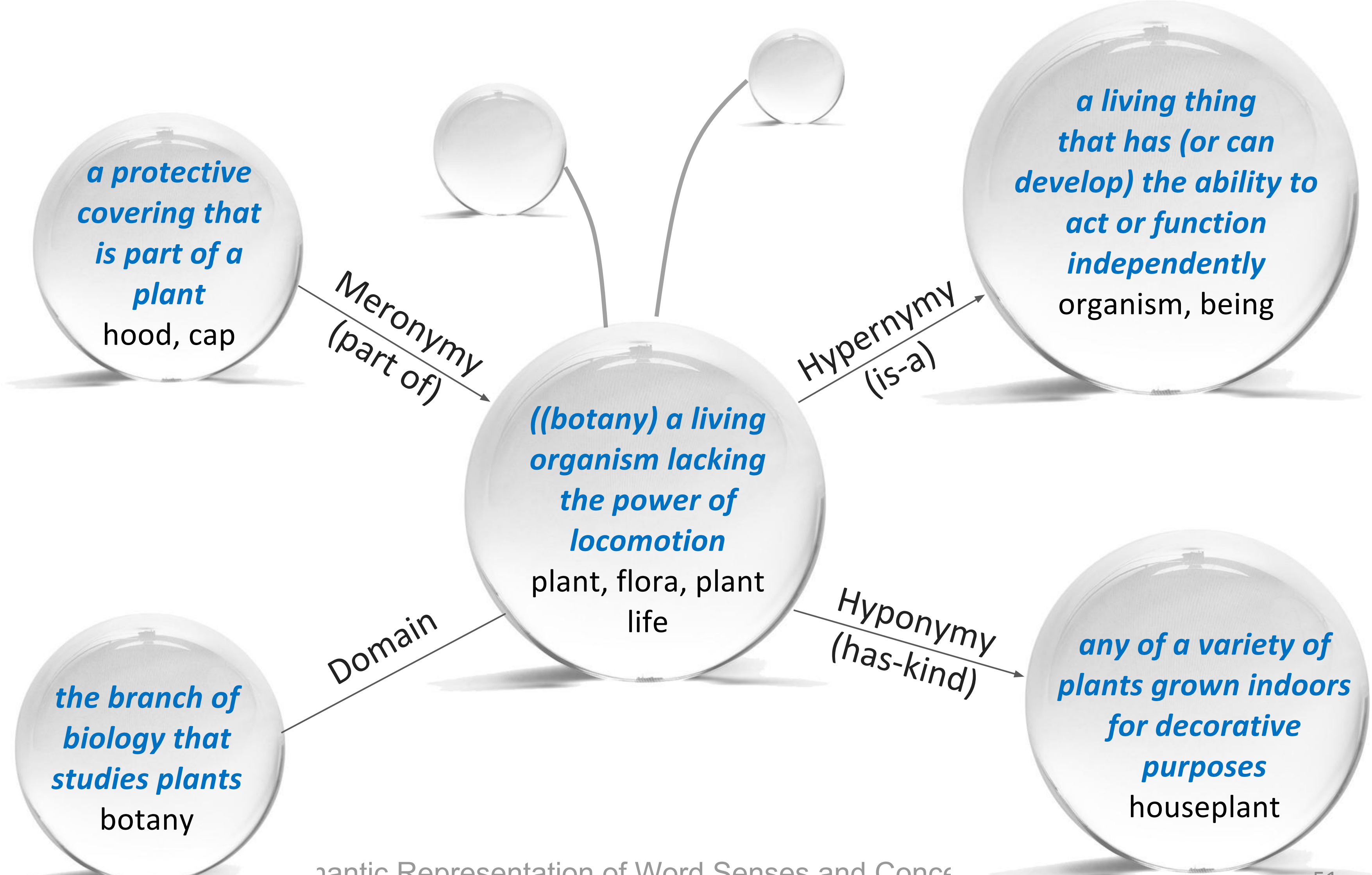
But also

WordNet

Main unit: synset (concept)



WordNet semantic relations



WordNet

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: plant [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **plant**, works, industrial plant (buildings for carrying on industrial labor) *"they built a large plant to manufacture automobiles"*
- S: (n) **plant**, flora, plant life ((botany) a living organism lacking the power of locomotion)
- S: (n) **plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- S: (n) **plant** (something planted secretly for discovery by another) *"the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"*

Verb

- S: (v) **plant**, set (put or set (seeds, seedlings, or plants) into the ground) *"Let's plant flowers in the garden"*
- S: (v) implant, engraft, embed, imbed, **plant** (fix or set securely or deeply) *"He planted a knee in the back of his opponent"; "The dentist implanted a tooth in the gum"*
- S: (v) establish, found, **plant**, constitute, institute (set up or lay the groundwork for) *"establish a new department"*
- S: (v) **plant** (place into a river) *"plant fish"*
- S: (v) **plant** (place something or someone in a certain position in order to secretly observe or deceive) *"Plant a spy in Moscow"; "plant bugs in the dissident's apartment"*
- S: (v) **plant**, implant (put firmly in the mind) *"Plant a thought in the students' minds"*

Link to online browser

Knowledge-based Sense Representations

X. Chen, Z. Liu, M. Sun: **A Unified Model for Word Sense Representation and Disambiguation** (EMNLP 2014)

☆ S. Rothe and H. Schutze: **AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes** (ACL 2015)

R. Johansson and L. Nieto Piña: **Embedding a Semantic Network in a Word Space** (NAACL 2015, short)

S. K. Jauhar, C. Dyer, E. Hovy: **Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models** (NAACL 2015)

M. T. Pilehvar and N. Collier, **De-Conflated Semantic Representations** (EMNLP 2016)

☆ M. T. Pilehvar, D. Jurgens and R. Navigli: **Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity** (ACL 2013)

Chen et al (2014)

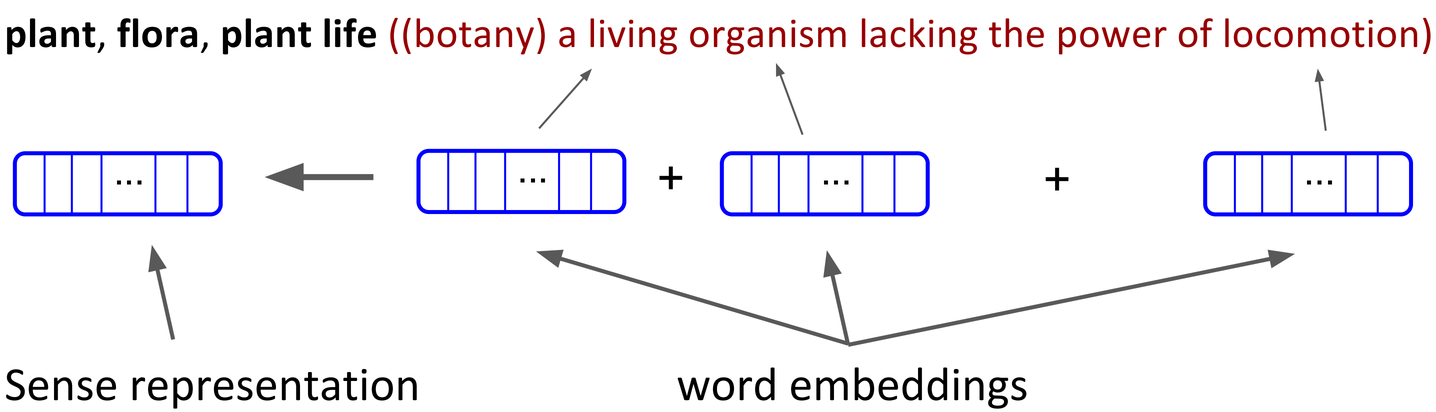
A Unified Model for Word Sense Representation and Disambiguation

Basic idea: word sense representation and Word Sense Disambiguation can benefit from each other

⇒ Joint word sense representation and disambiguation

Chen et al (2014)

1- Use a sense definition to initialize its representation



Chen et al (2014)

1- Use a sense definition to initialize its representation

2- Automatically disambiguate large amounts of text

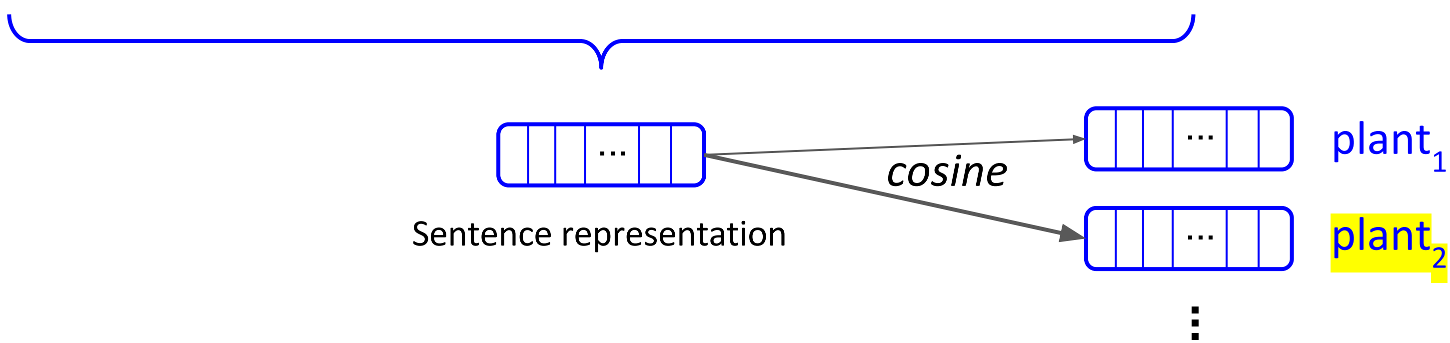
They **proposed simple disambiguation techniques** based on the obtained initial sense representations and used these disambiguation techniques to **disambiguate large amounts of texts**

Chen et al (2014)

Disambiguation Technique

To disambiguate a content word (plant):

water is absorbed by roots of a plant from the soil



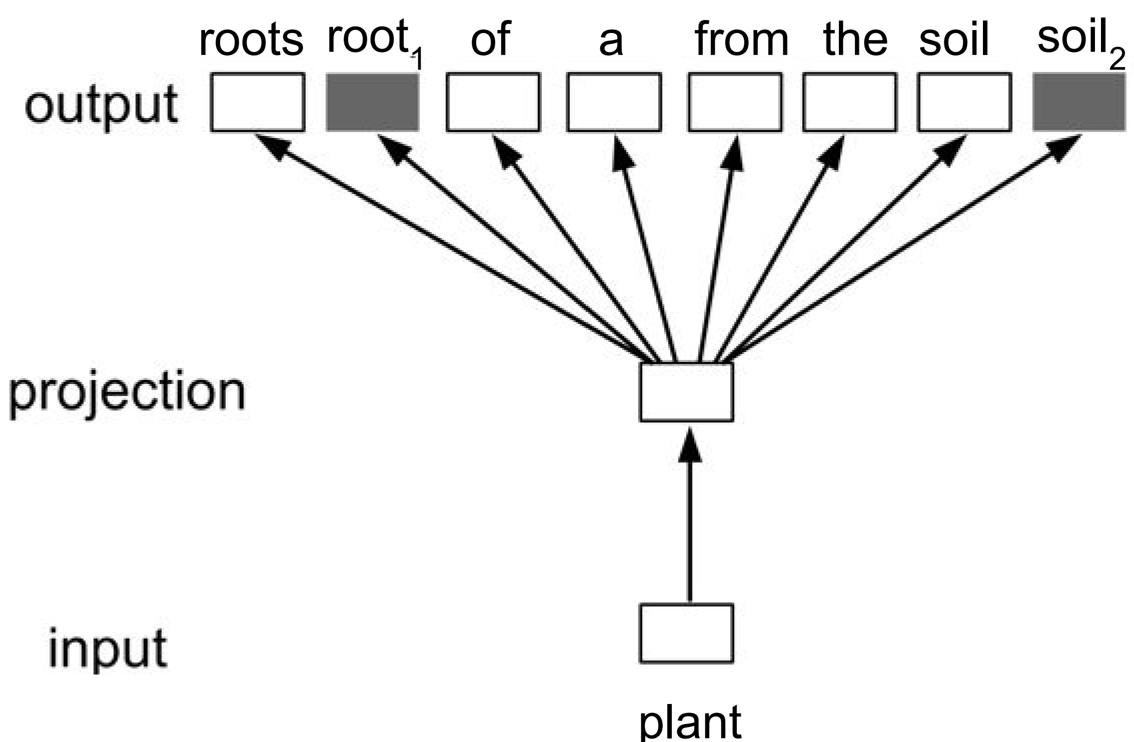
- Obtain the sentence representation (by averaging word embeddings)
- Pick the sense of *plant* which has the highest cosine similarity to the sentence vector

Chen et al (2014)

1- Use a sense definition to initialize its representation

2- Automatically disambiguate large amounts of text

3- Modify the objective of Skip-gram to learn sense representations



Chen et al (2014)

Experiments and evaluation

Word similarity measurement

The most commonly used benchmark for the evaluation of sense representation techniques

Chen et al (2014)

Experiments and evaluation

Word similarity measurement

RG-65
MC-30
TOEFL
MEN
WordSim-353
SimLex-999
SCWS
....and many more

also the **SemEval-2017 task on Multilingual and Cross-lingual Word Similarity**

The most comm he evaluation of
se ues

Chen et al (2014)

Experiments and evaluation

Word similarity measurement

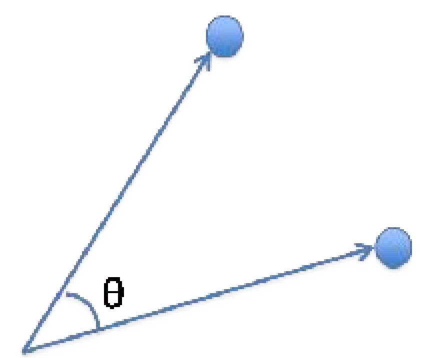
But:

- How **vector** representations are used to measure **semantic similarity**?
- How **sense** representations are used for measuring **word** similarity?

Vector Comparison
Cosine Similarity

The most commonly used measure for the similarity of vector space model (sense) representations

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



How are sense representations used for word similarity?

Usually, four techniques are used (Reisinger and Mooney, 2010):

1- MaxSim

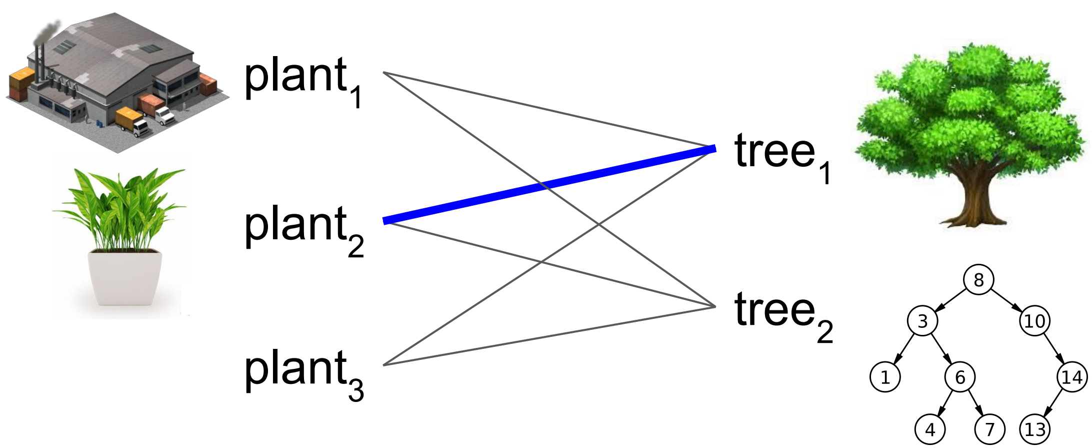
2- AvgSim

3- MaxSimC

4- AvgSimC

How are sense representations used for word similarity?

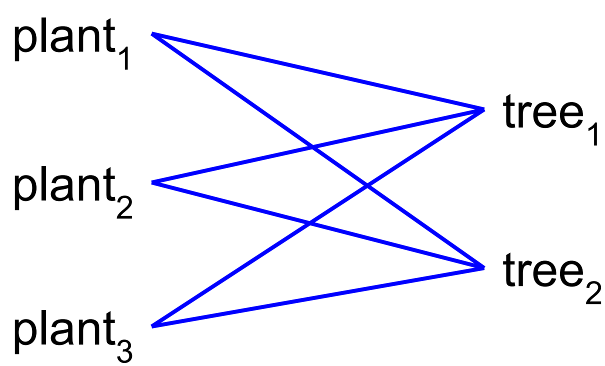
1- **MaxSim**: pick the similarity between the most similar senses across two words



$$\text{MaxSim}(w, w') \stackrel{\text{def}}{=} \max_{1 \leq j \leq K, 1 \leq k \leq K'} d(\pi_k(w), \pi_j(w'))$$

How are sense representations used for word similarity?

2- **AvgSim**: average the similarities between senses across two words



$$\text{AvgSim}(w, w') \stackrel{\text{def}}{=} \frac{1}{KK'} \sum_{j=1}^{K} \sum_{k=1}^{K'} d(\pi_k(w), \pi_j(w'))$$

How are sense representations used for word similarity?

For some datasets, words are provided with contexts, e.g., Stanford Contextual Word Similarity (SCWS)

plant

In a thermal power **plant** heat energy is converted to electric power.

tree

Almost 400 billion **trees** grow in the Amazon rainforest.

How are sense representations used for word similarity?

3- **MaxSimC**: the similarity between the "most appropriate" senses of the two words

In a thermal power $\boxed{\begin{array}{l}\text{plant}_1\\ \text{plant}_2\\ \text{plant}_3\end{array}}$ heat energy is converted to electric power.

Almost 400 billion $\boxed{\begin{array}{l}\text{tree}_1\\ \text{tree}_2\end{array}}$ grow in the Amazon rainforest.

$$\mathrm{MaxSimC}(w, w') \stackrel{\mathrm{def}}{=} d(\hat{\pi}(w), \hat{\pi}(w'))$$

The most appropriate sense of the word w given the context

How are sense representations used for word similarity?

4- **AvgSimC**: average of pairwise similarities weighted by their appropriateness in context

In a thermal power plant$_1$ / plant$_2$ / plant$_3$ heat energy is converted to electric power.

Almost 400 billion tree$_1$ / tree$_2$ grow in the Amazon rainforest.

The likelihood of a sense in the context

$$\text{AvgSimC}(w, w') \stackrel{\text{def}}{=} \frac{1}{KK'} \sum_{j=1}^{K} \sum_{k=1}^{K'} d_{c,w,k} \boxed{d_{c',w',j}} d(\pi_k(w), \pi_j(w'))$$

Chen et al (2014)

Results on the SCWS dataset:

word embeddings → Our Model-S

sense embeddings → Our Model-M

Model	$\rho \times 100$
Our Model-S	64.2
Our Model-M	**68.9**

Sense representations usually improve over word representations on word similarity benchmarks

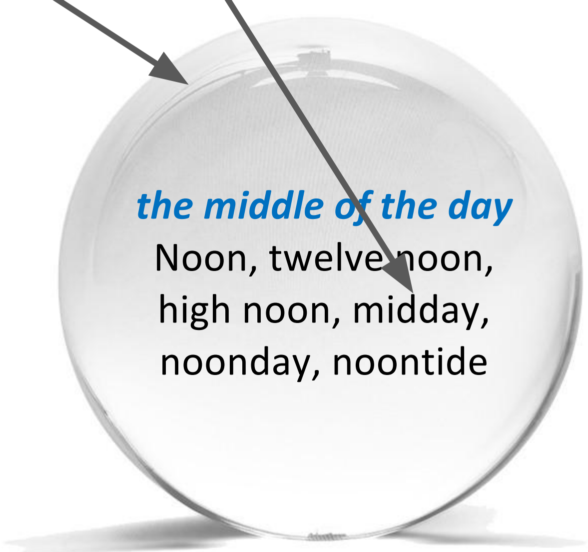
Chen et al (2014)

Limitations:

- Content words in definitions are not always enough for accurately pinpointing the semantics of a word sense
- The disambiguation technique is far from optimal which introduces noise to the representation procedure

Rothe and Schütze (2015)

AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Senses



Rothe and Schütze (2015)

AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Senses

Leverages WordNet properties (constraints) for learning sense representations

polysemy and synonymy

Rothe and Schütze (2015)

Two basic premises:

1- A word is the sum of its senses

e.g., embedding of plant is the sum of embeddings of plant(organism), plant(industry), etc.

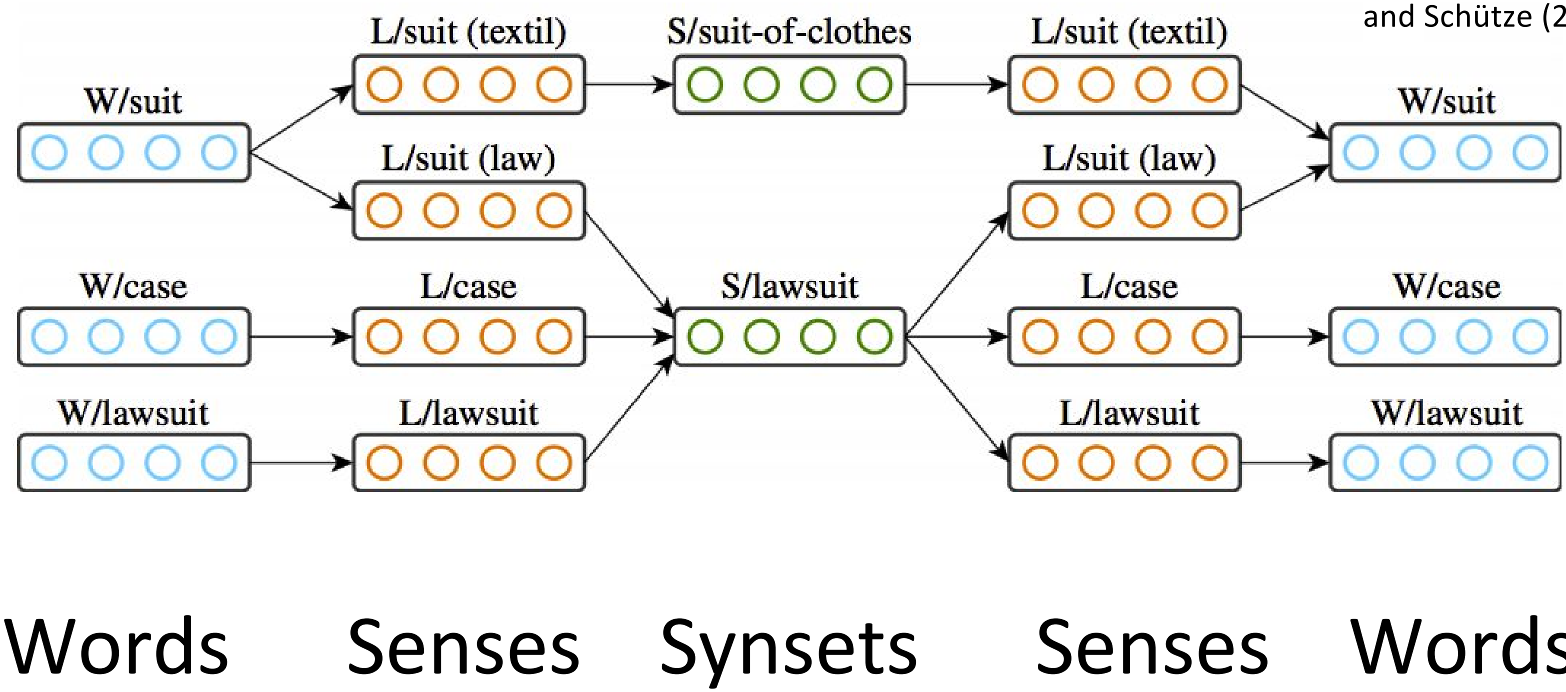
2- A synset is the sum of its senses

e.g., embedding of this synset is:
plant (organism) + flora (organism) + plant_life (organism)

Rothe and Schütze (2015)

An autoencoder framework for learning

Illustration from Rothe and Schütze (2015)



Rothe and Schütze (2015)

Word similarity experiments

Stanford Contextual Word Similarity

		AvgSim	AvgSimC
1	Huang et al. (2012)	62.8†	65.7†
2	Tian et al. (2014)	–	65.4†
3	Neelakantan et al. (2014)	67.2	69.3
4	Chen et al. (2014)	66.2†	68.9
5	words (word2vec)	66.6‡	66.6†
6	synsets	62.6†	63.7†
7	lexemes	**68.9**	**69.8**

Johansson and Nieto Piña (2015)

Embedding a Semantic Network in a Word Space

(NAACL 2015, short)

Learns sense embeddings in the same semantic space as (pre-trained) word embeddings

Applied to Swedish data:

SALDO semantic network

Johansson and Nieto Piña (2015)

target and neighbour sense representations

$$\underset{E,p}{\text{minimize}} \quad \sum_{i,j,k} w_{ijk} \Delta(E(s_{ij}), E(n_{ijk}))$$

$$\text{subject to} \quad \sum_{j} p_{ij} E(s_{ij}) = F(l_i) \qquad \forall i$$

word representation

The distances between neighbours to be minimized, while satisfying the <u>mix constraint</u> for each lemma

a word vector is a convex combination of its senses vectors

Johansson and Nieto Piña (2015)

Evaluation on

classifying frames in FrameNet

Frame	P	R	F
ANIMALS	0.741	0.643	0.689
FOOD	0.684	0.679	0.682
PEOPLE_BY_VOCATION	0.595	0.651	0.622
ORIGIN	0.789	0.691	0.737
PEOPLE_BY_ORIGIN	0.693	0.481	0.568
Overall	0.569	0.292	0.386

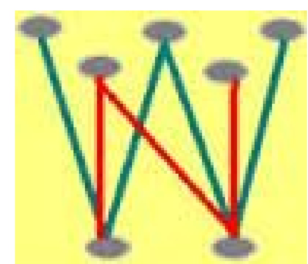
(a) Using lemma embeddings.

Frame	P	R	F
ANIMALS	0.826	0.663	0.736
FOOD	0.726	0.743	0.735
PEOPLE_BY_VOCATION	0.605	0.637	0.621
ORIGIN	0.813	0.684	0.742
PEOPLE_BY_ORIGIN	0.756	0.508	0.608
Overall	0.667	0.332	0.443

(b) Using sense embeddings.

Retrofitting (Faruqui et al., NAACL 2015)

Retrofitting Word Vectors to Semantic Lexicons. Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith (NAACL 2015)



Distributional approaches usually rely **only** on the **statistics** derived from text corpora

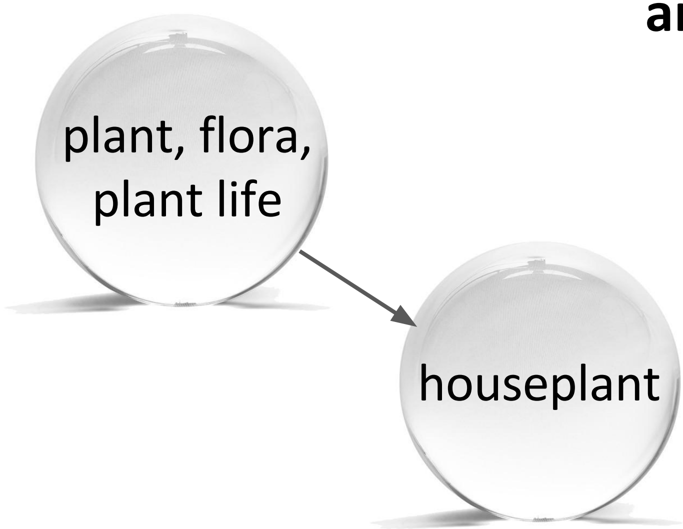
They usually **ignore** all the valuable information encoded in **knowledge resources**

Retrofitting (Faruqui et al., NAACL 2015)

WordNet

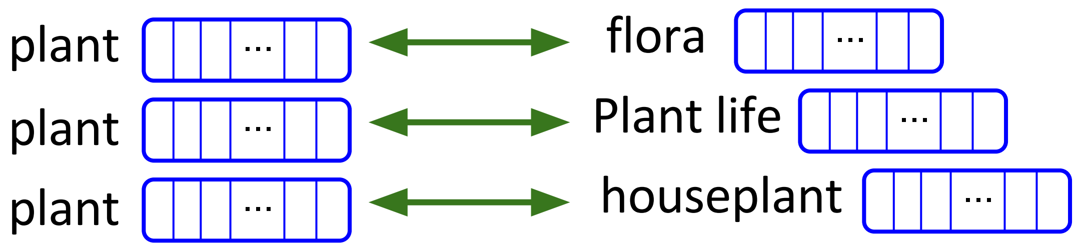
The Paraphrase Database

Benefit from synonymy and other semantic relationships in resources

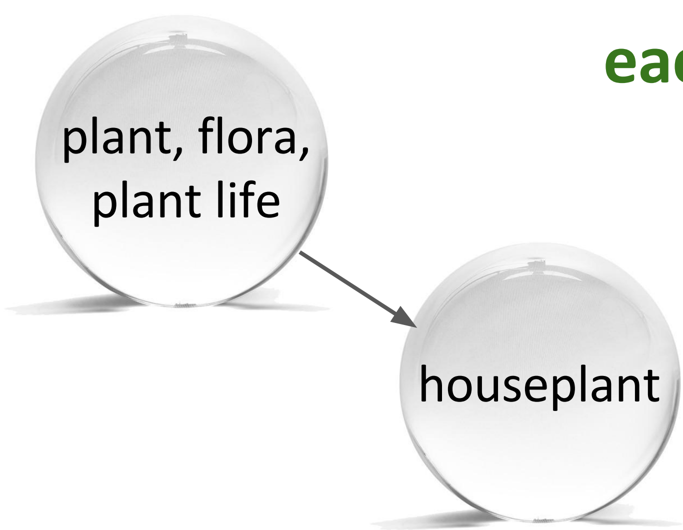


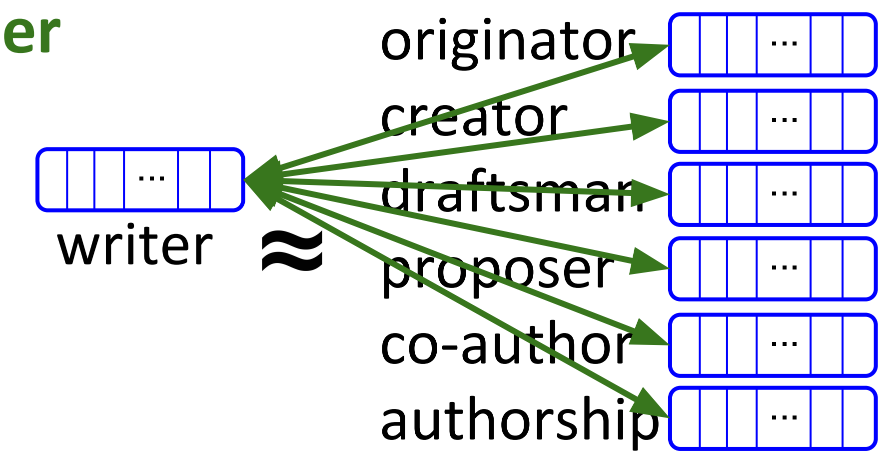
writer ≈ originator
creator
draftsman
proposer
co-author
authorship

Retrofitting (Faruqui et al., NAACL 2015)



Make these vectors more similar to each other





Jauhar et al. (NAACL 2015)

Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models (S. K. Jauhar, C. Dyer and E. Hovy)

Two techniques for learning sense-specific embeddings that are linked to WordNet: **Retro** and **EM**

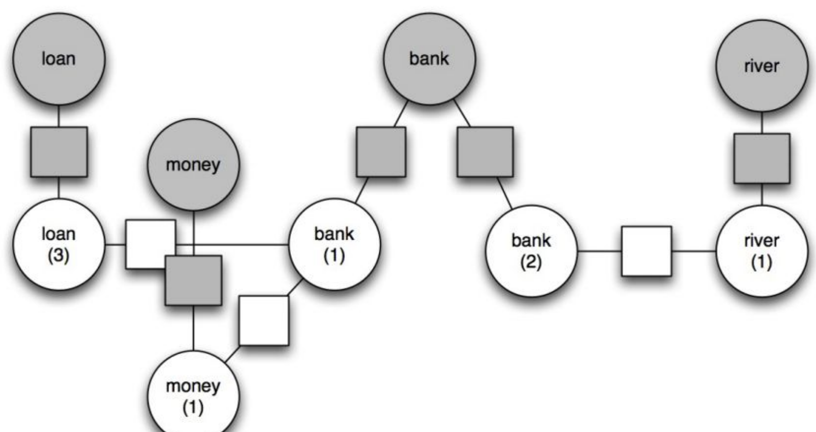
Jauhar et al. (NAACL 2015)

RETRO

$$C(V) = \arg\min_{V} \sum_{i-ij} \alpha \|\hat{u}_i - v_{ij}\|^2 + \sum_{ij-i'j'} \beta_r \|v_{ij} - v_{i'j'}\|^2$$

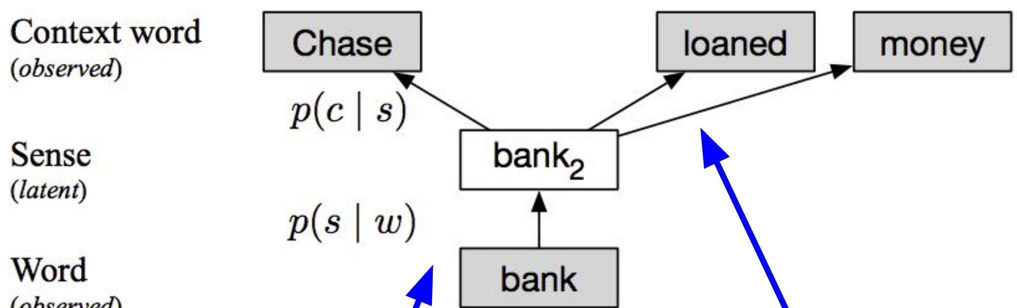
Initial word vectors

Sense vectors



Jauhar et al. (NAACL 2015)

EM: Extends the skip-gram model to learn ontologically-grounded sense vectors

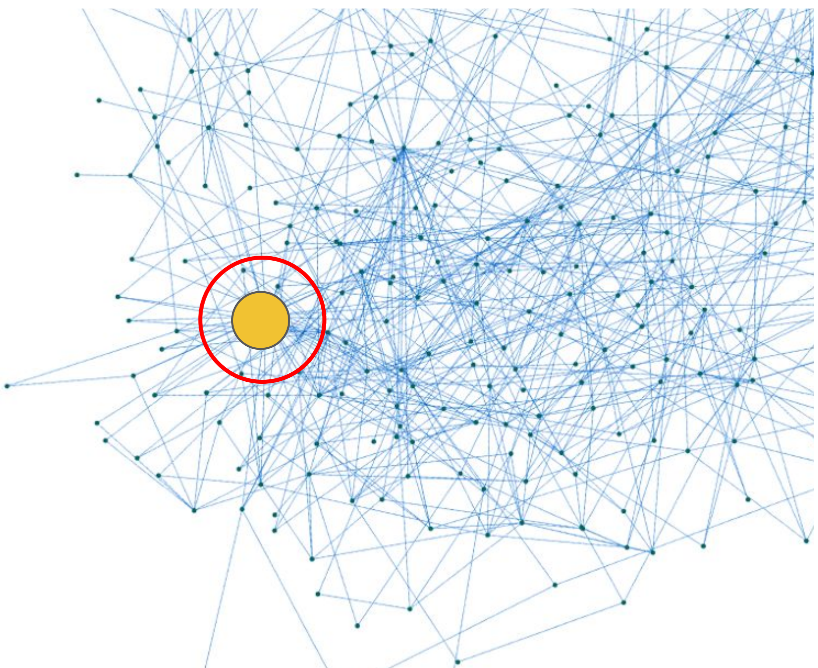


$$C(\theta) = \arg\max_{\theta} \sum_{(w_i,c_i)\in D} \log\Big(\sum_{s_{ij}} p(c_i \mid s_{ij};\theta)\times p(s_{ij} \mid w_i;\theta)\Big) - \gamma \sum_{ij-i'j'} \beta_r \|v_{ij} - v_{i'j'}\|^2$$

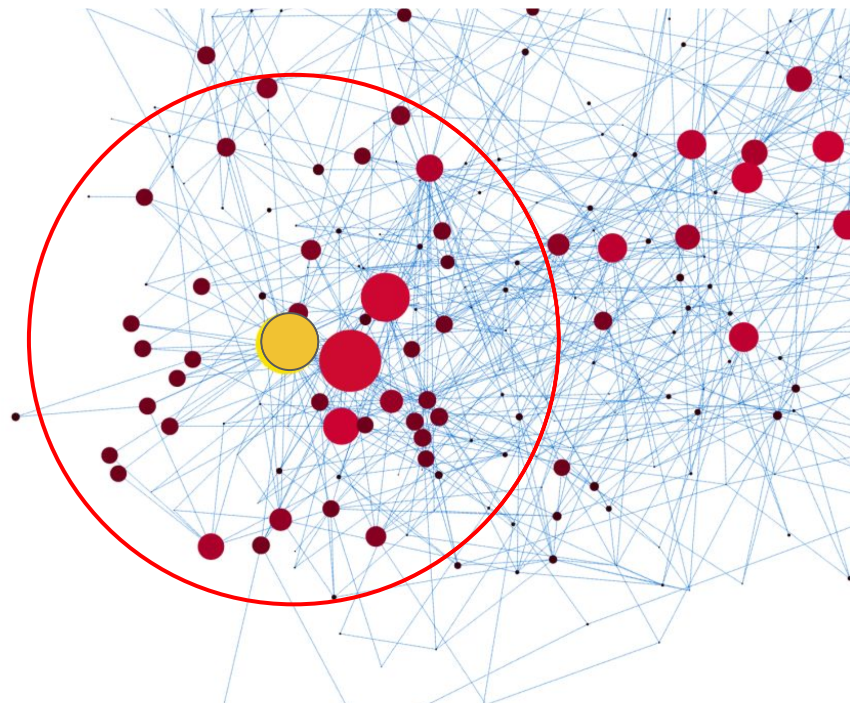
Ontological prior

De-Conflated Semantic Representations

Approaches so far



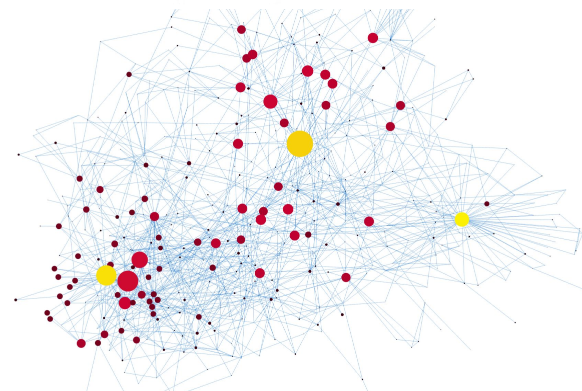
M. T. Pilehvar and N. Collier (EMNLP 2016)



De-Conflated Semantic Representations

Uses Personalized PageRank algorithm to exploit WordNet for sense specific information

$$\vec{v}^{(t)} = (1-\alpha)\, M\vec{v}^{(t-1)} + \alpha\, \vec{v}^{(0)}$$

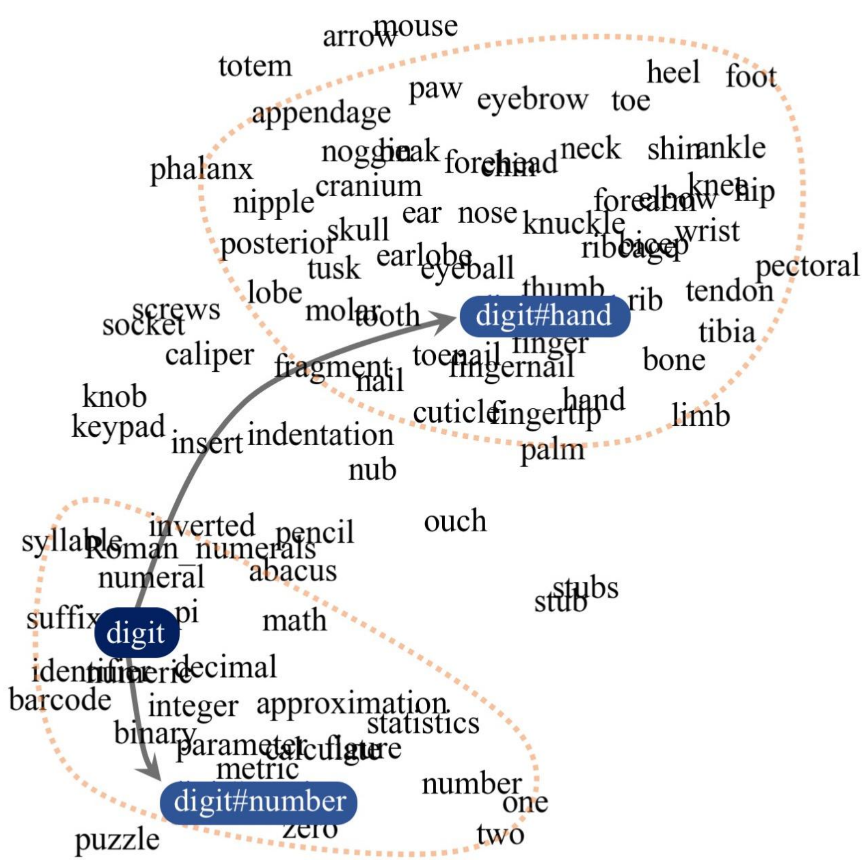


Digit

#	Sense biasing words
1	dactyl, finger, toe, thumb, pollex, body_part, nail, minimus, tarsier, webbed, extremity, appendage
2	figure, cardinal_number, cardinal, integer, whole_number, numeration_system, number_system, system_of_numeration, large_integer, constituent, element, digital

De-Conflated Semantic Representations

M. T. Pilehvar and N. Collier (EMNLP 2016)



De-Conflated Semantic Representations

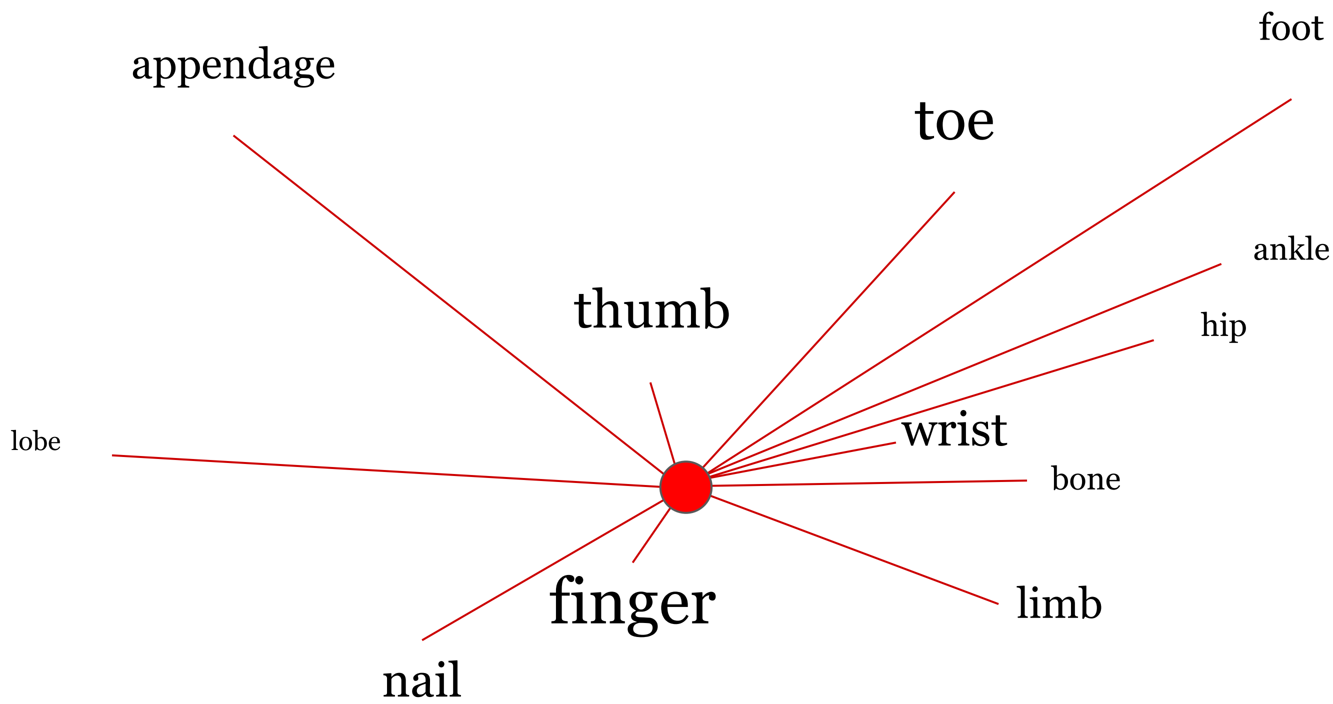
- Learns a representation $v^*_{s_i}$ for a sense s_i that is:

  - Close to its **lemma embedding**

  - Close to a weighted average of **embeddings of its sense biasing words**

$$\arg\min_{v^*} \alpha \, d(v^*_{s_i}, v_{s_i}) + \sum_{b_{ij} \in \mathcal{B}_i} \delta_{ij} \, d(v^*_{s_i}, v_{b_{ij}})$$

De-Conflated Semantic Representations



De-Conflated Semantic Representations

Evaluation: **Word Similarity**

Approach	Score AvgSim	Score AvgSimC
DECONF	**70.8**	**71.5**
Rothe and Schütze (2015) (best)	68.9	69.8
Neelakantan et al. (2014) (best)	67.3	69.3
Chen et al. (2014)	66.2	68.9
Liu et al. (2015) (best)	–	68.1
Huang et al. (2012)	62.8	65.7
Tian et al. (2014) (best)	–	65.7
Iacobacci et al. (2015)	62.4	–
Initial word vectors	65.1	

Dataset	Approach	Score r	Score ρ
MEN-3K	Iacobacci et al. (2015)	–	**80.5**
	DECONF	**78.0**	78.6
	Faruqui et al. (2015)	–	75.9
	Initial word vectors	72.3	73.2
RG-65	DECONF	**90.5**	**89.6**
	Iacobacci et al. (2015)	–	87.1
	Faruqui et al. (2015)	–	84.2
	Initial word vectors	77.2	76.1
YP-130	DECONF	**72.9**	**69.5**
	Iacobacci et al. (2015)	–	63.9
	Initial word vectors	58.0	55.9
SL-999	DECONF	**60.5**	**59.0**
	Goikoetxea et al. (2015)	–	55.2
	Initial word vectors	45.4	44.2

De-Conflated Semantic Representations

Evaluation: **Word to Sense Similarity** (SemEval-2014 task on Cross-Level Semantic Similarity)

Word similarity:

plant farm

Word to sense similarity:

plant#2 farm

System	MaxSim		AvgSim	
	r	ρ	r	ρ
DeConf*	**36.4**	**37.6**	**36.8**	**38.8**
Rothe and Schütze (2015)*	34.0	33.8	34.1	33.6
Iacobacci et al. (2015)*	19.1	21.5	21.3	24.2
Chen et al. (2014)*	17.7	18.0	17.2	16.8
DeConf	35.5	36.4	36.2	38.0
Iacobacci et al. (2015)	19.0	21.5	20.9	23.2

Align, Disambiguate, and Walk (ADW)

Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity (Pilehvar, Jurgens and Navigli, ACL 2013)

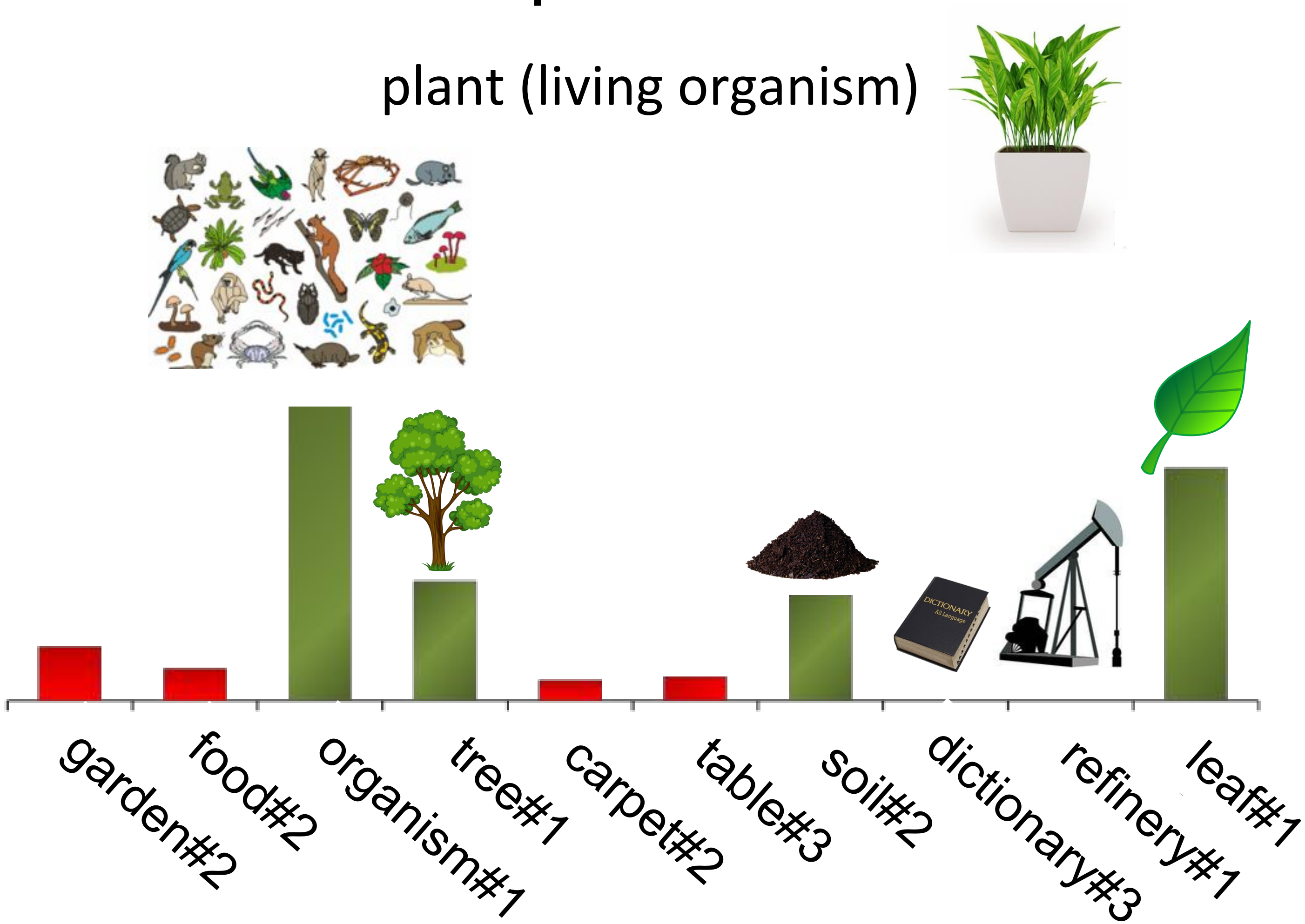
From Senses to Texts: An All-in-one Graph-based Approach for Measuring Semantic Similarity (Pilehvar and Navigli, 2015, Artificial Intelligence, 2015)

- Purely based on the knowledge derived from WordNet (no corpus statistics)
- Human-interpretable sense representations (all sense representations covered so far were non-interpretable)

ADW: Semantic Signature

Human-interpretable dimensions

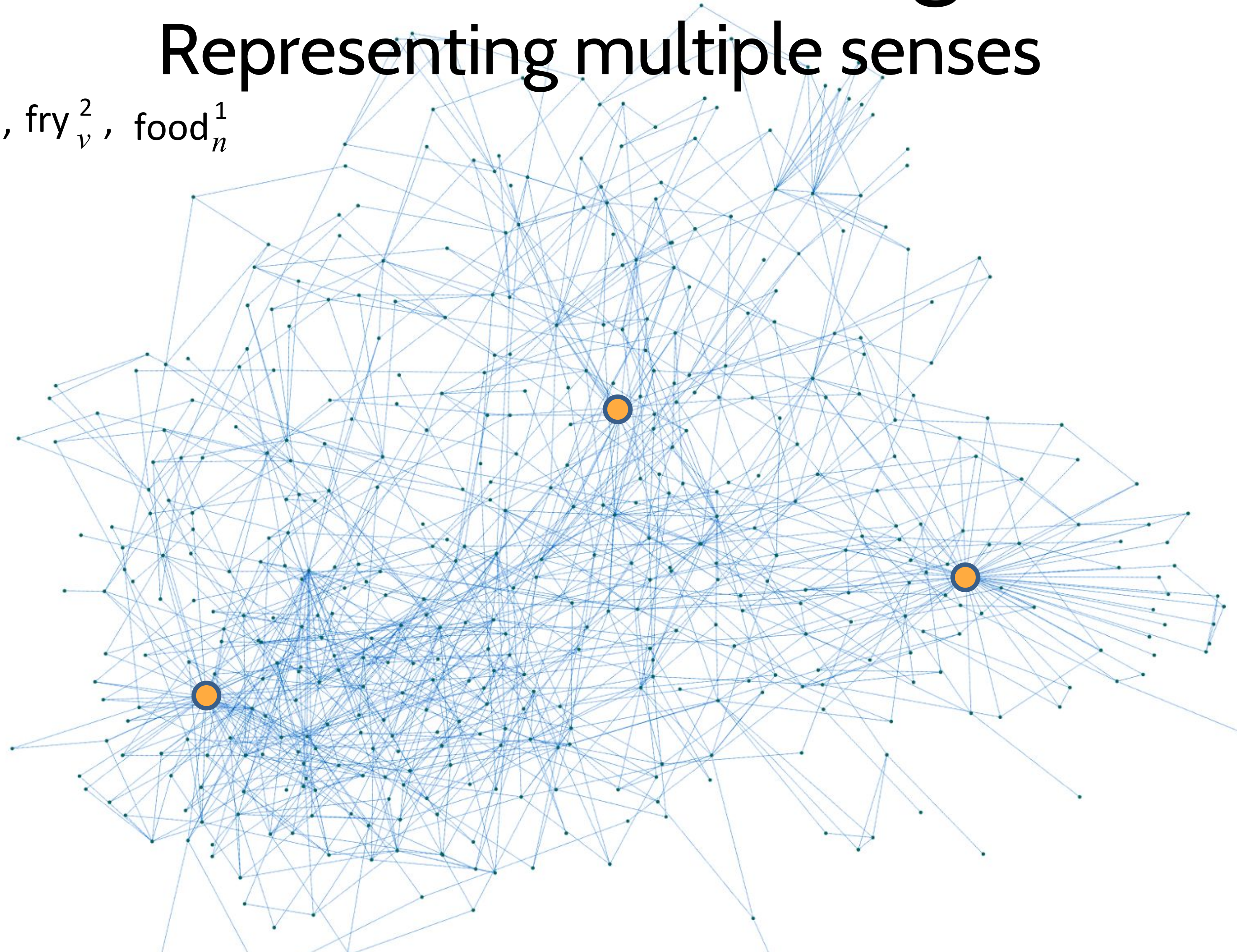
plant (living organism)



ADW: Personalized PageRank

Representing multiple senses

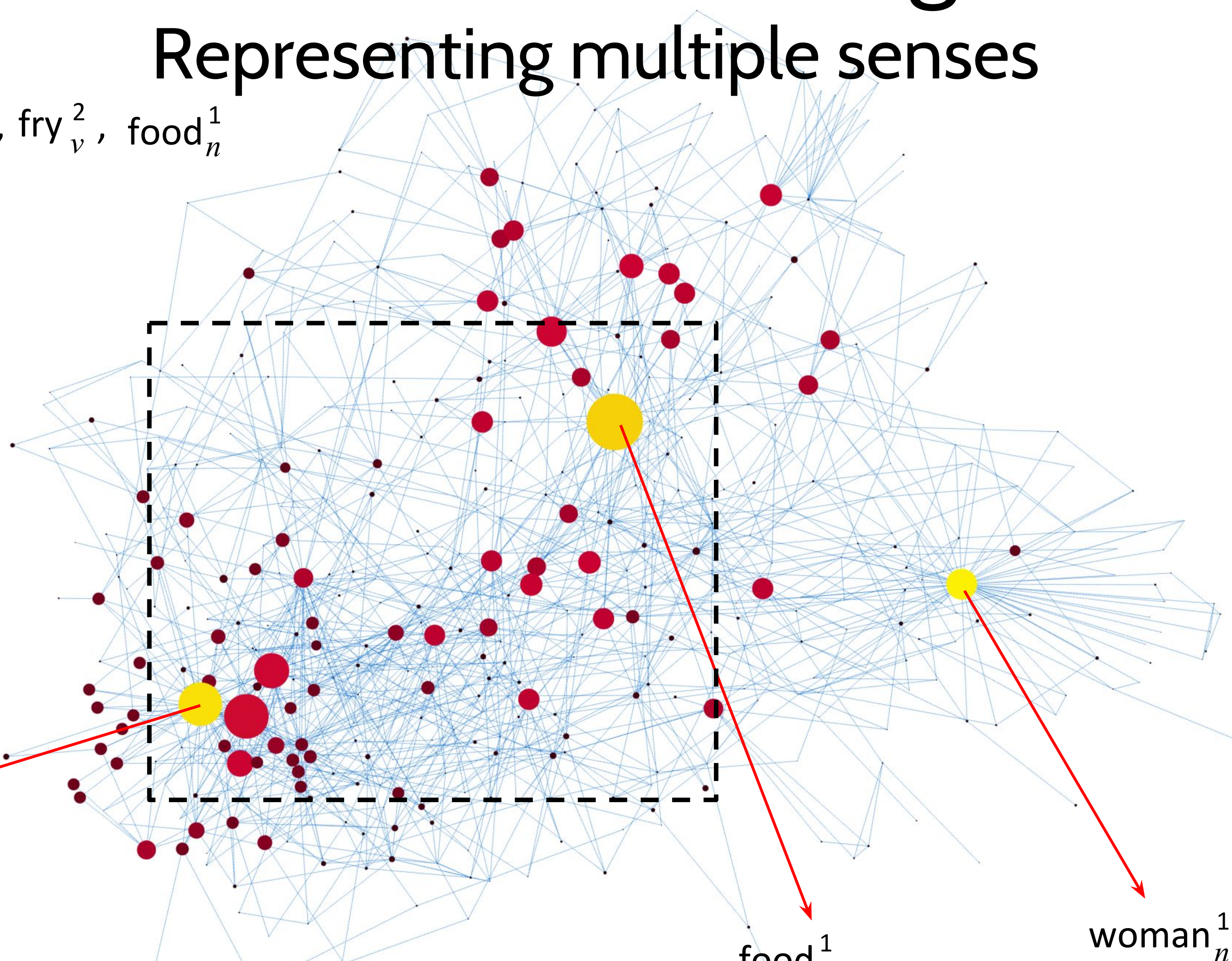
woman_n^1, fry_v^2, food_n^1

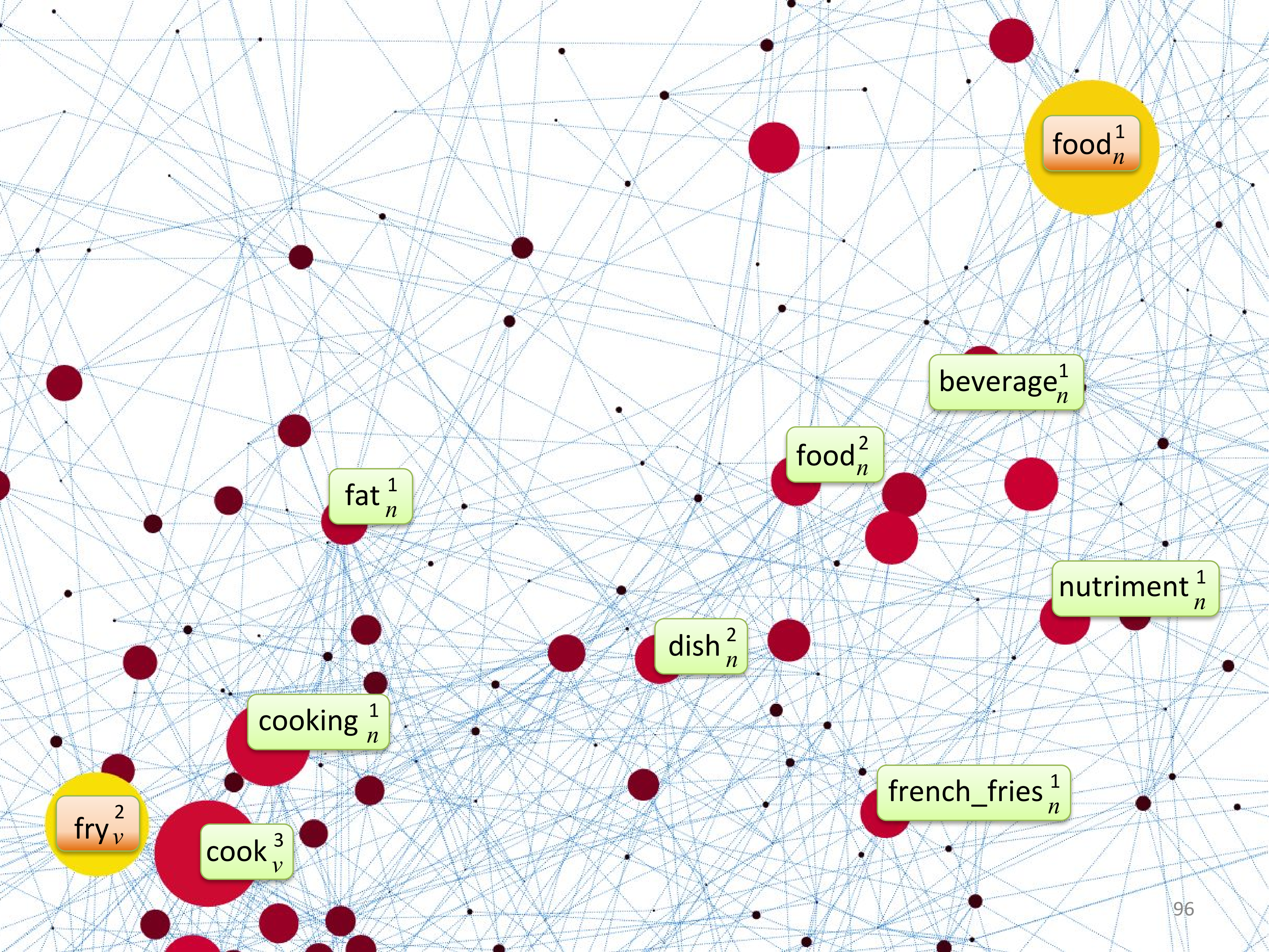


ADW: Personalized PageRank

Representing multiple senses

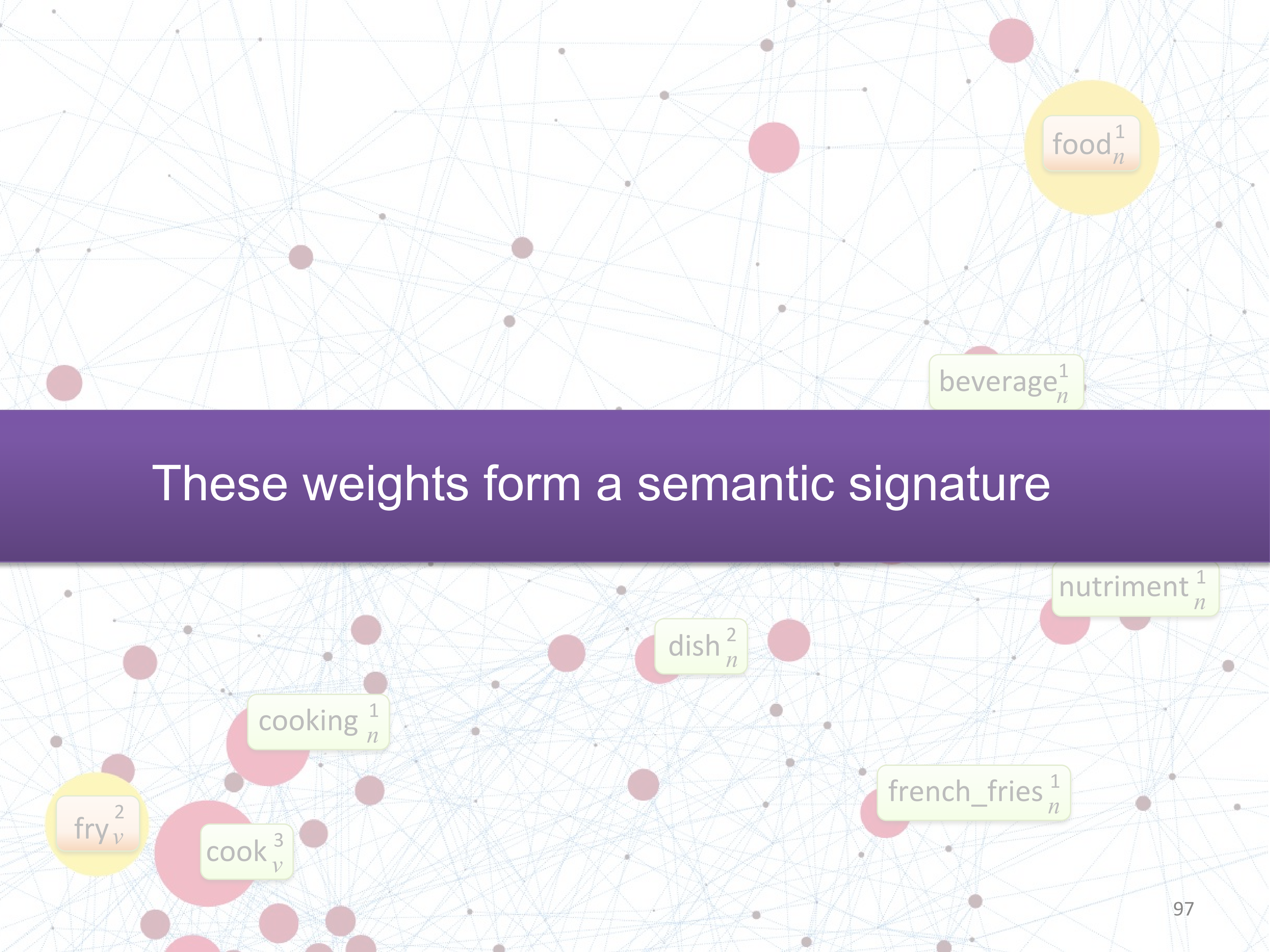
woman_n^1, fry_v^2, food_n^1





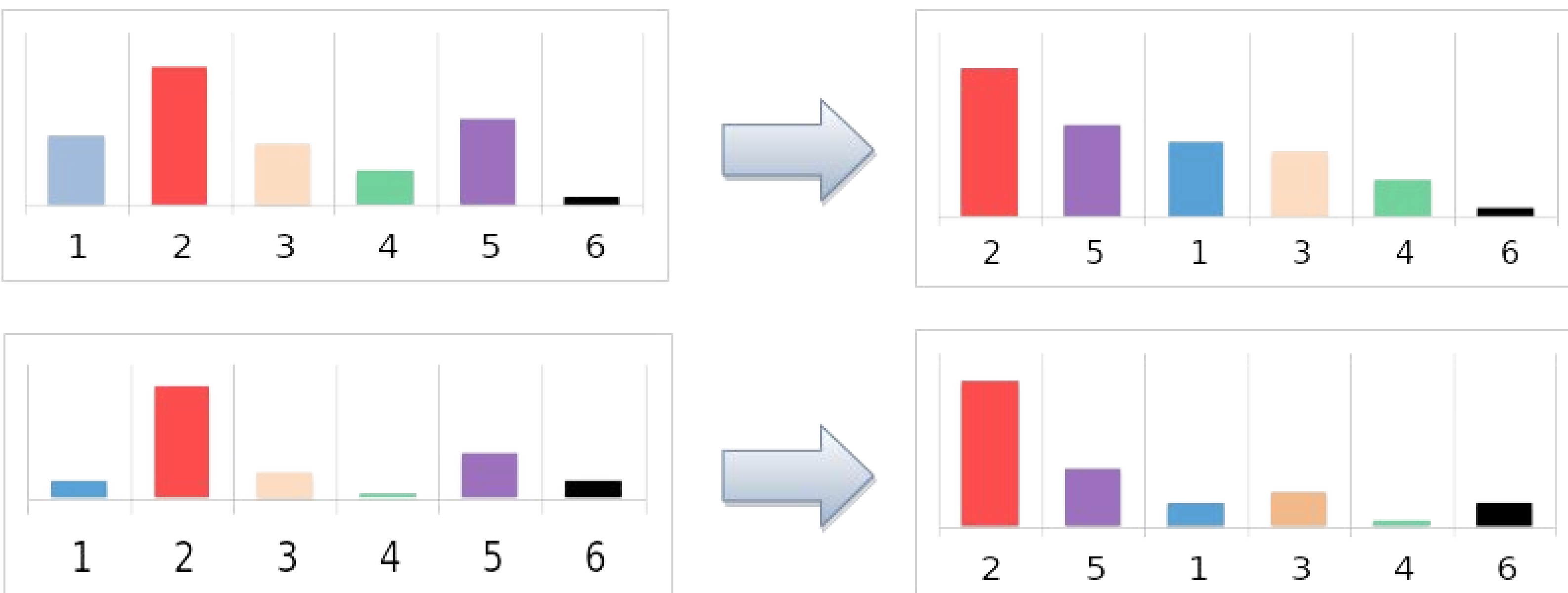
food$_n^1$
beverage$_n^1$
food$_n^2$
fat$_n^1$
nutriment$_n^1$
dish$_n^2$
cooking$_n^1$
french_fries$_n^1$
fry$_v^2$
cook$_v^3$

These weights form a semantic signature



Vector Comparison
Weighted Overlap

$$WO(v_1, v_2) = \frac{\sum_{q \in O} \left(rank(q, v_1) + rank(q, v_2)\right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}}$$



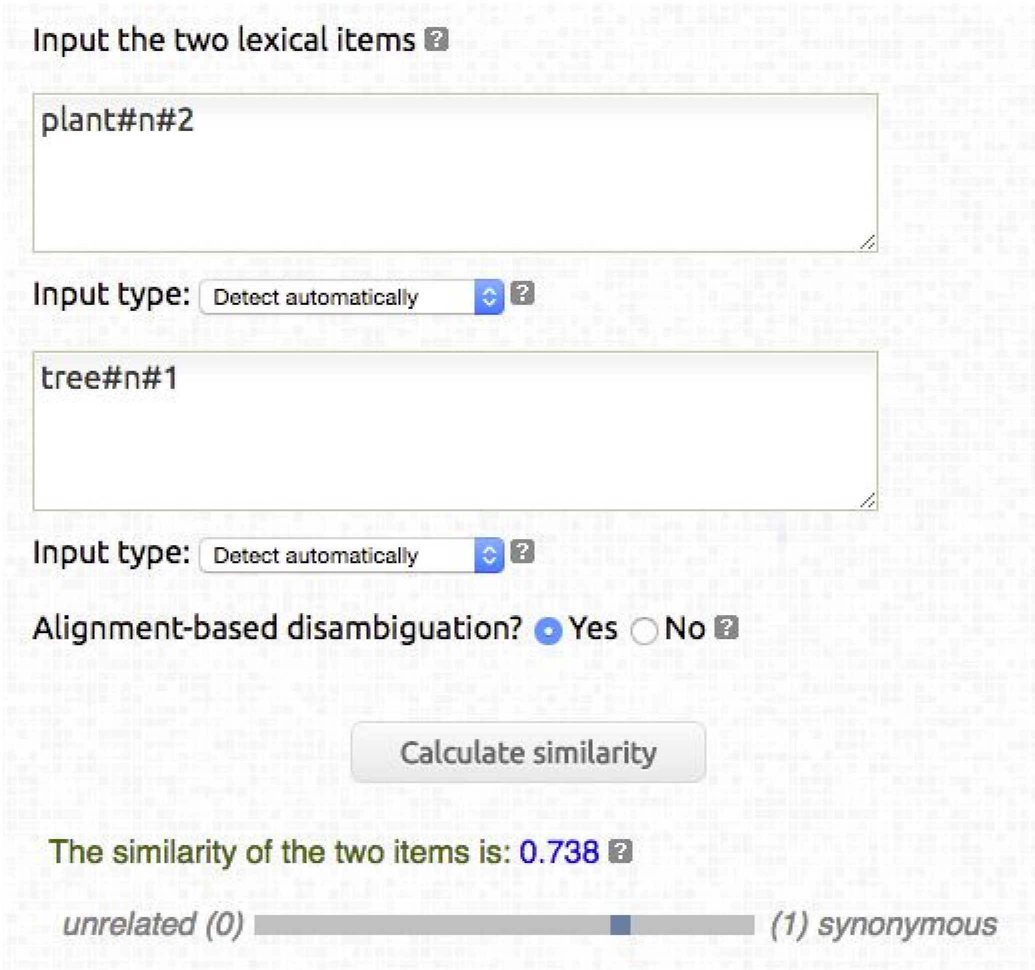
ADW

Alignment-based disambiguation

a simple technique for using sense representations for measuring semantic similarity of word, phrase or sentence pairs.

ADW

Online demo: http://lcl.uniroma1.it/adw/



ADW
Advantages and limitation

- \+ Interpretable dimensions
- \+ Unified representation for all lexical levels: senses, words, phrases and sentences
- \+ Uses only WordNet as its knowledge resource
- \+ Rich and highly accurate representations: state-of-the-art performance on multiple NLP tasks and datasets

- \- Limited coverage (that of WordNet)
  > Solution: use large-scale lexical resources

BabelNet

Large knowledge resources

Large knowledge resources

Wikipedia

BabelNet

FreeBase

Wikipedia

WIKIPEDIA
The Free Encyclopedia

Wikipedia

High coverage of **named entities** and **specialized concepts** from different domains

Wikipedia
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book

Article Talk Read Edit View history Search

University of California, Los Angeles

From Wikipedia, the free encyclopedia

Coordinates: 34°04'20.00"N 118°26'38.75"W

"UCLA", "Ucla", and "U.C.L.A." redirect here. For other uses, see UCLA (disambiguation).

The **University of California, Los Angeles** (**UCLA**) is a public research university located in the Westwood neighborhood of Los Angeles, California, United States. It became the University of California Southern Branch in 1919, making it the second-oldest undergraduate campus of the ten-campus system after the original University of California campus in Berkeley (1873).[11] It offers 337 undergraduate and graduate degree programs in a wide range of disciplines.[12] With an approximate enrollment of 30,000 undergraduate and 12,000 graduate students, UCLA has the highest enrollment of any university in California[6] and is the most applied to university in the United States with over 112,000 applications for fall 2015.[13]

The university is organized into five undergraduate colleges, seven professional schools, and four professional health science schools. The undergraduate colleges are the College of Letters and Science; Henry Samueli School of Engineering and Applied Science (HSSEAS); School of the Arts and Architecture; School of Theater, Film, and Television; and School of Nursing. Fifteen[14][15] Nobel laureates, one Fields Medalist,[16] and three Turing Award winners[17] have been faculty, researchers, or alumni. Among the current faculty members, 55 have been elected to the National Academy of Sciences, 28 to the National Academy of Engineering, 39 to the Institute of Medicine, and 124 to the American Academy of Arts and Sciences.[18] The university was elected to the Association of American Universities in 1974.[19]

UCLA student-athletes compete as the Bruins in the Pacific-12 Conference. The Bruins have won 125 national championships, including 112 NCAA team championships.[20][21] UCLA student-athletes have won 250 Olympic medals: 125 gold, 65 silver and 60 bronze.[22] The Bruins have competed in every Olympics since 1920 with one exception (1924), and have won a gold medal in every Olympics that the United States has participated in since 1932.[23]

Contents [hide]
1 History

University of California, Los Angeles

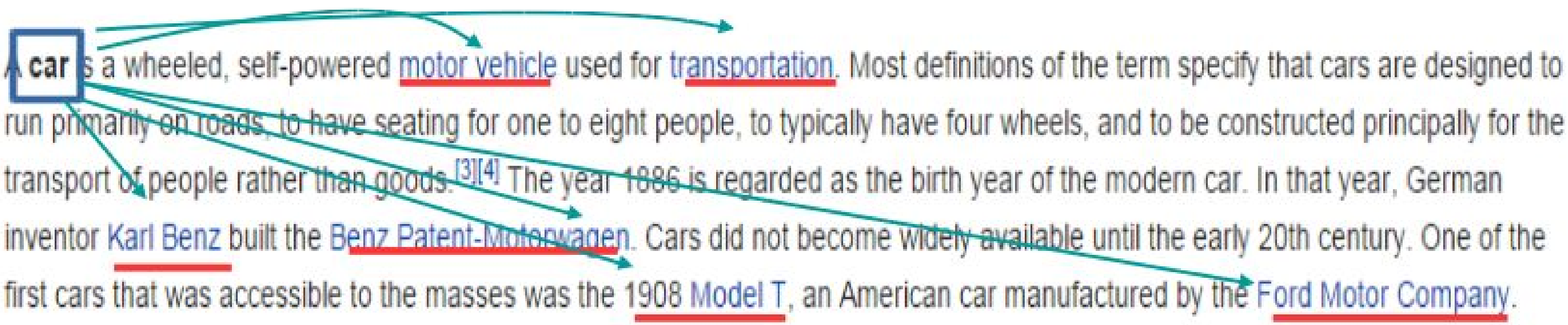
UCLA official seal

Former names	State Normal School at Los Angeles (1882-1919) University of California Southern Branch (1919–1927) University of California at Los Angeles (1927–1958)
Motto	Fiat lux (Latin)
Motto in English	Let there be light

Wikipedia hyperlinks

A **car** is a wheeled, self-powered motor vehicle used for transportation. Most definitions of the term specify that cars are designed to run primarily on roads, to have seating for one to eight people, to typically have four wheels, and to be constructed principally for the transport of people rather than goods.[3][4] The year 1886 is regarded as the birth year of the modern car. In that year, German inventor Karl Benz built the Benz Patent-Motorwagen. Cars did not become widely available until the early 20th century. One of the first cars that was accessible to the masses was the 1908 Model T, an American car manufactured by the Ford Motor Company.

Wikipedia hyperlinks



A **car** is a wheeled, self-powered motor vehicle used for transportation. Most definitions of the term specify that cars are designed to run primarily on roads, to have seating for one to eight people, to typically have four wheels, and to be constructed principally for the transport of people rather than goods.[3][4] The year 1886 is regarded as the birth year of the modern car. In that year, German inventor Karl Benz built the Benz Patent-Motorwagen. Cars did not become widely available until the early 20th century. One of the first cars that was accessible to the masses was the 1908 Model T, an American car manufactured by the Ford Motor Company.

Wikipedia

~4.3M *
Wikipages

>71M *
hyperlinks

977M *
lemmas

WIKIPEDIA
The Free Encyclopedia

*English dump 11/2014

[Based on the slides of Raganato and Delli Bovi (2016)]

Wikipedia

~**4.3M** *
Wikipages

>**71M** *
hyperlinks

977M *
lemmas

WIKIPEDIA
The Free Encyclopedia

Constantly updating and growing!

270+ active languages!

*English dump 11/2014

[Based on the slides of Raganato and Delli Bovi (2016)]

Wikipedia

~**4.3M** *
Wikipages

>**71M**
hyperlinks

977M
lemmas

WIKIPEDIA
The Free Encyclopedia

Named Entity Disambiguation (**Wikification**)

Semantic similarity

Information Extraction

Taxonomies, ontologies and semantic networks

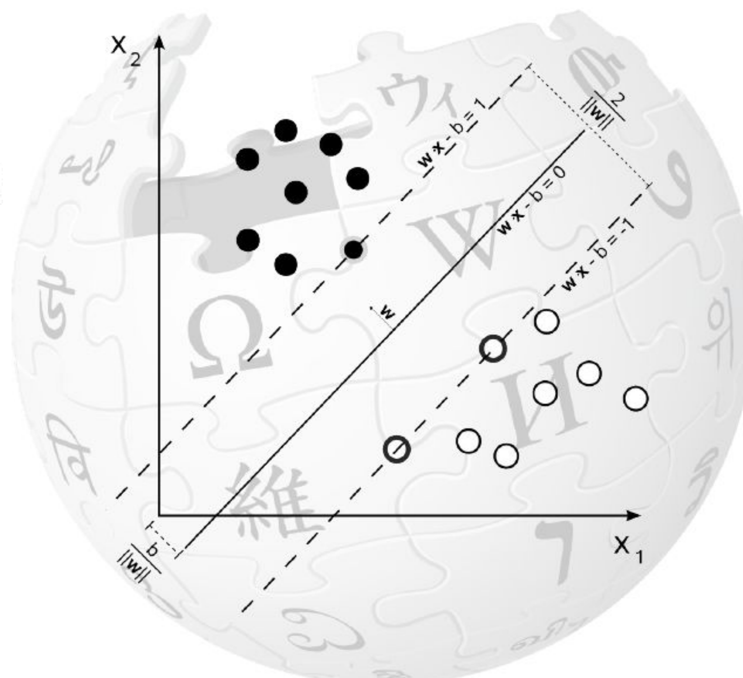
*English dump 11/2014

[Based on the slides of Raganato and Delli Bovi (2016)]

Wikipedia as a sense-annotated corpus

~**4.3M** *
concepts or entities

>**71M**
sense annotations



Named Entity Disambiguation (**Wikification**)

Semantic similarity

Information Extraction

Taxonomies, ontologies and semantic networks

*English dump 11/2014

[Based on the slides of Raganato and Delli Bovi (2016)]

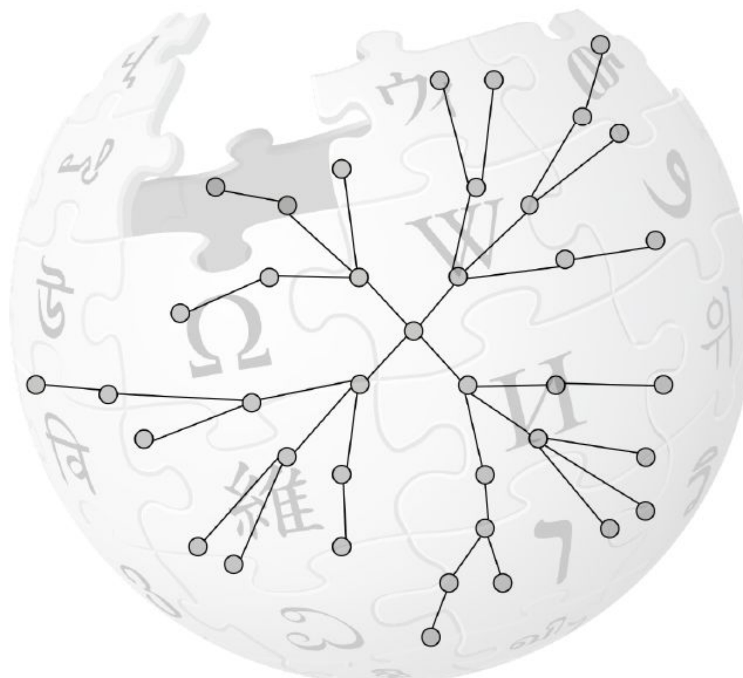
Wikipedia as a semantic network

~**4.3M** *

concept nodes

>**71M**

semantic connections



Named Entity Disambiguation (**Wikification**)

Semantic similarity

Information Extraction

Taxonomies, ontologies and semantic networks

*English dump 11/2014

[Based on the slides of Raganato and Delli Bovi (2016)]

Semantic Representations exploiting Wikipedia

- **SSA** (Hassan and Mihalcea, AAAI 2011)
- **SaSA** (Wu and Giles, AAAI 2015)

SSA: Salient Semantic Analysis

(Hassan and Mihalcea, AAAI 2011)

It exploits Wikipedia as a **sense-annotated corpus** using its hyperlinks

It increases the number of links by exploiting the **one sense per page** heuristic.

SSA: Salient Semantic Analysis

(Hassan and Mihalcea, AAAI 2011)

It exploits Wikipedia as a **sense-annotated corpus** using its hyperlinks

It increases the number of links by exploiting the **one sense per page** heuristic.

This property and other structural properties of Wikipedia have been exploited in Raganato et al. (IJCAI 2016) to build a large sense-annotated corpus.

SSA: Salient Semantic Analysis

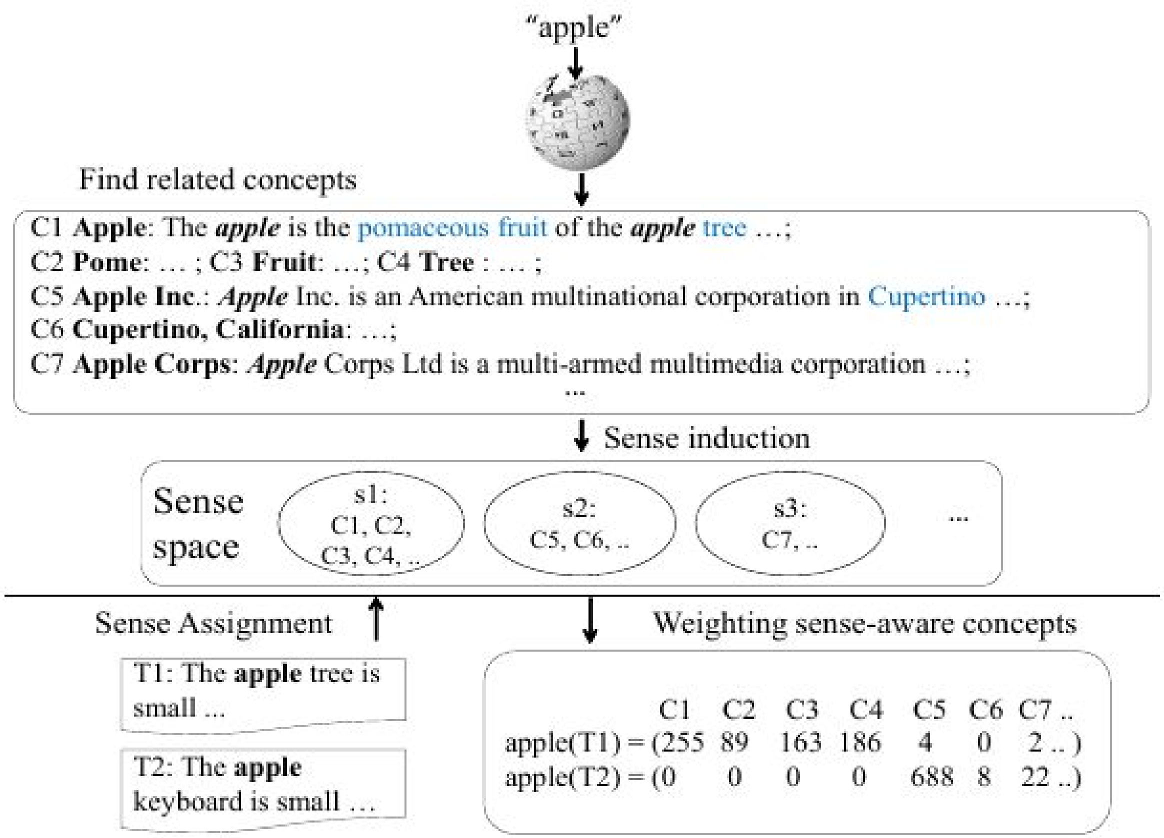
(Hassan and Mihalcea, AAAI 2011)

For a given word, it constructs an explicit vector where **dimensions are co-occurring Wikipedia pages** (weights correspond to normalized frequencies).

Strong results in **word, sentence and document relatedness**.

SaSA: Sense-aware Semantic Analysis

(Wu and Giles, AAAI 2015)



To be explained in the next section of "Unsupervised sense representations"!

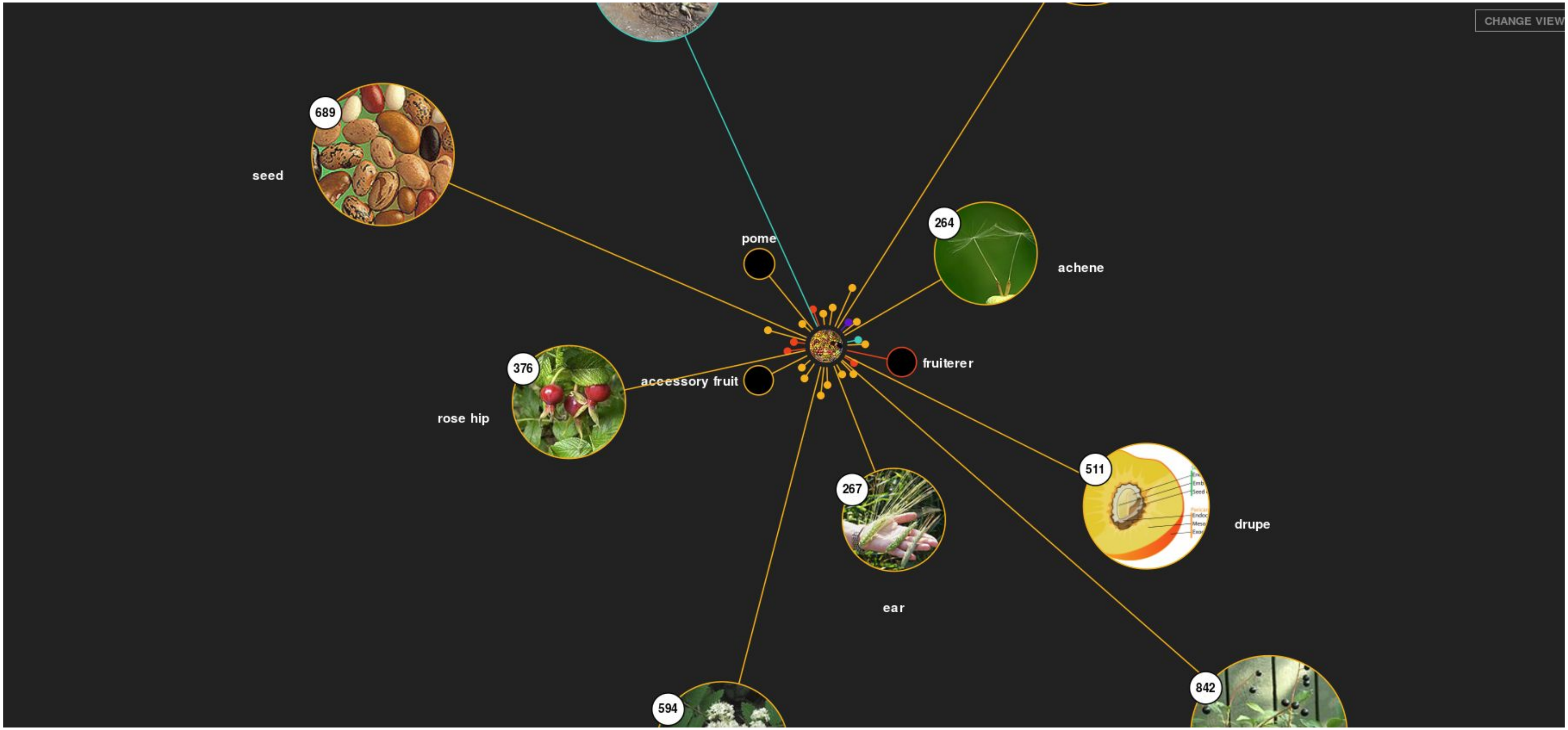
BabelNet

BabelNet

(Navigli and Ponzetto, AIJ 2012)

Thanks to an automatic mapping algorithm, it merges Wikipedia and WordNet, among other resources (Wiktionary, OmegaWiki, WikiData, VerbNet, FrameNet)

BabelNet as a very large semantic network (13.8M synsets and 380M relations)



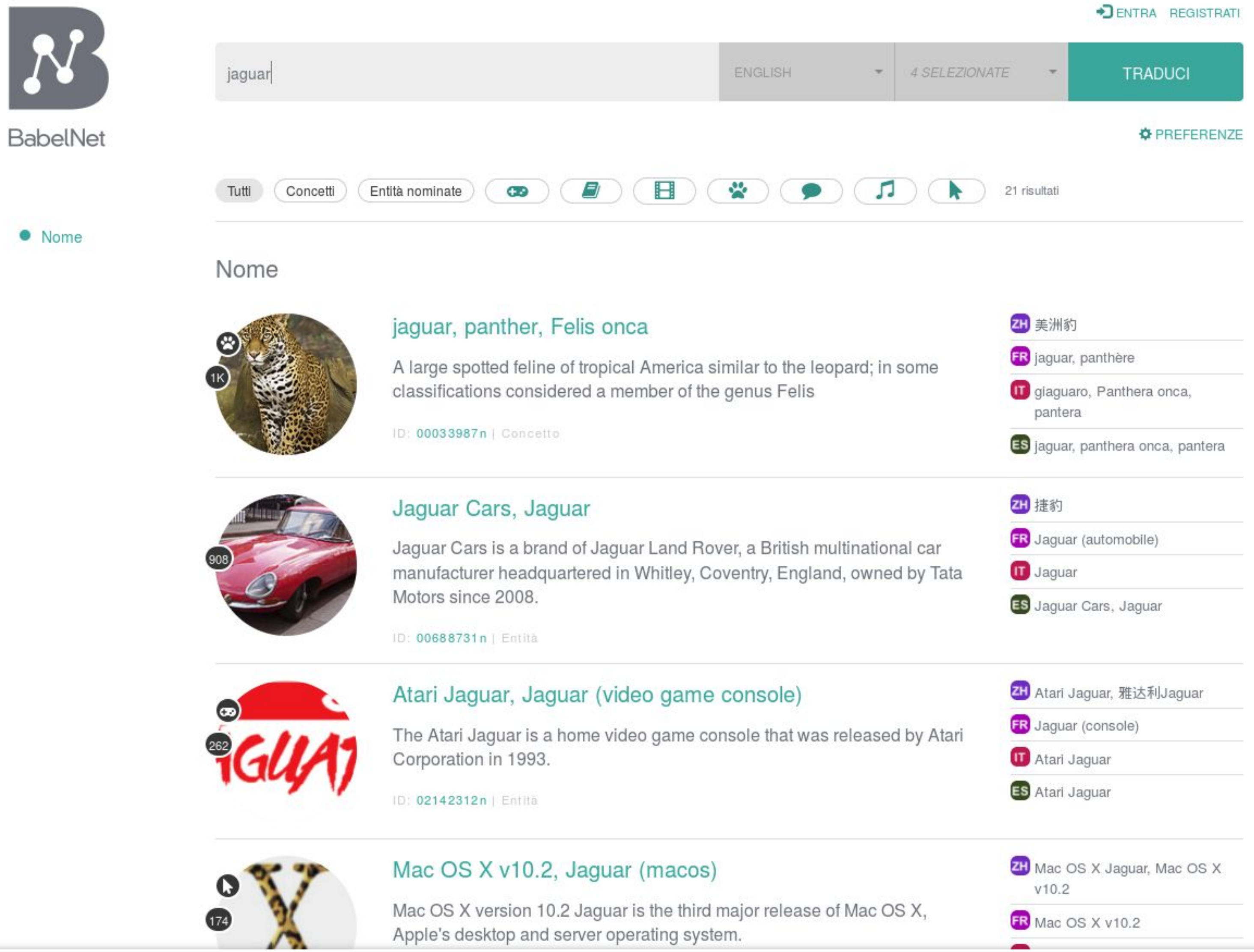
BabelNet

(Navigli and Ponzetto, AIJ 2012)

Other features:

- **Multilinguality:** 270+ languages
- Integration of **encyclopedic** (named entities) and **lexicographic knowledge** (concepts)
- Synsets associated with **images, domains, definitions, examples,** etc.

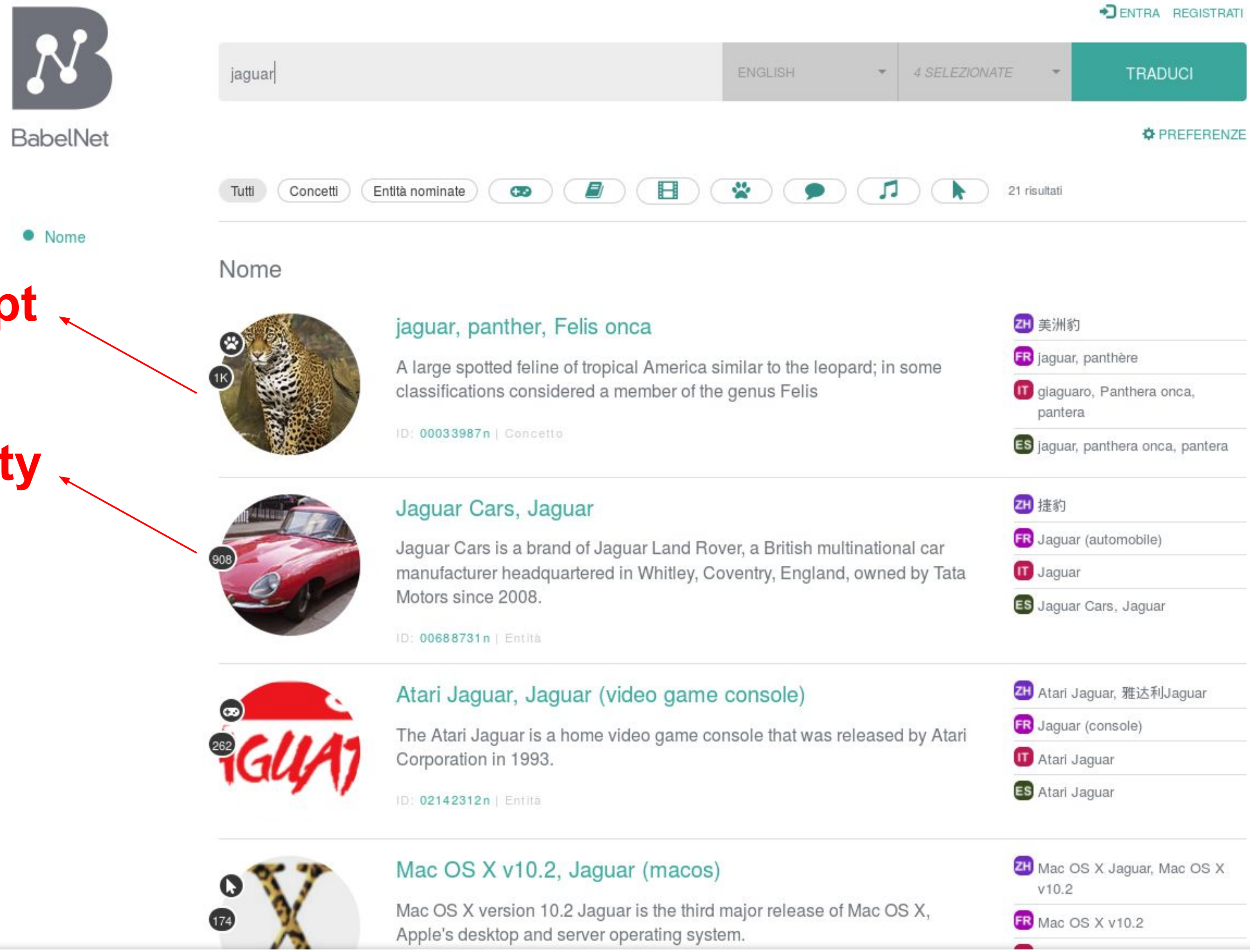
BabelNet


BabelNet
ENTRA REGISTRATI
jaguar
ENGLISH
4 SELEZIONATE
TRADUCI
PREFERENZE
Tutti
Concetti
Entità nominate
21 risultati
Nome
Nome
jaguar, panther, Felis onca
A large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis
ID: 00033987n | Concetto
1K
ZH 美洲豹
FR jaguar, panthère
IT giaguaro, Panthera onca, pantera
ES jaguar, panthera onca, pantera
Jaguar Cars, Jaguar
Jaguar Cars is a brand of Jaguar Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by Tata Motors since 2008.
ID: 00688731n | Entità
908
ZH 捷豹
FR Jaguar (automobile)
IT Jaguar
ES Jaguar Cars, Jaguar
Atari Jaguar, Jaguar (video game console)
The Atari Jaguar is a home video game console that was released by Atari Corporation in 1993.
ID: 02142312n | Entità
262
ZH Atari Jaguar, 雅达利Jaguar
FR Jaguar (console)
IT Atari Jaguar
ES Atari Jaguar
Mac OS X v10.2, Jaguar (macos)
Mac OS X version 10.2 Jaguar is the third major release of Mac OS X, Apple's desktop and server operating system.
174
ZH Mac OS X Jaguar, Mac OS X v10.2
FR Mac OS X v10.2

BabelNet

Concept

Entity

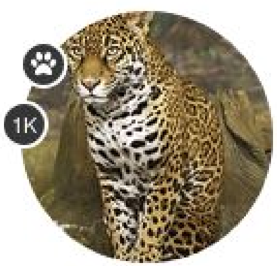


BabelNet

It follows the same structure of WordNet:

synsets are the main units

Nome



jaguar, panther, Felis onca

A large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis

ID: 00033987n | Concetto

ZH 美洲豹

FR jaguar, panthère

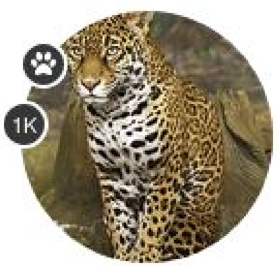
IT giaguaro, Panthera onca, pantera

ES jaguar, panthera onca, pantera

BabelNet

In this case, **synsets are multilingual**

Nome



jaguar, panther, Felis onca

A large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis

ID: 00033987n | Concetto

ZH 美洲豹

FR jaguar, panthère

IT giaguaro, Panthera onca, pantera

ES jaguar, panthera onca, pantera

Knowledge-based sense representations exploiting Wikipedia and BabelNet

- **NASARI** (Camacho-Collados et al.; NAACL and ACL 2015, AIJ 2016)

- **SensEmbed** (Iacobacci et al. ACL 2015)

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Goal

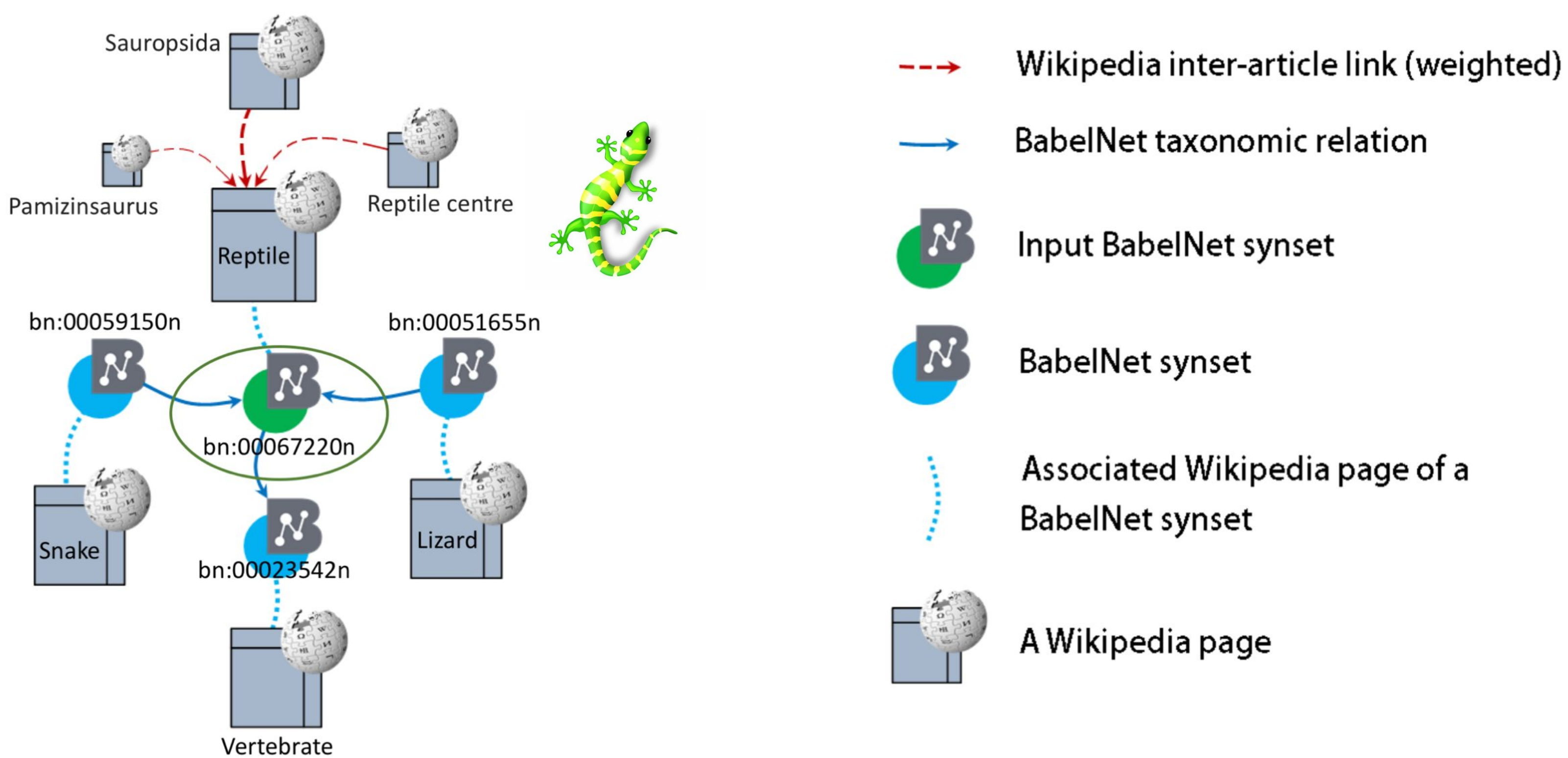
Build vector representations for multilingual BabelNet synsets.

How?

It exploits **Wikipedia semantic network** and the **WordNet taxonomy** to construct a subcorpus contextual information for any given BabelNet synset.

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)



Process of obtaining contextual information for a BabelNet synset exploiting BabelNet taxonomy and Wikipedia as a semantic network

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words)
- **Unified** (dimensions are multilingual BabelNet synsets)
- **Embedded** (latent dimensions)

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words): Dimensions are weighted via **lexical specificity** (statistical measure based on the hypergeometric distribution)

- **Unified** (dimensions are multilingual BabelNet synsets): This representation uses a **hypernym-based clustering technique** and can be used in **cross-lingual** applications

- **Embedded** (latent dimensions)

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words): Dimensions are weighted via **lexical specificity** (statistical measure based on the hypergeometric distribution)

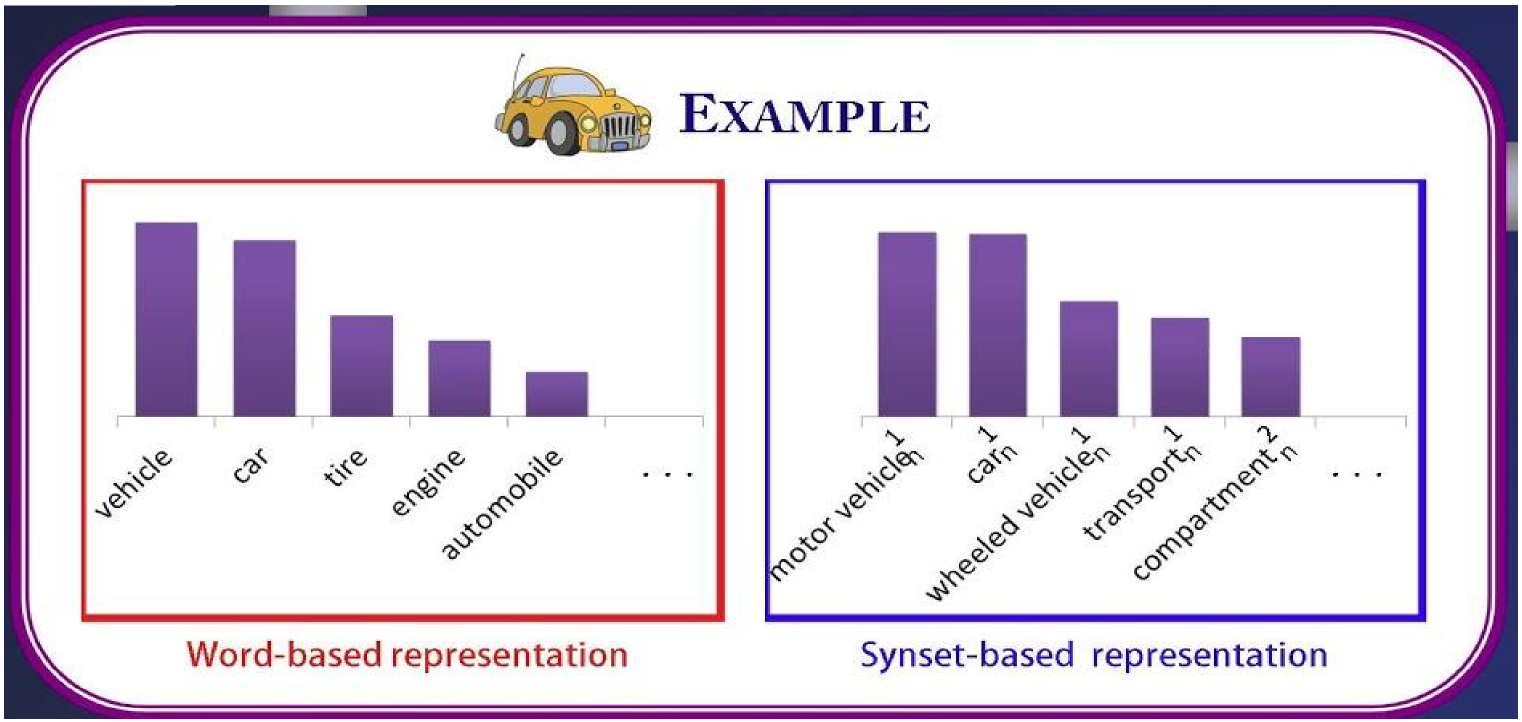
- **Unified** (dimensions are multilingual BabelNet synsets): This representation uses a **hypernym-based clustering technique** and can be used in **cross-lingual** applications

- **Embedded** (latent dimensions)

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

Interpretable dimensions

(Camacho-Collados et al., AIJ 2016)



NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words)
- **Unified** (dimensions are multilingual BabelNet synsets)

- **Embedded**: Low-dimensional vectors (latent) exploiting **word embeddings** obtained from text corpora. This representation is obtained by plugging word embeddings on the lexical vector representations.

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words)
- **Unified** (dimensions are multilingual BabelNet synsets)

- **Embedded**: Low-dimensional vectors (latent) exploiting **word embeddings** obtained from text corpora. This representation is obtained by plugging word embeddings on the lexical vector representations.

Word and synset embeddings share the same vector space!

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

High coverage of concepts and named entities in several languages (covers all Wikipedia pages).

Useful for **multilingual and cross-lingual semantic similarity**, **Sense Clustering, Domain Labeling and Word Sense Disambiguation**.

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

English	r	ρ	French	r	ρ	German	r	ρ	Spanish	r	ρ
NASARI	0.81	0.78	NASARI	**0.82**	0.73	NASARI	0.69	0.65	NASARI	**0.85**	0.79
$\text{NASARI}_{\text{lexical}}$	0.80	0.78	$\text{NASARI}_{\text{lexical}}$	0.80	0.70	$\text{NASARI}_{\text{lexical}}$	0.69	0.67	$\text{NASARI}_{\text{lexical}}$	**0.85**	0.79
$\text{NASARI}_{\text{unified}}$	0.80	0.76	$\text{NASARI}_{\text{unified}}$	**0.82**	**0.76**	$\text{NASARI}_{\text{unified}}$	**0.71**	**0.68**	$\text{NASARI}_{\text{unified}}$	0.82	0.77
$\text{NASARI}_{\text{embed}}$	**0.82**	**0.80**	–	–	–	–	–	–	$\text{NASARI}_{\text{embed}}$	0.79	0.77
SOC-PMI	0.61	–	SOC-PMI	0.19	–	SOC-PMI	0.27	–	–	–	–
PMI	0.41	–	PMI	0.34	–	PMI	0.40	–	–	–	–
LSA-Wiki	0.65	0.69	LSA-Wiki	0.57	0.52	–	–	–	–	–	–
Wiki-wup	0.59	–	–	–	–	Wiki-wup	0.65	–	–	–	–
Word2Vec	–	0.73	Word2Vec	–	0.47	Word2Vec	–	0.53	Best-Word2Vec	0.80	**0.80**
Retrofitting	–	0.77	Retrofitting	–	0.61	Retrofitting	–	0.60	–	–	–
$\text{NASARI}_{\text{poly-embed}}$	0.74	0.77	$\text{NASARI}_{\text{poly-embed}}$	0.60	0.69	$\text{NASARI}_{\text{poly-embed}}$	0.46	0.52	$\text{NASARI}_{\text{poly-embed}}$	0.68	0.74
Polyglot-embed	0.51	0.55	Polyglot-embed	0.38	0.35	Polyglot-embed	0.18	0.15	Polyglot-embed	0.51	0.56
IAA	0.85°	-	IAA	-	-	IAA	0.81	-	IAA	0.83	-

Multilingual Word Similarity

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Measure	EN-FR		EN-DE		EN-ES		FR-DE		FR-ES		DE-ES		Average	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
NASARI$_{\text{unified}}$	**0.84**	0.79	**0.79**	0.79	**0.84**	0.82	0.75	0.70	**0.86**	0.78	**0.81**	**0.80**	**0.82**	0.78
CL-MSR-2.0	0.30	–	–	–	–	–	–	–	–	–	–	–	–	–
NASARI$_{pivot}$	0.79	0.69	0.78	0.76	0.80	0.74	**0.79**	0.70	0.80	0.67	0.72	0.68	0.78	0.71
ADW$_{pivot}$	0.80	0.82	0.73	**0.82**	0.78	**0.84**	0.72	**0.77**	0.81	**0.81**	0.68	0.72	0.75	**0.80**
Word2Vec$_{pivot}$	0.77	0.82	0.70	0.73	0.76	0.80	0.65	0.70	0.75	0.76	0.64	0.63	0.71	0.74
Best-Word2Vec$_{pivot}$	0.75	**0.84**	0.69	0.76	0.75	0.82	0.77	0.73	0.74	0.79	0.64	0.64	0.72	0.76
Best-PMI-SVD$_{pivot}$	0.76	0.76	0.72	0.74	0.77	0.77	0.65	0.69	0.76	0.74	0.62	0.61	0.71	0.72

Cross-lingual Word Similarity

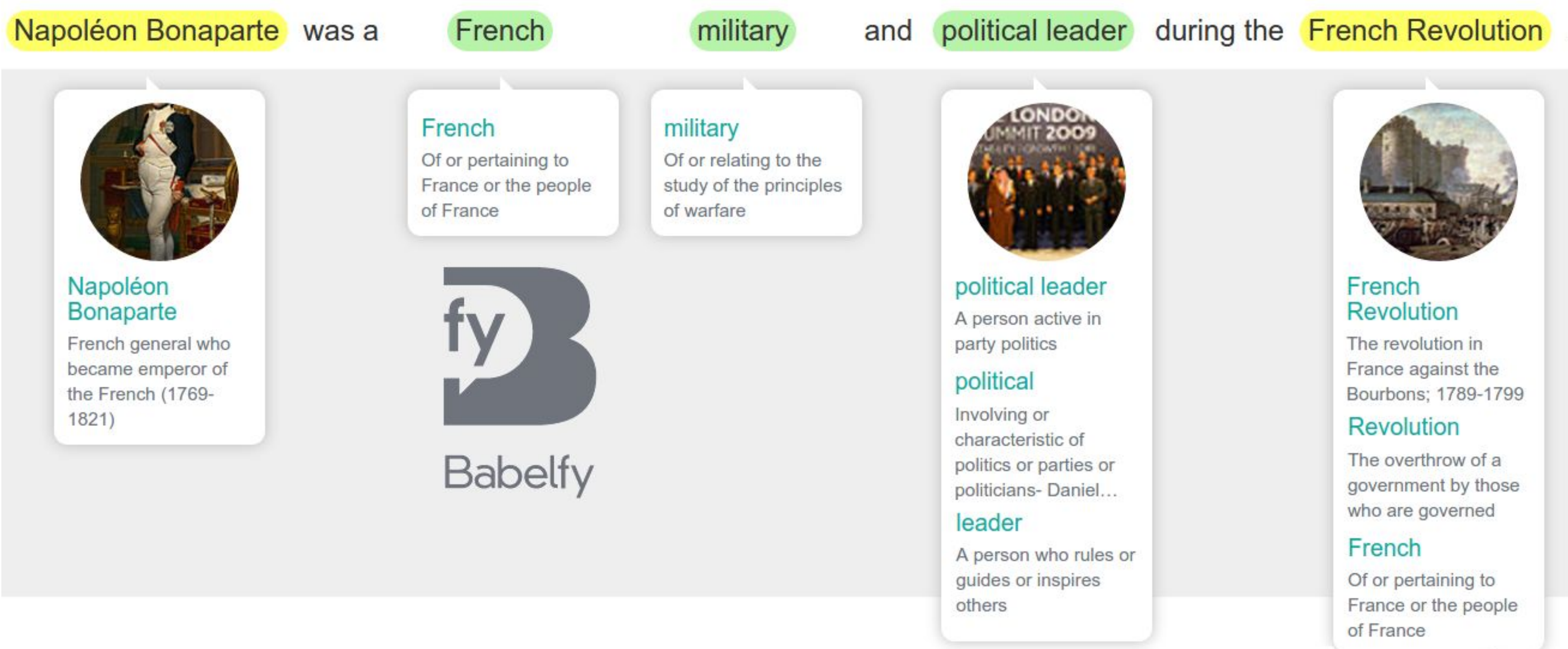
SensEmbed (Iacobacci et al., ACL 2015)

It leverages **BabelNet** and **Word2Vec** to build sense embeddings. Two steps:

- First, it uses **Babelfy** (Moro et al., TACL 2014), a multilingual joint disambiguation and entity linking system, to disambiguate a corpus.

Babelfy (Moro et al. TACL 2014)

Disambiguation and Entity Linking



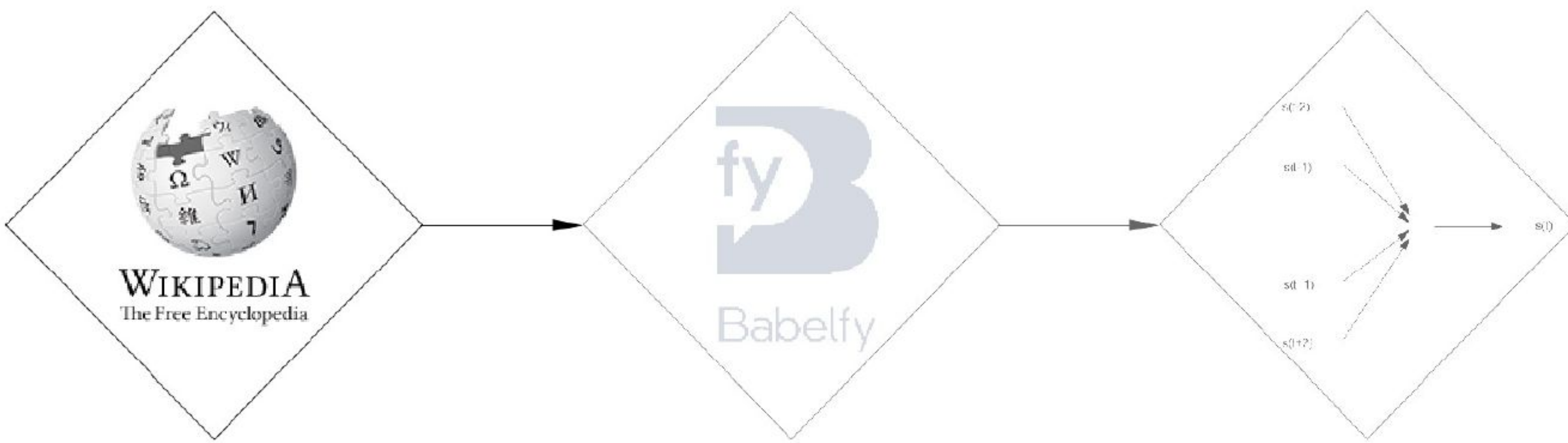
SensEmbed (Iacobacci et al., ACL 2015)

It leverages **BabelNet** and **Word2Vec** to build sense embeddings. Two steps:

- First, it uses **Babelfy** (Moro et al., TACL 2014), a multilingual joint disambiguation and entity linking system, to disambiguate a corpus.
- Then, it uses **Word2Vec** to learn sense embeddings from the sense-annotated corpus.

SensEmbed (Iacobacci et al., ACL 2015)

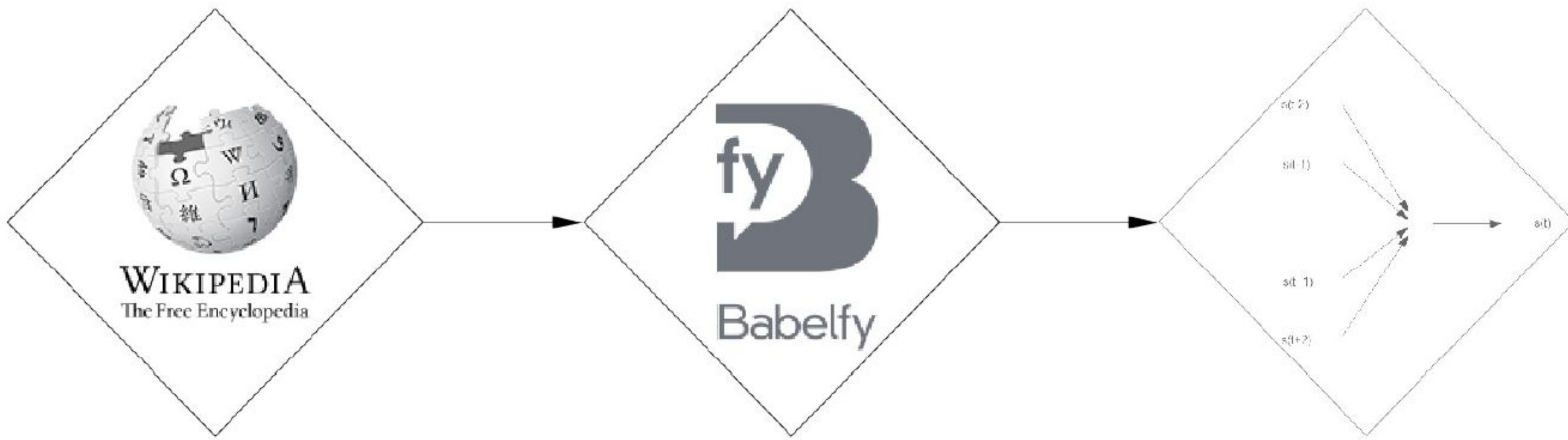
SensEmbed construction




```
...survey on the relationship between the banks and our industry , in preparation for a forthcoming forum.
             ...and it stands on the right bank  of the Drava River , bounded by the river to the north...
         ... If you have dividend or receive bank  or building society interest on which tax has been paid ,
   ...workplaces and unions. Corporations, banks and trusts controlled a great deal and , although machines...
           ...The critical decision for the banks will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth banks with trees on top. The heavily wooded area was criss-crossed...
```

SensEmbed (Iacobacci et al., ACL 2015)

SensEmbed construction



...survey on the relationship between the **banks** and our industry , in preparation for a forthcoming forum.
...and it stands on the right **bank** of the Drava River , bounded by the river to the north...
... If you have dividend or receive **bank** or building society interest on which tax has been paid ,
...workplaces and unions. Corporations, **banks** and trusts controlled a great deal and , although machines...
...The critical decision for the **banks** will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth **banks** with trees on top. The heavily wooded area was criss-crossed...

SensEmbed (Iacobacci et al., ACL 2015)

SensEmbed construction

...survey on the relationship between the **banks** and our industry , in preparation for a forthcoming forum.
...and it stands on the right **bank** of the Drava River , bounded by the river to the north...
... If you have dividend or receive **bank** or building society interest on which tax has been paid ,
...workplaces and unions. Corporations, **banks** and trusts controlled a great deal and , although machines...
...The critical decision for the **banks** will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth **banks** with trees on top. The heavily wooded area was criss-crossed...

-2.19067 1.16642 -1.91385 -0.269672 0.712771 -0.623024 -3.20115 0.560895 0.891554 0.145258 1.26956 -0.221078 -0.0733777 2.08072 -3.30558 -0.727272 -0.902202 -1.84578 -1.38985 -0.0791954 0.989769 -1.34631 1.10242 -1.59836 -1.37341 -1.42038 0.238941 -2.98729 -0.730938 0.267584 0.0560677 -0.722721 2.23752 -2.99094 -1.45598 -0.645446 0.278277 2.28877 -0.926191 2.89934 -1.17254 1.38449 2.38617 -0.0838845 -1.80698 0.622097 0.223875 0.870654 -0.33808 -0.41957

1.16672 0.811884 -0.115492 -2.59049 -1.50286 1.2536 1.44281 0.0136615 0.131499 2.04445 -0.425782 1.29676 0.0996086 1.52687 -0.0951281 -0.715488 -0.71172 0.453871 1.08481 1.55074 0.385158 -0.116754 -0.582987 -1.56923 -0.488404 -1.07999 0.0447149 -0.733387 0.765212 2.67995 2.51105 0.192151 1.49743 2.91849 1.86901 0.23101 0.381663 1.20355 0.126758 1.57204 -0.372069 -2.45076 0.514557 -1.4028 -1.20396 0.726036 2.41265 -0.104843 2.26862 1.21729

SensEmbed (Iacobacci et al., ACL 2015)

It leverages the BabelNet semantic network and the sense embeddings for **word and relational similarity**, tasks in which SensEmbed proves to be very competitive.

SensEmbed (Iacobacci et al., ACL 2015)

Measure	Dataset					
	RG-65	WS-Sim	WS-Rel	YP-130	MEN	Average
Pilehvar et al. (2013)	0.868	0.677	0.457	0.710	0.690	0.677
Zesch et al. (2008)	0.820	—	—	0.710	—	—
Collobert and Weston (2008)	0.480	0.610	0.380	—	0.570	—
Word2vec (Baroni et al., 2014)	0.840	0.800	0.700	—	0.800	—
GloVe	0.769	0.666	0.559	0.577	0.763	0.737
ESA	0.749	—	—	—	—	—
PMI-SVD	0.738	0.659	0.523	0.337	0.726	0.695
Word2vec	0.732	0.707	0.476	0.343	0.665	0.644
SensEmbed$_{closest}$	**0.894**	0.756	0.645	**0.734**	0.779	0.769
SensEmbed$_{weighted}$	0.871	**0.812**	**0.703**	0.639	**0.805**	**0.794**

Word Similarity (Spearman correlation)

SensEmbed (Iacobacci et al., ACL 2015)

Measure	MaxDiff	Spearman
Com	45.2	0.353
PairDirection	45.2	—
RNN-1600	41.8	0.275
UTD-LDA	—	0.334
UTD-NB	39.4	0.229
UTD-SVM	34.7	0.116
PMI baseline	33.9	0.112
Word2vec	43.2	0.288
$\text{SENSEMBED}_{closest}$	**45.9**	**0.358**

Relational Similarity

SensEmbed (Iacobacci et al., ACL 2015)

It has also shown its effectiveness in **Taxonomy Learning** (Espinosa-Anke et al. AAAI, 2016) and **Open Information Extraction** (Delli Bovi et al., EMNLP 2015) tasks.

We will see more about this on the “Applications” section!

Freebase™

WIKIDATA

FreeBase

FreeBase was a **large collaborative knowledge base**.

It was finally shut down on May 2016, but the data was transferred to **WikiData**.

It is the core of the **Google Knowledge Graph**.

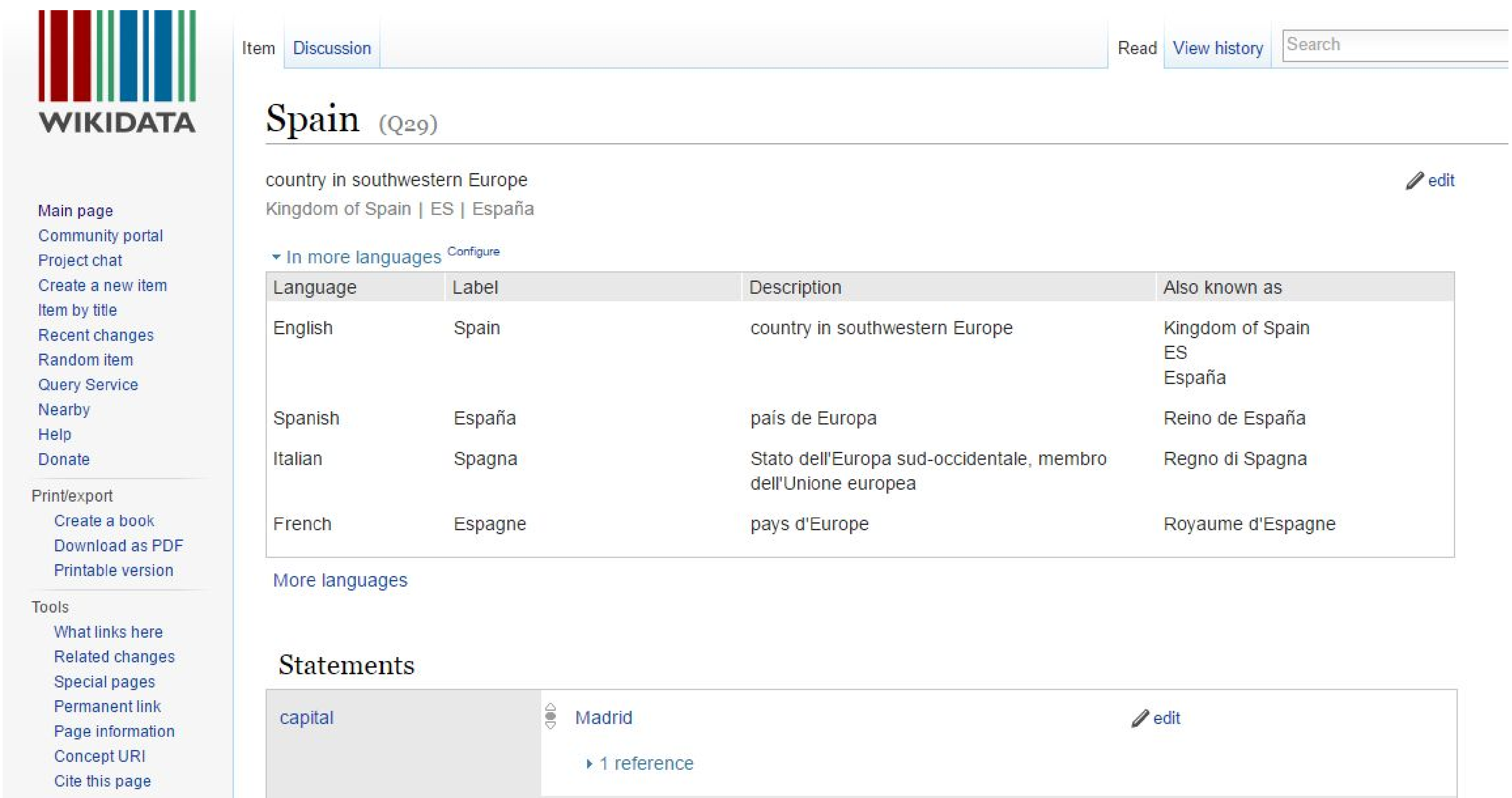
WikiData

WikiData is a large collaborative knowledge base (**18M items**).

It is based on Wikipedia and it provides a large set of relations (including a large taxonomy) among item. It exploits **Wikipedia infoboxes**.

Example: Madrid *capital of* Spain

WikiData



WIKIDATA

Main page
Community portal
Project chat
Create a new item
Item by title
Recent changes
Random item
Query Service
Nearby
Help
Donate

Print/export
Create a book
Download as PDF
Printable version

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page

Item | Discussion | Read | View history | Search

Spain (Q29)

country in southwestern Europe

edit

Kingdom of Spain | ES | España

In more languages Configure

Language	Label	Description	Also known as
English	Spain	country in southwestern Europe	Kingdom of Spain ES España
Spanish	España	país de Europa	Reino de España
Italian	Spagna	Stato dell'Europa sud-occidentale, membro dell'Unione europea	Regno di Spagna
French	Espagne	pays d'Europe	Royaume d'Espagne

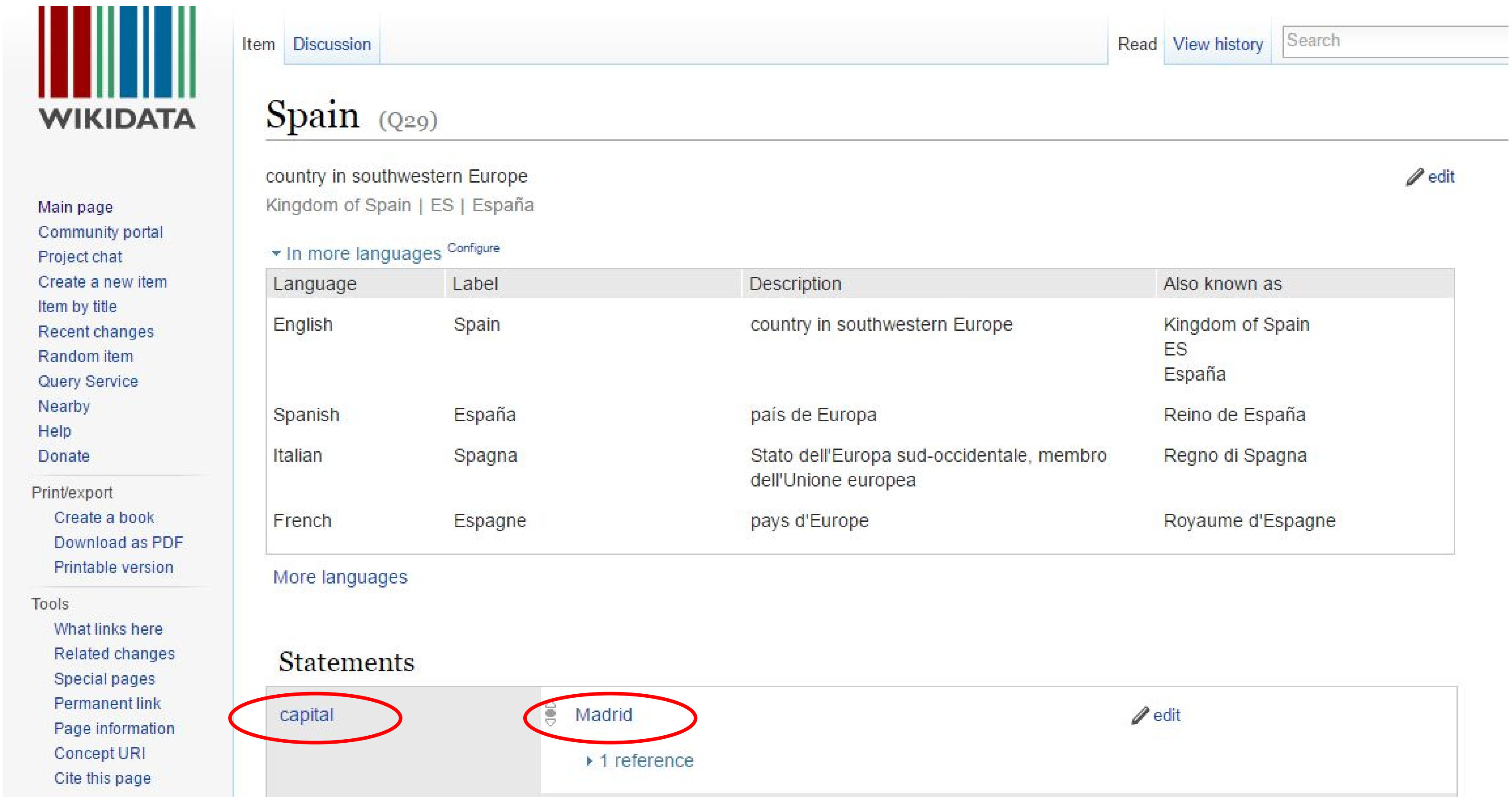
More languages

Statements

capital — Madrid — edit

1 reference

WikiData


WIKIDATA
Main page
Community portal
Project chat
Create a new item
Item by title
Recent changes
Random item
Query Service
Nearby
Help
Donate
Print/export
Create a book
Download as PDF
Printable version
Tools
What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page
Item
Discussion
Read
View history
Search
Spain (Q29)
country in southwestern Europe
edit
Kingdom of Spain | ES | España
In more languages
Configure
Language
Label
Description
Also known as
English
Spain
country in southwestern Europe
Kingdom of Spain
ES
España
Spanish
España
país de Europa
Reino de España
Italian
Spagna
Stato dell'Europa sud-occidentale, membro dell'Unione europea
Regno di Spagna
French
Espagne
pays d'Europe
Royaume d'Espagne
More languages
Statements
capital
Madrid
edit
1 reference

TransE: Translating Embeddings for Modeling Multi-relational Data

(Bordes et al., NIPS 2013)

Idea: Learn representations in a vector space not only for **entities** but also for **relations**.

TransE: Translating Embeddings for Modeling Multi-relational Data

(Bordes et al., NIPS 2013)

Relations become Translations

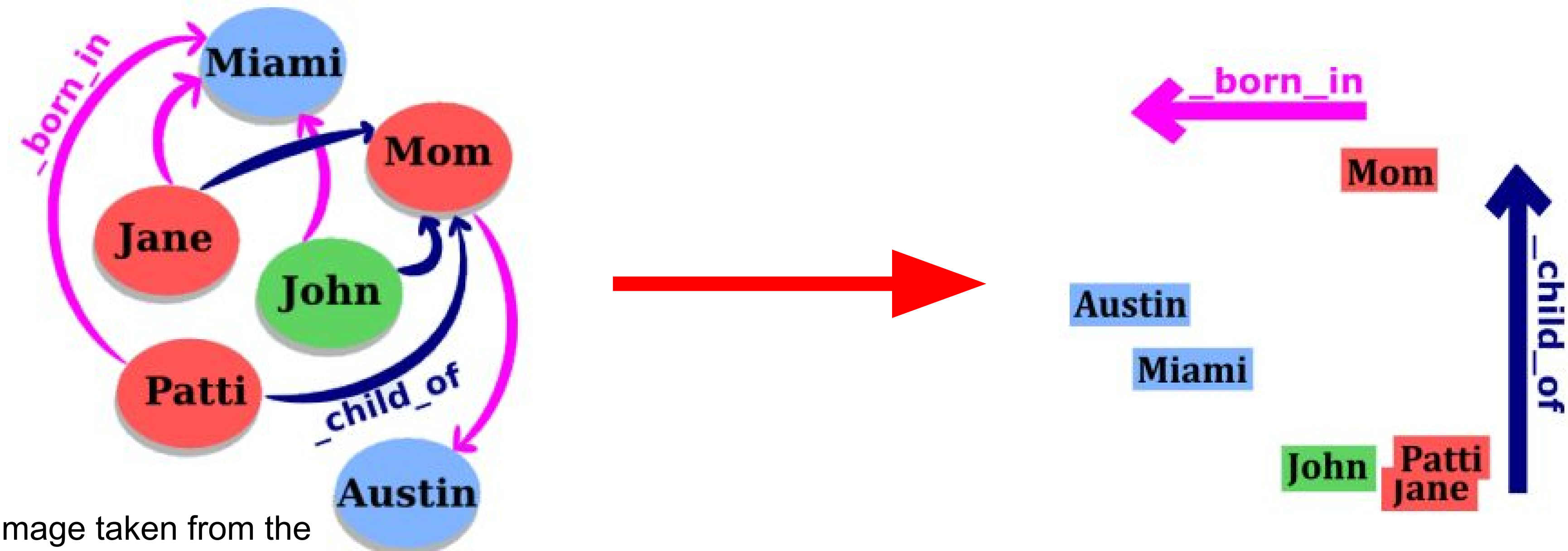


Image taken from the poster presentation of Bordes et al., [2015]

TransE: Translating Embeddings for Modeling Multi-relational Data

(Bordes et al., NIPS 2013)

Given a training set of triples (h,l,t), TransE learns embeddings for entities and their relationships by **minimizing the following loss function**:

$$\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+$$

TransE: Translating Embeddings for Modeling Multi-relational Data

(Bordes et al., NIPS 2013)

It has proved its effectiveness in learning relations in two different lexical resources: **WordNet** and **FreeBase**.

TransE: Translating Embeddings for Modeling Multi-relational Data

(Bordes et al., NIPS 2013)

New works based on the original TransE:

- **pTransE:** Joint embedding of words and entities (Wang et al., EMNLP 2014)
- **TransH:** Improving the relation mapping (Wang et al., AAAI 2014)
- **TransR:** Learning embeddings of entities and relations in separate spaces (Lin et al., AAAI 2015)
- **TransD:** Dynamic mapping for each entity-relation pair in separated spaces (Ji et al., ACL 2015)

Unsupervised sense representations

Multi-prototype Representations

Why Unsupervised?

Why do we need them?

What for?

Unsupervised Learning

"Given a set of observations [...] the goal is to directly infer the properties of this probability density without the help of a supervisor or teacher providing correct answers or degree-of-error for each observation."

Hastie, Friedman, Tibshirani, 2001

Unsupervised Learning

Most commonly-used techniques:

- **Clustering** or data segmentation has the goal of **grouping** a collection of objects into subsets or "clusters," such that those within each cluster are more **closely related**
- **Principal components** are a sequence of **projections** of features which are **mutually uncorrelated** and ordered in variance

Distributional Hypothesis

"words that occur in the same contexts tend to have similar meanings"

Harris, 1954

"a word is characterized by the company it keeps"

Firth, 1957

Unsupervised Word Sense Disambiguation

It aims to divide "the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same sense or not"

Schütze 1998

Unsupervised Word Sense Disambiguation

Main approaches

- Based on **Clustering**
- **Joint training** of multiple prototypes
- Exploiting **bilingual corpora**

Cluster-based sense representations

Cluster-based sense representations

- They are generally split in **two steps**:
  - Discrimination of senses
  - Single/Multiple prototype training

- They have a bounded (fixed) amount of prototypes

- Generally clustering considers no overlaps between clusters

Multi-Prototype Vector-Space Models of Word Meaning

Reisinger and Mooney, NAACL 2010

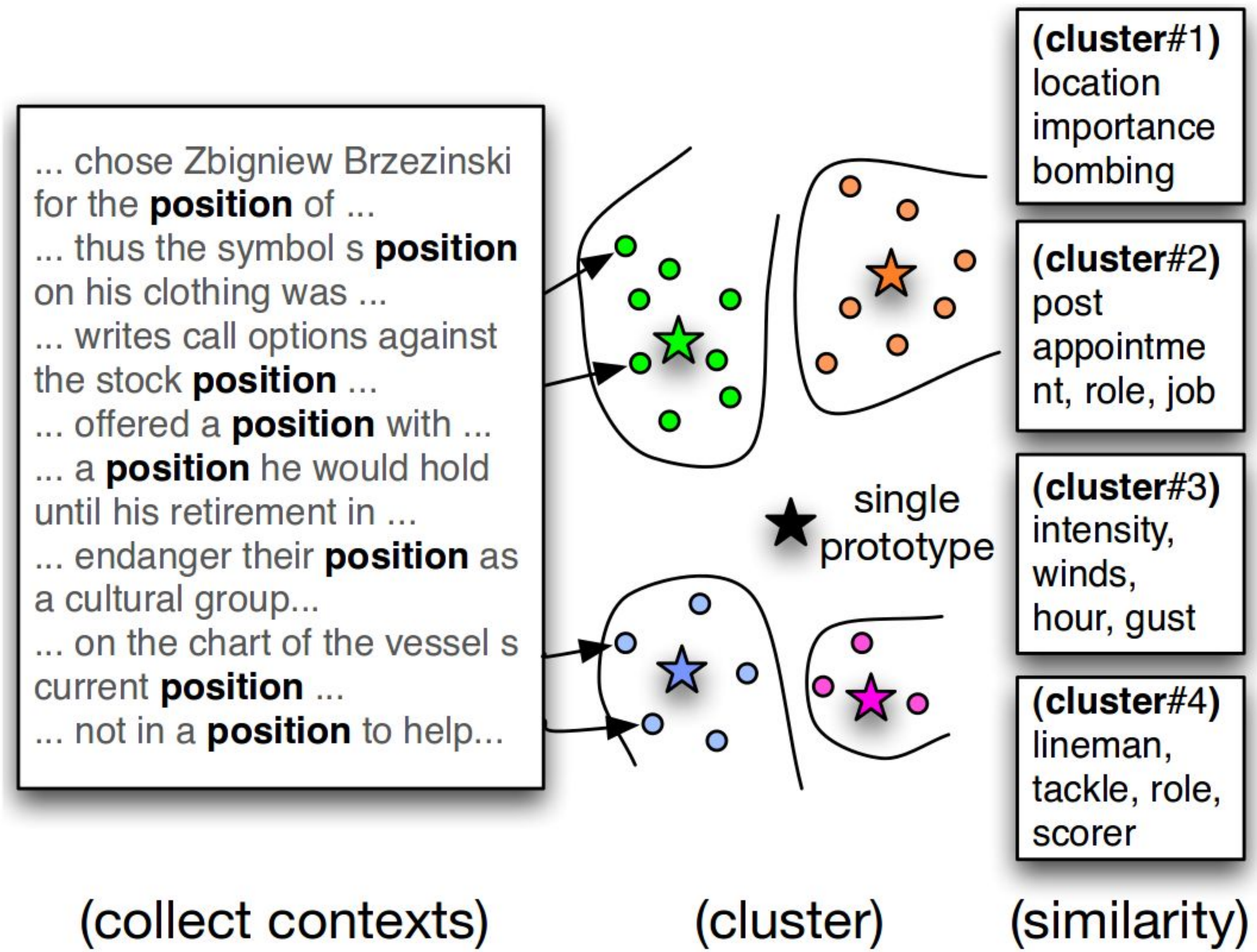
It presents a vector-space model that represents a word's meaning by a set of distinct "**sense specific**" **vectors.**

The set of vectors for a word is determined by **clustering** the contexts in which a word appears.

Explicit **feature vectors** based on unigrams

Multi-Prototype Vector-Space Models of Word Meaning

Reisinger and Mooney, NAACL 2010



Multi-Prototype Vector-Space Models of Word Meaning

Reisinger and Mooney, NAACL 2010

It measures similarity between two words, *w* and *w'*, by calculating the **minimum distance** in terms of **cosine similarity** between *w* and *w'* sense vectors:

$$\text{MaxSim}(w, w') \stackrel{\text{def}}{=} \max_{1 \leq j \leq K, 1 \leq k \leq K} d(\pi_k(w), \pi_j(w'))$$

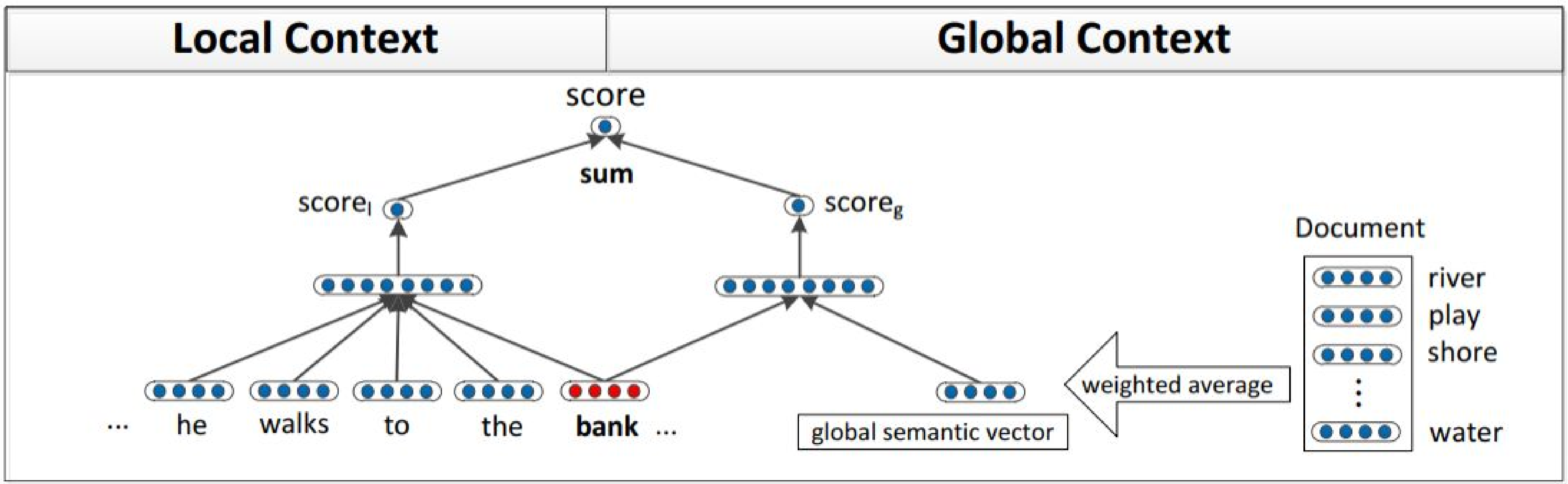
Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012

It presents a model that unlike Reisinger and Mooney, where only **local context** (i.e., co-occurrences) is used, leverages also **global context** (i.e. document topics) for learning multiple prototype vectors

Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012



Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012

Senses are represented with latent features in a 50-dimensional embedding space.

The representations are **clustered** via fixed-size context windows in order to **discriminate** the single-prototype representation into its different meanings.

Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012

Center Word	Nearest Neighbors
bank_1	corporation, insurance, company
bank_2	shore, coast, direction
star_1	movie, film, radio
star_2	galaxy, planet, moon
cell_1	telephone, smart, phone
cell_2	pathology, molecular, physiology
left_1	close, leave, live
left_2	top, round, right

Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012

It also includes a **new dataset** for measuring Multi-Prototype representations that has become the ***de facto*** evaluation for sense-based representations: Stanford Contextual Word Similarity or **SCWS**.

Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012

Model	$\rho \times 100$
C&W-S	57.0
Our Model-S	58.6
Our Model-M AvgSim	62.8
Our Model-M AvgSimC	**65.7**
tf-idf-S	26.3
Pruned *tf-idf*-S	62.5
Pruned *tf-idf*-M AvgSim	60.4
Pruned *tf-idf*-M AvgSimC	60.5

Reisinger and Mooney, 2010

Sense-Aware Semantic Analysis: A Multi-Prototype Word Representation Model Using Wikipedia

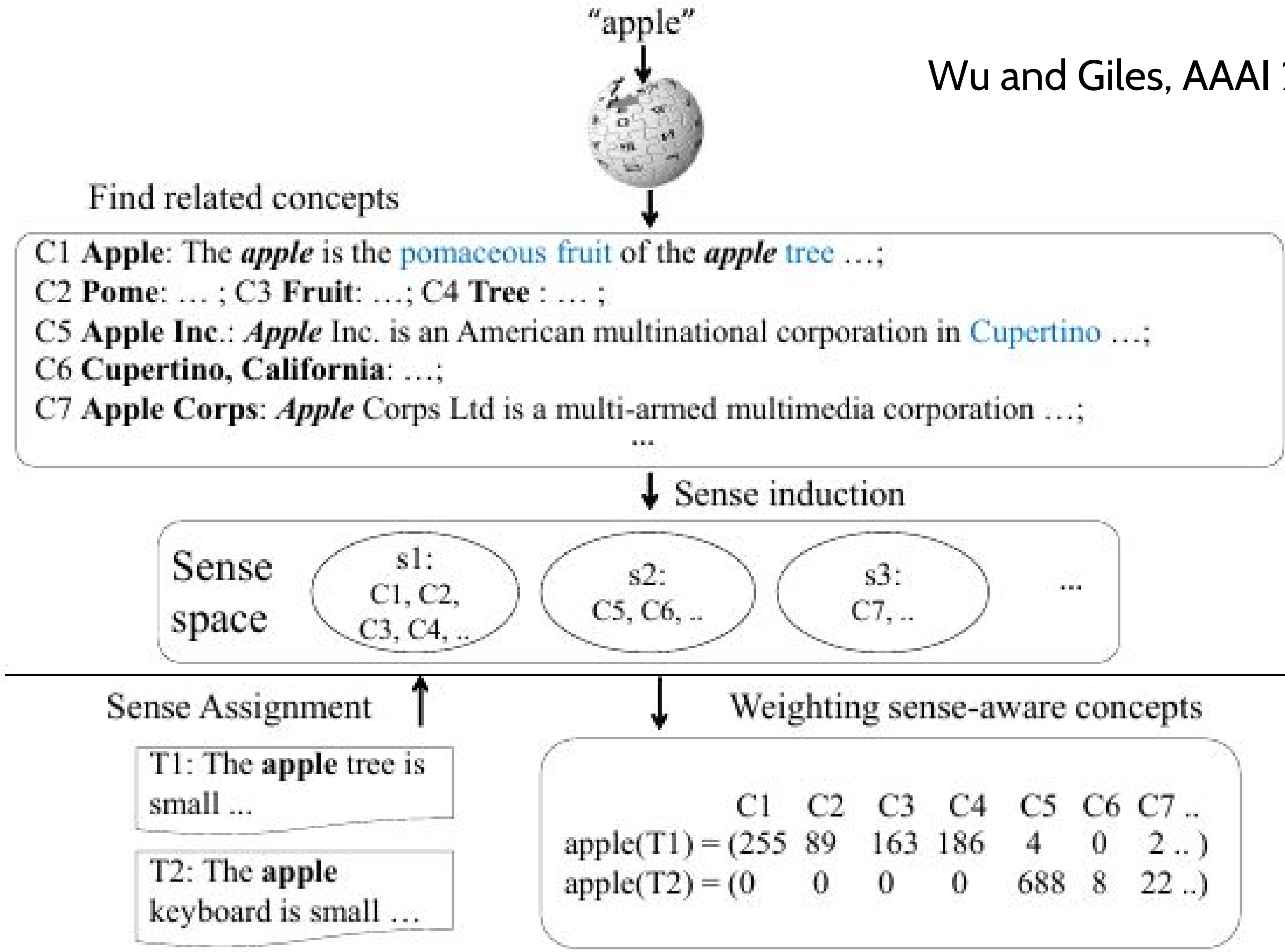
Wu and Giles, AAAI 2015

It provides "sense-specific" prototypes of a word by **clustering Wikipedia pages** based on both **local** (i.e. co-occurrences) and **global contexts** (i.e. links and categories) of the word in Wikipedia.

Each dimension of the vector space is a Wikipedia concept or article where a word appears or co-occurs with.

Sense-Aware Semantic Analysis: A Multi-Prototype Word Representation Model Using Wikipedia

Wu and Giles, AAAI 2015



Sense-Aware Semantic Analysis: A Multi-Prototype Word Representation Model Using Wikipedia

Wu and Giles, AAAI 2015

Model	ρ
ESA	0.518
SSA	0.509
Pruned tfidf-M	0.605
Huang et al. 2012	0.657
$SaSA_1$	**0.662**
$SaSA_K$	**0.664**

Reisinger and Mooney, 2010

K-Embeddings: Learning Conceptual Embeddings for Words using Context

Vu and Parker, NAACL 2016

It proposes an extension of **word embedding** as an iterative algorithm.

It has **latent representations** based on the chosen word embeddings model.

K-Embeddings: Learning Conceptual Embeddings for Words using Context

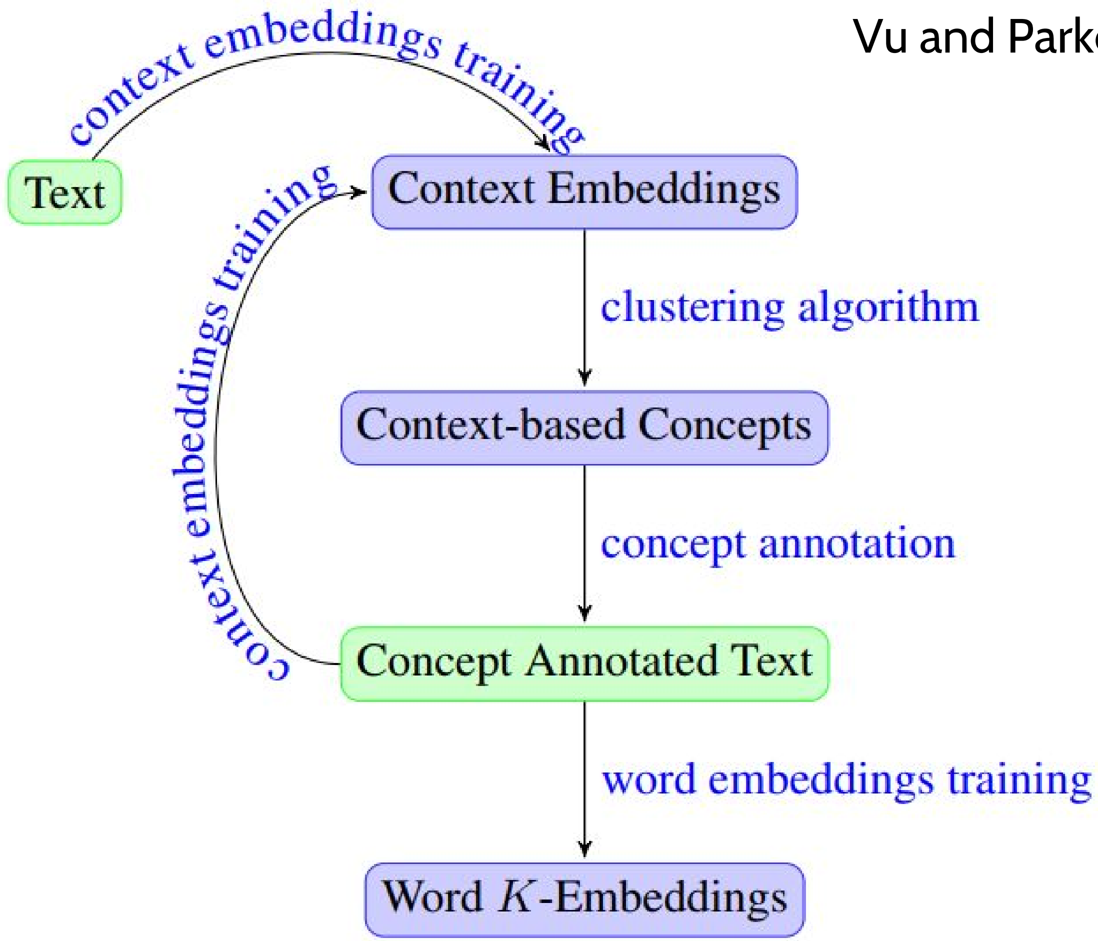
Vu and Parker, NAACL 2016

It **clusters** the **context embeddings** and uses those clusters as **sense annotations** for training sense embeddings

The resulting **annotation** could be used as **input** to **refine the clusters** (*iterative*)

K-Embeddings: Learning Conceptual Embeddings for Words using Context

Vu and Parker, NAACL 2016



K-Embeddings: Learning Conceptual Embeddings for Words using Context

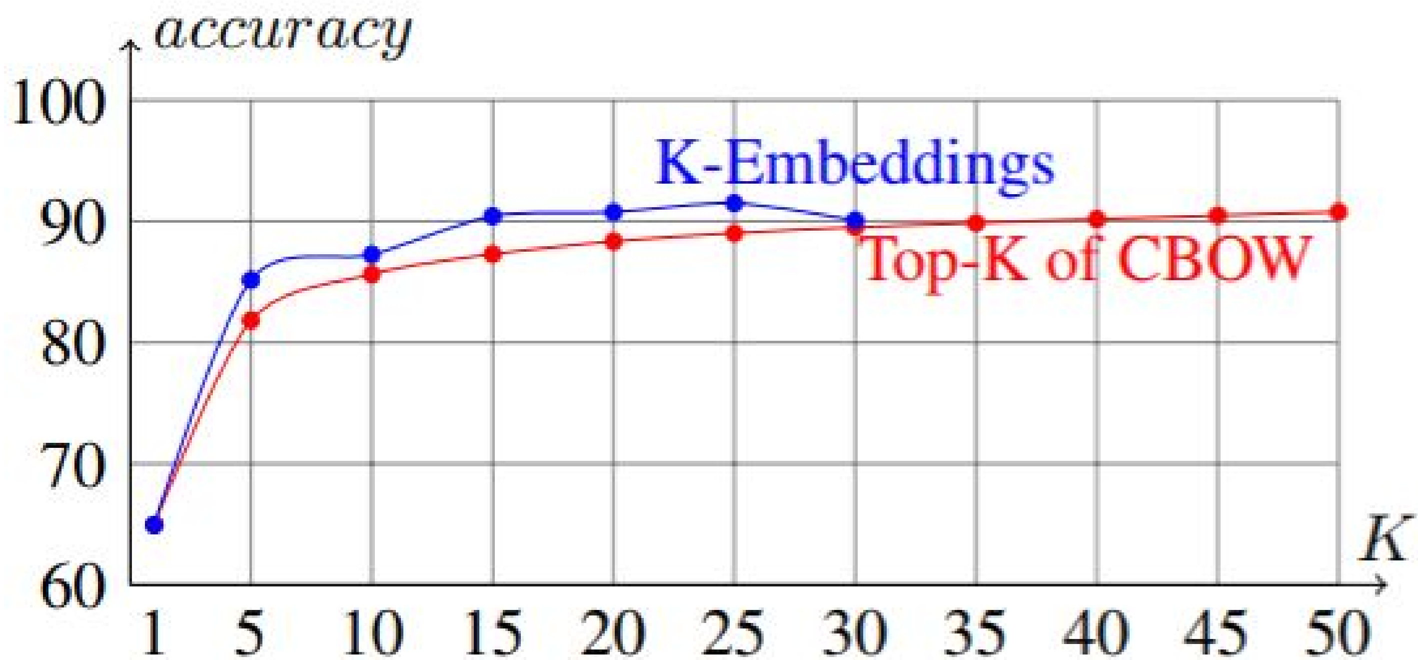
Vu and Parker, NAACL 2016

The convergence of the number of prototypes

K	total embeddings	vocabulary size	ratio
1	1,965,139	1,965,139	1.00
5	2,807,016	1,443,061	1.95
10	2,740,351	1,474,704	1.86
15	3,229,945	1,374,055	2.35
20	3,236,882	1,410,521	2.29
25	3,382,722	1,383,162	2.45
30	3,404,150	1,418,027	2.40

K-Embeddings: Learning Conceptual Embeddings for Words using Context

Vu and Parker, NAACL 2016



Accuracy on Microsoft Research Syntactic Analogies Dataset (Mikolov et al., 2013)

Joint training of sense representations

Joint training of sense representations

- The training is done in a single step
- No assumption on sense overlap (unlike cluster-based techniques)
- No assumption in the number of prototypes
- Allows to have a shared space of words and senses as an emergent behavior of the model

Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space

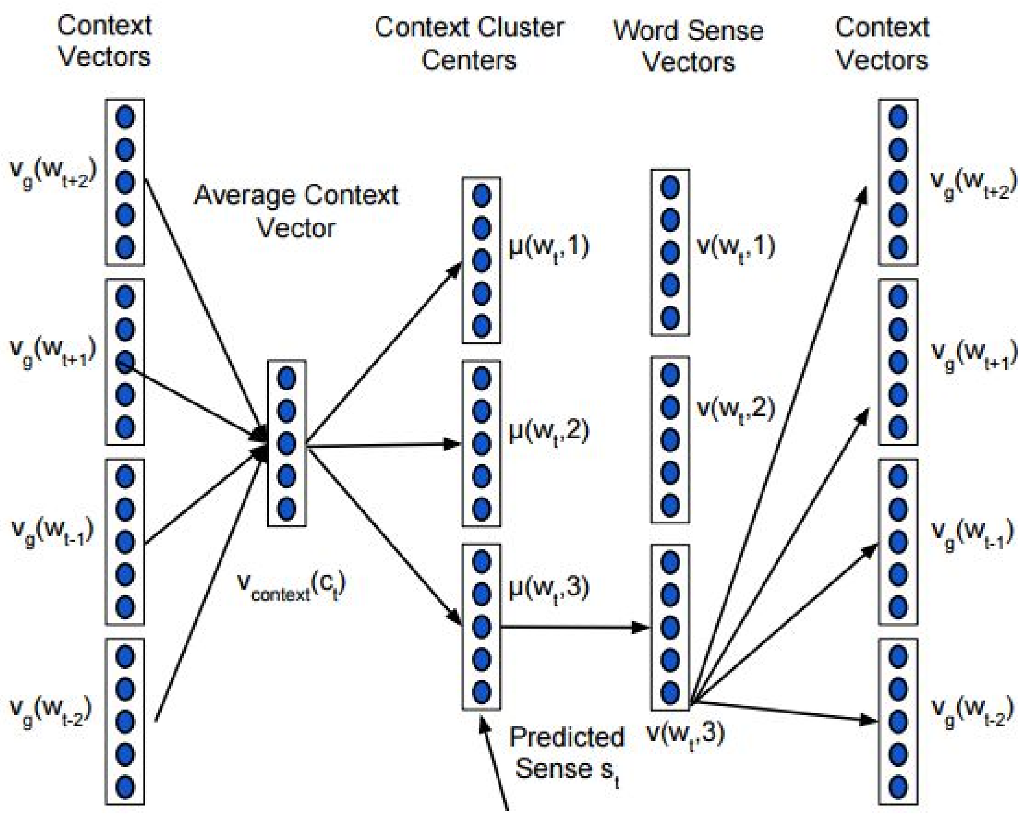
Neelakantan et al., EMNLP 2014

- An **extension** of **Skip-gram** model
- It allows to learn **multiple embeddings** per word type with **no assumptions** about the **number of senses** per word type.
- Improve the computational expense of the two-step (cluster-based) process.

Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space

Neelakantan et al., EMNLP 2014

MSSG: Multi-Sense Skip-gram



Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space

Neelakantan et al., EMNLP 2014

NP-MSSG: Multi-Sense Skip-gram

Similar to MSSG but instead of choosing across the k possible sense vectors, if the Context Cluster Center is not **similar enough** (given a threshold) a new cluster is created.

Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space

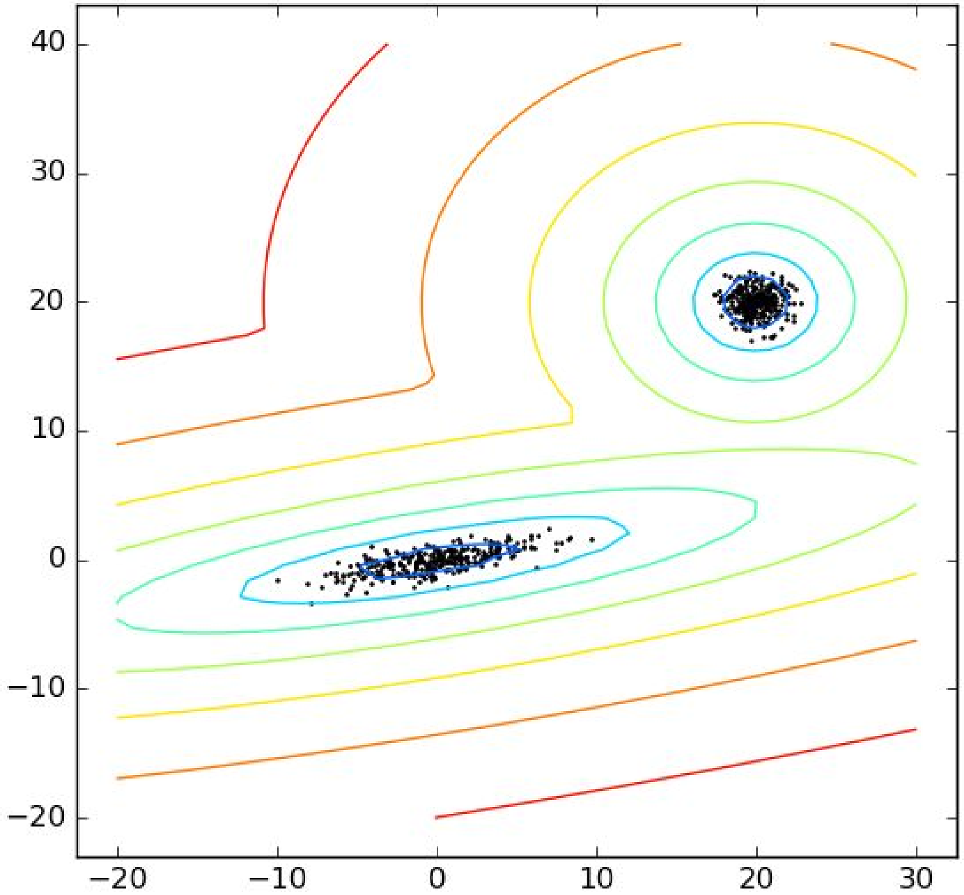
Neelakantan et al., EMNLP 2014

Model	avgSimC
Pruned TF-IDF	60.5
Huang et al-50d	65.7
MSSG-50d	66.9
MSSG-300d	**69.3**
NP-MSSG-50d	66.1
NP-MSSG-300d	69.1

A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Tian et al., COLING 2014

A Mixture model



A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Tian et al., COLING 2014

The main idea is to combine

- Skip-Gram Model
  - Provides less parameters
  - Only needs local context

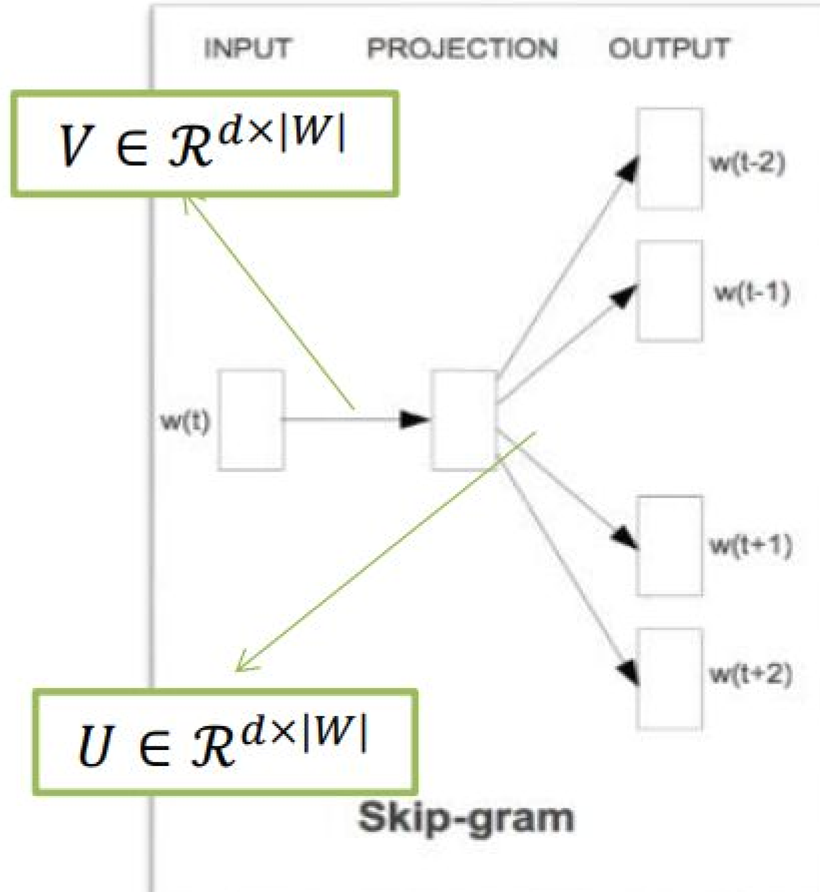
- Mixture Model
  - Provides a probabilistic framework
  - Avoid additional clustering efforts

A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Tian et al., COLING 2014

Skip-gram Model

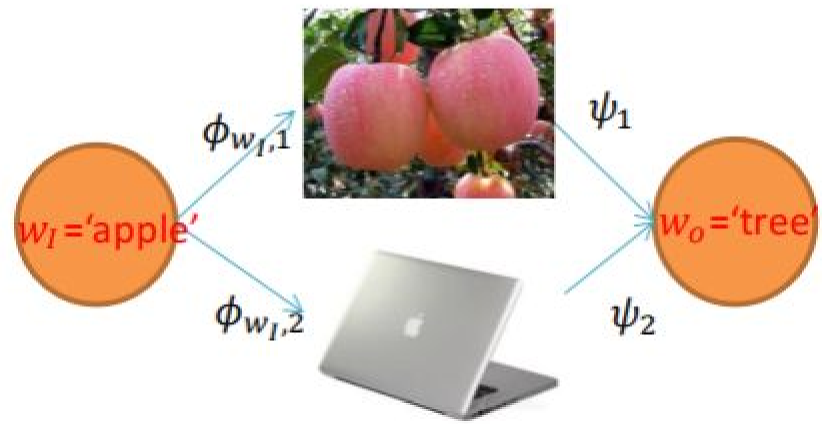
$$P(w_O|w_I) = \frac{\exp(V_{w_I}^T U_{w_O})}{\sum_{w \in \mathcal{W}} \exp(V_{w_I}^T U_w)}$$



A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Tian et al., COLING 2014

Multi-Prototype Skip-gram Model



$$p(w_O|w_I) = \sum_{i=1}^{N_{w_I}} P(w_O|h_{w_I} = i, w_I) P(h_{w_I} = i|w_I)$$

$$= \sum_{i=1}^{N_{w_I}} \frac{exp(U_{w_O}^T V_{w_I,i})}{\sum_{w \in W} exp(U_w^T V_{w_I,i})} P(h_{w_I} = i|w_I),$$

- Suppose $N_{apple} = 2$
  - $h_{apple} = 1$: 'apple' is a fruit
  - $h_{apple} = 2$: 'apple' is a company
- Denote
  $\psi_i = P(tree|h_{apple} = i, apple)$

A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Tian et al., COLING 2014

Model	$\rho \times 100$
Word2Vec	61.7
EHModel	65.7
Model_M	63.6
Model_W	65.4

Huang et al., 2012

A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Tian et al., COLING 2014

Model	EHModel	Our Model
#parameters	$dn_{words} + dn_{embeddings} + (dn_{window} + 1)h_l + (2d + 1)h_g$	$dn_{words} + dn_{embeddings}$

Topical Word Embeddings

Liu et at., AAAI 2015

It proposes a multi-prototype word embeddings model with interpretable dimensions

The dimensions are **topics**, rather than words, obtained by **latent Dirichlet allocation** (LDA)

It is based on the assumption that words will have **different embeddings** under **different topics**

Topical Word Embeddings

Liu et at., AAAI 2015

Three models

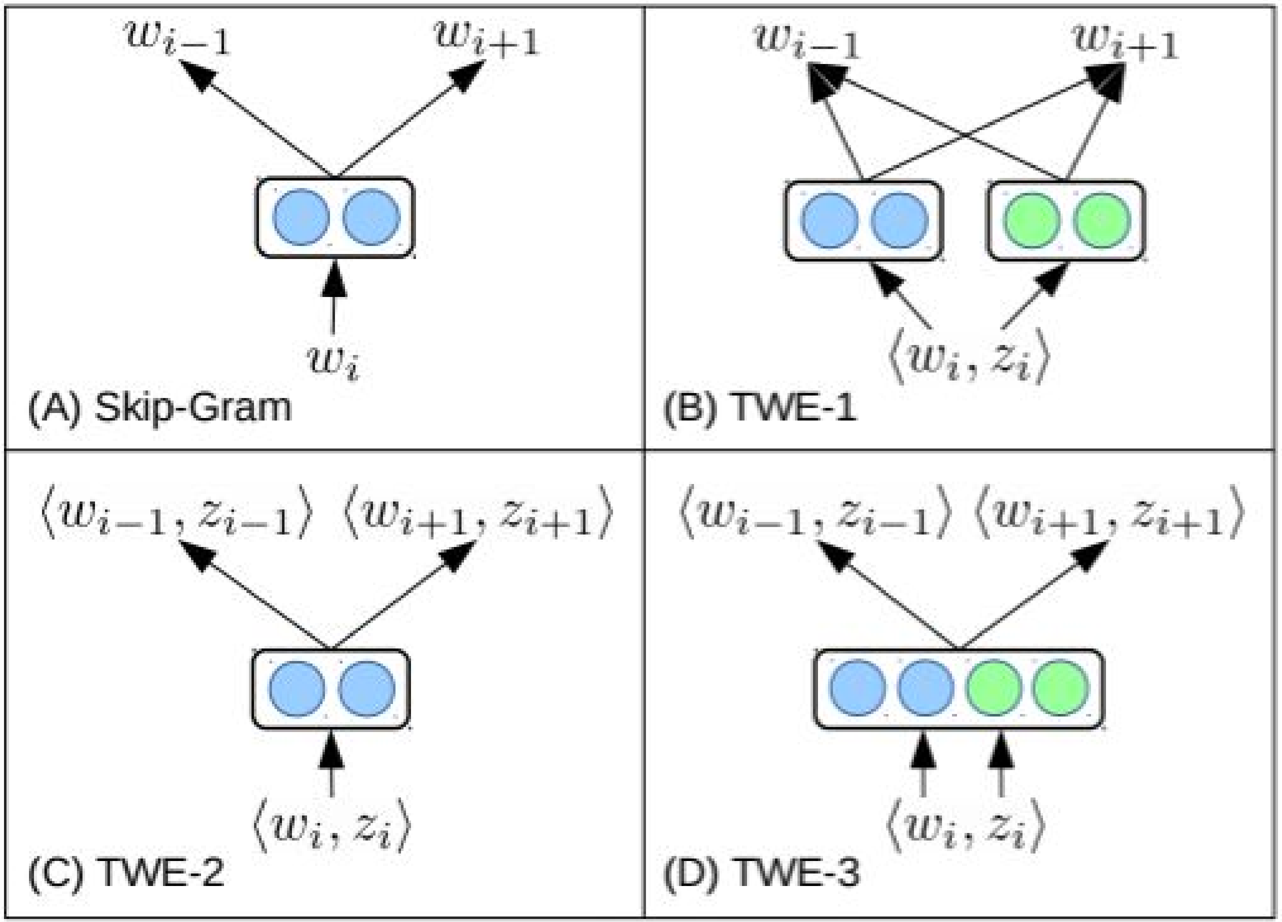
- TWE-1. Each **topic** is treated as an **extra word**. Embeddings of words and topics are **learned separately.** The topical embeddings are build with **both contributions**

- TWE-2. Each **word-topic pair** is considered as a **pseudo word**, and learn topical word embeddings directly

- TWE-3. Words and topics are **separate but learned jointly**. The embedding of each word-topic pair is the **concatenation** of **both word and topic** embeddings

Topical Word Embeddings

Liu et at., AAAI 2015



Topical Word Embeddings

Liu et at., AAAI 2015

Model	$\rho \times 100$	
C&W	57.0	
TFIDF	26.3	
Pruned TFIDF	62.5	
LDA-S	56.9	
LDA-C	50.4	
Skip-Gram	65.7	
	`AvgSimC`	`MaxSimC`
Pruned TFIDF-M	60.5	60.4
Tian	65.4	63.6
Huang	65.3	58.6
TWE-1	**68.1**	**67.3**
TWE-2	67.9	63.6
TWE-3	67.1	65.5

Do Multi-Sense Embeddings Improve Natural Language Understanding?

Li and Jurafsky, EMNLP 2015

It **criticizes multi-prototype** models by **questioning** if there is clear evidence how these models **improve** single-prototype approaches on real NLU tasks.

It introduces a multisense embeddings model based on **Chinese Restaurant Processes**

Do Multi-Sense Embeddings Improve Natural Language Understanding?

Li and Jurafsky, EMNLP 2015

Chinese restaurant process

A restaurant where a new **customer** finds **table** and is likely to choose those tables which are **more populated**.

Do Multi-Sense Embeddings Improve Natural Language Understanding?

Li and Jurafsky, EMNLP 2015

Idea:

A word is associated with a **new sense vector** just when **evidence** in the context suggests that it is sufficiently different from its early senses.

Do Multi-Sense Embeddings Improve Natural Language Understanding?

Li and Jurafsky, EMNLP 2015

Model	SCWS Correlation
SkipGram	66.4
SG+Greedy	69.1
SG+Expect	69.7
Chen	68.4
Neelakantan	69.3

Sense-based representations by exploiting bilingual resources

Sense-based representations by exploiting bilingual resources

"The other major potential source of sense-tagged data comes from **parallel aligned bilingual corpora**. Here, translation distinctions can provide a practical correlate to sense distinctions, as when instances of the English word"

Resnik & Yarowsky, 1997

Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources

Guo et al., COLING 2014

It proposes a method for learning **sense-specific word embeddings** by using **bilingual parallel data**.

It is supported by a **language model** based on **neural networks**

“same word in the source language with different senses […] has different translations in the foreign language”

Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources

Guo et al., COLING 2014

Idea:

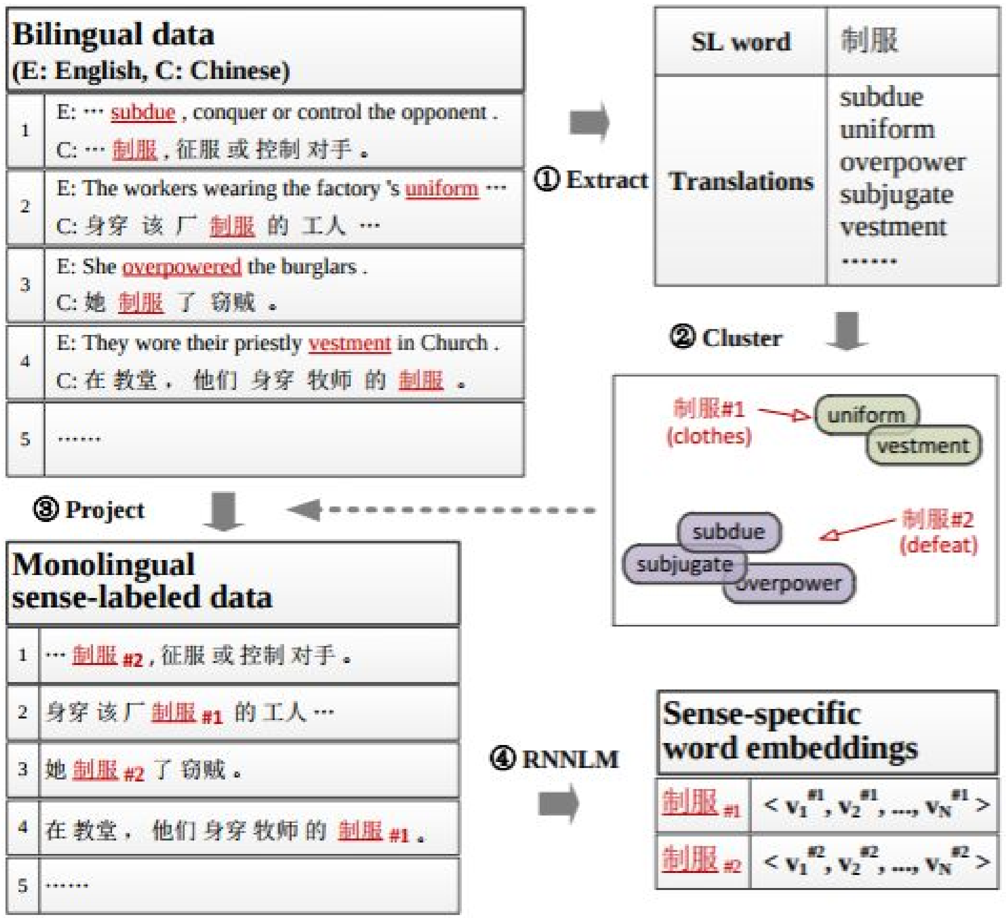
The words in the **source language** are **tagged** with their **translation** in the **foreign language**

The translations are **clustered**, exhibiting **different senses** in **different clusters**

The **sense-annotated** data is used to learn **sense-specific word embeddings**

Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources

Guo et al., COLING 2014



Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources

Guo et al., COLING 2014

System	MaxSim		AvgSim	
	$\rho \times 100$	$\tau \times 100$	$\rho \times 100$	$\tau \times 100$
Ours	**55.4**	**40.9**	49.3	35.2
SingleEmb	42.8	30.6	42.8	30.6
Multi-prototype	40.7	29.1	38.3	27.4

Spearman and Kendall correlation SemEval 2012 Task 4 EvaluatingChinese Word Similarity (Jin & Wu, 2012)

Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Šuster et al., NAACL 2016

Uses **second-language** embeddings as a supervisory signal in learning multisense representations in the first language

"**Polysemy** in one language can be at least **partially resolved** by looking at the translation of the word and its context in **another language**"

Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Šuster et al., NAACL 2016

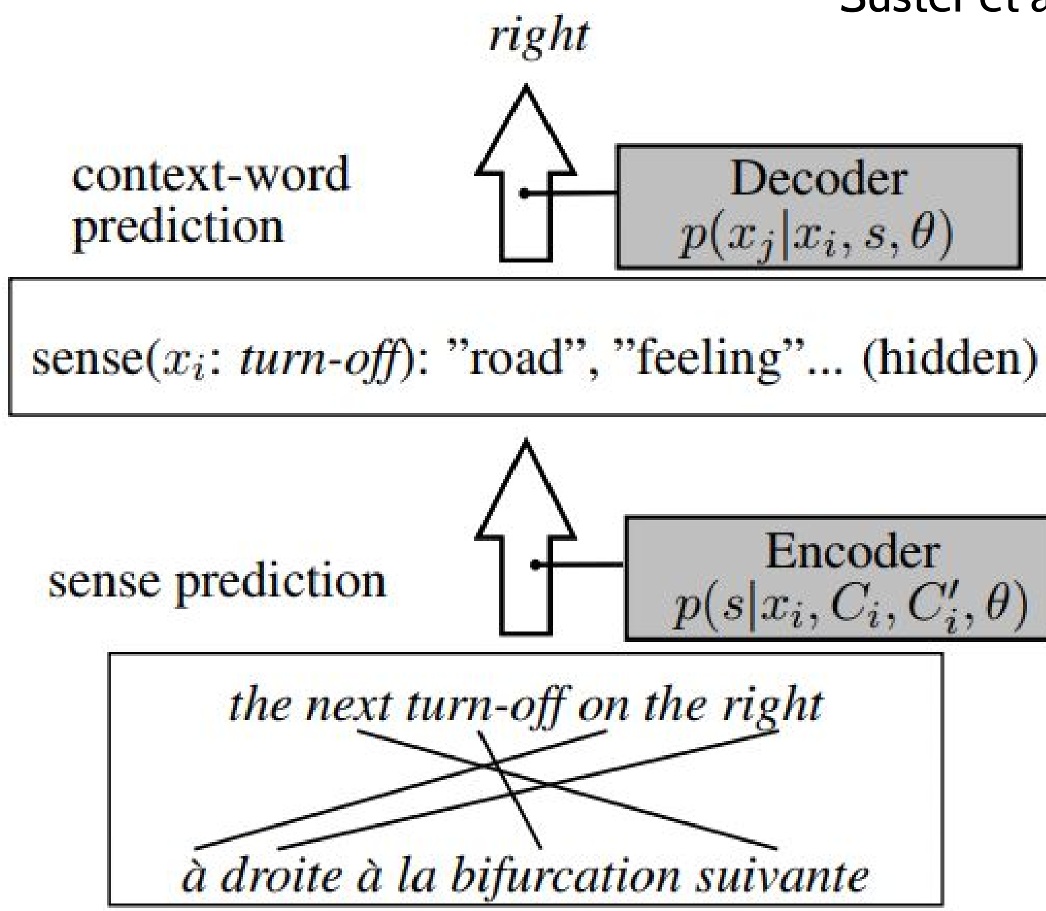
It is designed as an **autoencoder**: a feed forward neural network model that learns to **mimic** its **input layer** in the **output layer**.

Two parts:

- An encoding part which **assigns a sense** to a pivot word given the **word** and the **context in both languages**
- A reconstruction (decoding) part recovering **context words** based on the pivot **word** and its **sense**

Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Šuster et al., NAACL 2016



Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Šuster et al., NAACL 2016

Model (300-dim.)	SCWS
SG	65.0
MU	66.7
BIMU	69.0
Chen et al. (2014)	68.4
Neelakantan et al. (2014)	69.3
Li and Jurafsky (2015)	69.7

Omitting the Bilingual corpora

Advantages and limitations of both types of sense representations

Advantages and Limitations

Knowledge-based sense representations

Advantages

- The learned sense representations are **linked to sense inventories**
  - This in turn might enable **multilinguality** (see BabelNet)
  - May exploit extra-information available in the underlying resource
- The number of senses for each word varies and is decided by **expert lexicographers**

Advantages and Limitations

Knowledge-based sense representations

Disadvantages

- Require sense inventories
  - Might **not be available/complete in some languages**
- Given that sense inventories are fixed, to **cover emerging senses** the inventory needs to be updated before we can create a vector representation

Advantages and Limitations

Unsupervised sense representations

Advantages

- **Fully unsupervised** (no need for external knowledge resources) and allows to have an entire **end-to-end approach**

- Can be **adapted to specific corpora and domains**

Advantages and Limitations

Unsupervised sense representations

Disadvantages

- The learned sense representations are **not linked to any sense inventory**
- Usually assume the **number of senses to be fixed for all words**
- The representations are generally **not fine grained** and **difficult to evaluate**.
- **Rare words and less frequent meanings** are not represented properly

Applications

Applications

- Semantic Similarity *(used in other applications)*
- Word Sense Disambiguation / Entity Linking
- Link Prediction
- Ontology learning
- Information Extraction
- Sense Clustering
- Alignment of Lexical Resources

Sense-based Semantic Similarity

Based on the semantic similarity between senses.

Two main measures:

- **Cosine similarity** for low-dimensional vectors
- **Weighted Overlap** for sparse high-dimensional vectors (usually interpretable)

Sense-based Semantic Similarity: Words

Different sense-based measures as explained in the previous section.

Sense-based similarity **performs on par or better than word-based approaches**.

How to compose vectors for sentence/document representation?

Averaging word vectors is the most common approach

Drawbacks:

- **Word order** is not taken into account (new **neural network** approaches take word order into account, e.g. LSTMs)
- **Syntax** is not taken into account
- **Ambiguity** is not taken into account

How to model sentences and documents using sense representations?

There are some interesting **compositionality** ideas and approaches to test the use of sense representations to model sentences and documents: e.g. ADW or Li and Jurafsky (2015).

However, sense-based representation of sentences and documents remains an **open problem** (same applies to word-based).

Word Sense Disambiguation

Two ways to use sense representations for WSD:

- Integrated as a feature in a supervised disambiguation system (Rothe and Schütze, ACL 2015)
- Knowledge-based disambiguation (Camacho-Collados et al., ACL 2015)

Integration of sense representations in a supervised WSD system

(Rothe and Schütze, ACL 2015)

IMS (Zhong and Ng, ACL 2010 demo) is a state-of-the-art supervised disambiguation system. It is a **SVM classifier** which uses features based on the surrounding words of the target word (**local context**).

Idea: Use **word and sense embeddings of the surrounding words** and add it as a new feature.

Integration of sense representations in a supervised WSD system

(Rothe and Schütze, ACL 2015)

			Senseval-2	Senseval-3
IMS feature sets	1	POS	53.6	58.0
	2	surrounding word	57.6	65.3
	3	local collocation	58.7	64.7
	4	S_{naive}-product	56.5	62.2
	5	S-cosine	55.5	60.5
	6	S-product	58.3	64.3
	7	S-raw	56.8	63.1
system comparison	8	MFS	$47.6^{\dagger}$	$55.2^{\dagger}$
	9	Rank 1 system	$64.2^{\dagger}$	72.9
	10	Rank 2 system	$63.8^{\dagger}$	72.6
	11	IMS	$65.2^{\ddagger}$	$72.3^{\ddagger}$
	12	IMS + S_{naive}-prod.	$62.6^{\dagger}$	$69.4^{\dagger}$
	13	IMS + S-cosine	$65.1^{\ddagger}$	$72.4^{\ddagger}$
	14	IMS + S-product	**66.5**	**73.6**
	15	IMS + S-raw	$62.1^{\dagger}$	$66.8^{\dagger}$
	16	IMS + $S_{optimized}$-prod.	*66.6*	*73.6*

WSD using WordNet as sense inventory (lexical sample)

Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

Basic idea

Select the sense which is semantically closer to the semantic representation of the whole document (**global context**).

$$\hat{d}(s) = \operatorname*{argmax}_{d \in D} WO(\vec{\text{NASARI}}_{lex}(s), \vec{v}_{lex}(d))$$

Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

Kobe, which is one of Japan's largest cities, [...]

?

Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

Kobe, which is one of Japan's largest cities, [...]

X

Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

Kobe, which is one of Japan's largest cities, [...]

Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

System	English	French	Italian	German	Spanish	Average
NASARI	86.3	**76.2**	83.7	**83.2**	82.9	**82.5**
MUFFIN	84.5	71.4	81.9	83.1	**85.1**	81.2
Babelfy	**87.4**	71.6	**84.3**	81.6	83.8	81.7
UMCC-DLSI	54.8	60.5	58.3	61.0	58.1	58.5
MFS	80.2	74.9	82.2	83.0	82.1	79.3

Multilingual WSD using Wikipedia as sense inventory (all-words)

Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

System	SemEval-2013	SemEval-2007
NASARI	66.7	66.7
NASARI+IMS	67.0	**68.5**
MUFFIN	66.0	66.0
Babelfy	65.9	62.7
UKB	61.3	56.0
UMCC-DLSI	64.7	–
Multi-Objective	**72.8**	66.0
IMS	65.3	67.3
MFS	63.2	65.8

WSD using WordNet as sense inventory (All-Words)

Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

System	SemEval-2013	SemEval-2007
NASARI	66.7	66.7
NASARI+IMS	67.0	**68.5**
MUFFIN	66.0	66.0
Babelfy	65.9	62.7
UKB	61.3	56.0
UMCC-DLSI	64.7	–
Multi-Objective	**72.8**	66.0
IMS	65.3	67.3
MFS	63.2	65.8

WSD using WordNet as sense inventory (All-Words)

Word Sense Disambiguation

Open problem

Integration of **knowledge-based** (exploiting global contexts) and **supervised** (exploiting local contexts) systems to overcome the *knowledge-acquisition bottleneck*.

Link Prediction

Bordes et al. (NIPS 2013)

Add automatically relations between entities in a knowledge base.

How?

Embedding entities and relationships together -> TransE

Link Prediction

Bordes et al. (NIPS 2013)

DATASET	WN				FB15K			
METRIC	MEAN RANK		HITS@10 (%)		MEAN RANK		HITS@10 (%)	
Eval. setting	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Filt.*
Unstructured [2]	315	304	35.3	38.2	1,074	979	4.5	6.3
RESCAL [11]	1,180	1,163	37.2	52.8	828	683	28.4	44.1
SE [3]	1,011	985	68.5	80.5	273	162	28.8	39.8
SME(LINEAR) [2]	545	533	65.1	74.1	274	154	30.7	40.8
SME(BILINEAR) [2]	526	509	54.7	61.3	284	158	31.3	41.3
LFM [6]	469	456	71.4	81.6	283	164	26.0	33.1
TransE	**263**	**251**	**75.4**	**89.2**	**243**	**125**	**34.9**	**47.1**

Link Prediction results in WordNet and FreeBase

Taxonomy Learning

Espinosa-Anke et al. (AAAI 2016)

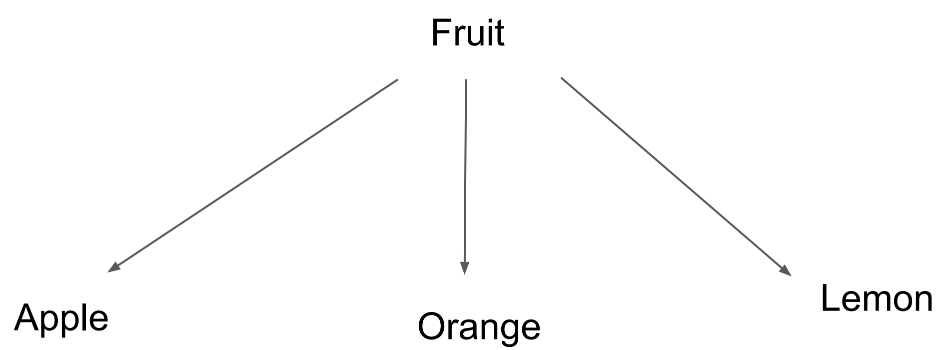
Global approach which exploits a large semantic network to **extend**, **taxonomize** and **semantify** domain terminologies.

How are sense representations used?

Taxonomy Learning

Espinosa-Anke et al. (AAAI 2016)

It uses sense representations to disambiguate and provide semantic coherence for taxonomies



Taxonomy Learning

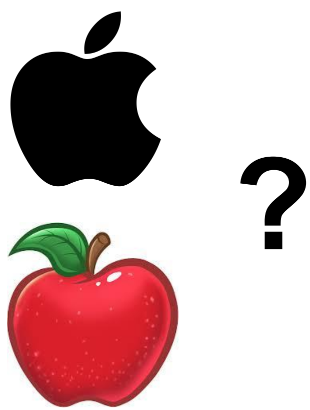
Espinosa-Anke et al. (AAAI 2016)

It uses sense representations to disambiguate and provide semantic coherence for taxonomies

Fruit

Apple   Orange   Lemon

Apple



Apple

Taxonomy Learning

Espinosa-Anke et al. (AAAI 2016)

It uses sense representations to disambiguate and provide semantic coherence for taxonomies

Fruit

Apple | Orange | Lemon

Apple

Semantically incoherent according to the sense vectors (SensEmbed)

Apple

Similar to the root (*fruit*) according to cosine similarity Semantically coherent

Open Information Extraction

Delli Bovi et al. (EMNLP 2015)

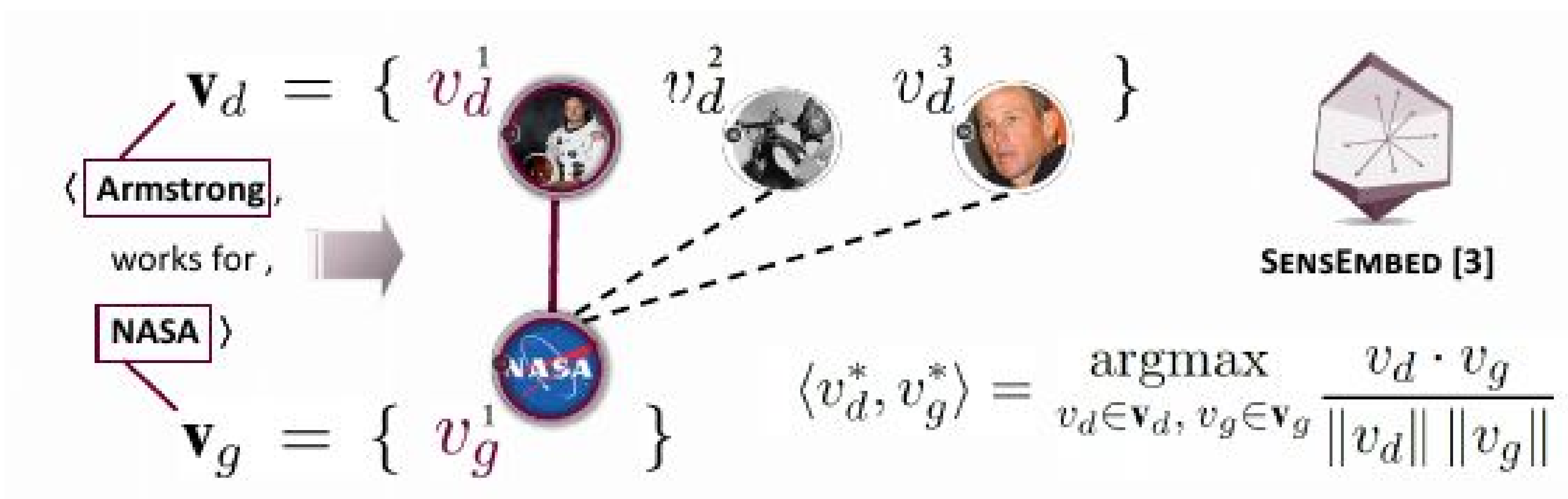
Idea

Integrate the output of different Open Information Extraction systems into a single **unified and fully disambiguated knowledge repository**.

Open Information Extraction

Delli Bovi et al. (EMNLP 2015)

Similarly to the taxonomy learning approach, it uses sense representations to **disambiguate** and give a **semantic coherence to the extracted relations**.



Sense Clustering

- Current sense inventories suffer from the **high granularity** of their sense inventories.
- A meaningful clustering of senses would help **boost the performance on downstream applications**

(Hovy et al., AIJ 2013)

- Examples:
  - Street (with sidewalks or without sidewalks) *in WordNet*
  - Parameter (computer programming) - Parameter *in Wikipedia*

Sense Clustering

Basic approach

Using a clustering algorithm based on the **semantic similarity between sense vectors**

Sense Clustering

- ADW (Pilehvar et al. ACL 2013) for **WordNet**

- NASARI (Camacho-Collados et al. AIJ 2016) for **Wikipedia**

Sense Clustering

(Pilehvar et al., ACL 2013)

	Onto		SE-2			Onto + SE-2	
Method	**Noun**	**Verb**	**Noun**	**Verb**	**Adj**	**Noun**	**Verb**
$\mathcal{R}_{Cos}$	0.406	0.522	0.450	0.465	0.484	0.441	0.485
$\mathcal{R}_{WO}$	**0.421**	**0.544**	**0.483**	**0.482**	**0.531**	**0.470**	**0.503**
$\mathcal{R}_{Jac}$	0.418	0.531	0.478	0.473	0.501	0.465	0.493
SVM	0.370	0.455	NA	NA	0.473	0.423	0.432
ODE	0.218	0.396	NA	NA	0.371	0.331	0.288

Clustering of WordNet senses (F-Measure)

Sense Clustering

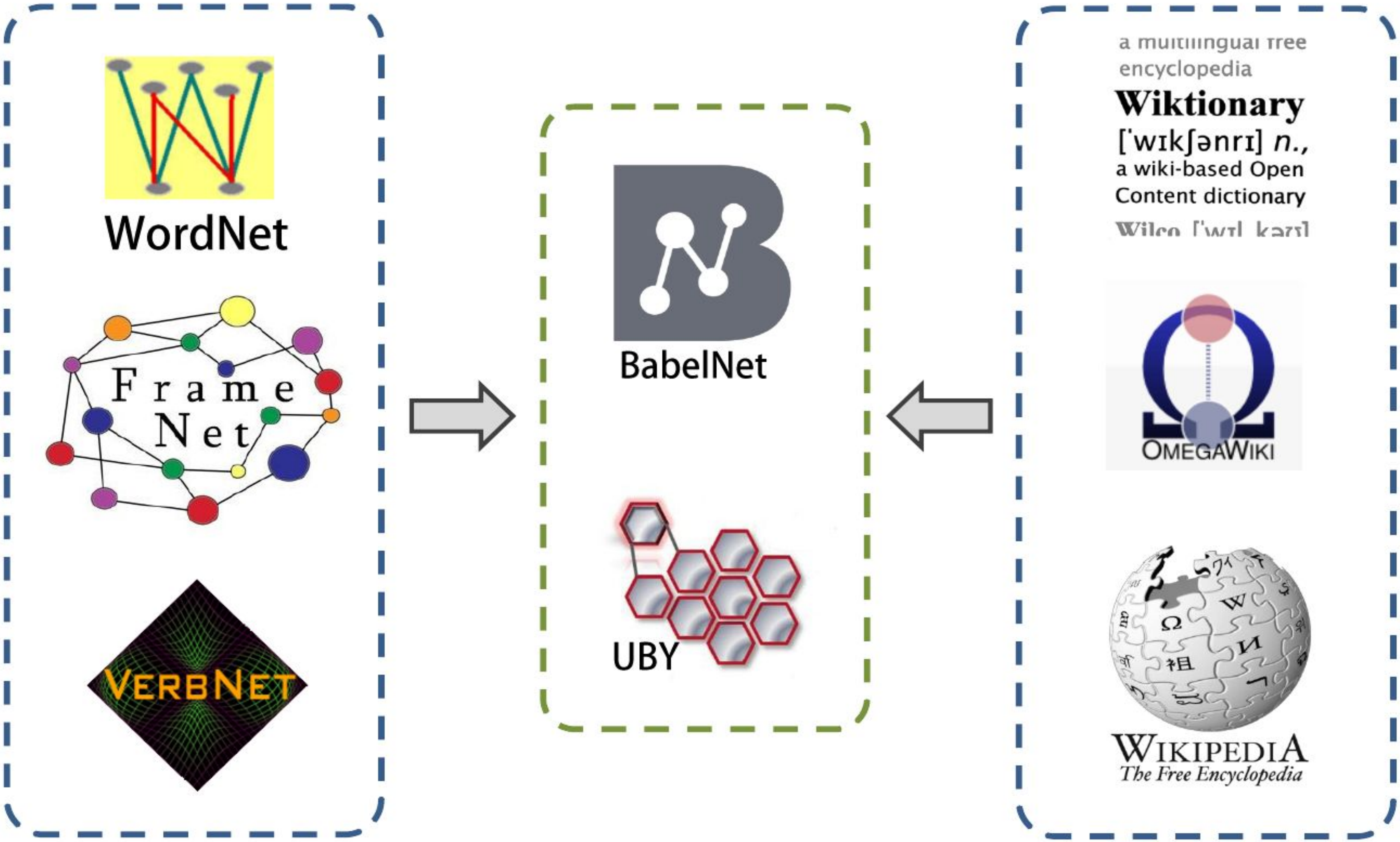
(Camacho-Collados et al., AIJ 2016)

Measure	System type	500-pair		SemEval	
		Acc.	F1	Acc.	F1
NASARI	unsupervised	83.8	**70.5**	**87.4**	**63.1**
$\text{NASARI}_{\text{lexical}}$	unsupervised	81.6	65.4	85.7	57.4
$\text{NASARI}_{\text{unified}}$	unsupervised	82.6	69.5	87.2	**63.1**
$\text{NASARI}_{\text{embed}}$	unsupervised	81.2	65.9	86.3	45.5
SVM-monolingual	supervised	77.4	-	83.5	-
SVM-multilingual	supervised	**84.4**	-	85.5	-
$\text{Baseline}_{\text{no-cluster}}$	-	71.4	0.0	82.5	0.0
$\text{Baseline}_{\text{cluster}}$	-	28.6	44.5	17.5	29.8

Clustering of Wikipedia pages

Alignment of Lexical Resources

Pilehvar and Navigli (ACL 2014)



Alignment of Lexical Resources

Pilehvar and Navigli (ACL 2014)

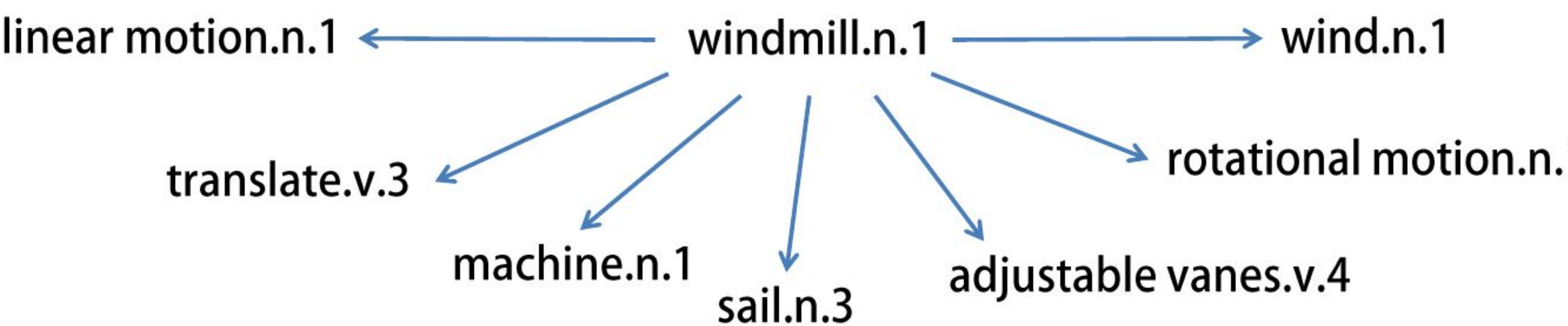
Idea: Ontologization of lexical resources to build a graph (semantic network) for each resource

Wiktionary
Show quotations
In other languages
Dansk
Eesti
Ελληνικά
Español
Français
한국어
Հայերեն
Ido
Italiano
Magyar

Definition page for *windmill*

Read Edit History Search

1. A machine which translates linear motion of wind to rotational motion by means of adjustable vanes called sails.



Alignment of Lexical Resources

Pilehvar and Navigli (ACL 2014)

Once the graph for each resource is constructed, **PageRank** is used to build a **sense representation** (i.e. semantic signature) for each concept.

Finally, sense representations with a very high degree of similarity are aligned.

Open Problems and Future Work

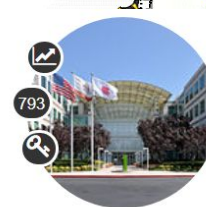
Open Problems and Future Work

1. Improve evaluation
 - Move from word similarity gold standards to end-to-end applications
   - Integration in Natural Language Understanding tasks (Li and Jurafsky, EMNLP 2015)
   - SemEval task? see e.g. WSD & Induction within an end user application @ SemEval 2013

Open Problems and Future Work

2. Make semantic representations more meaningful

- unsupervised representations are hard to inspect (clustering is hard to evaluate)
- but also knowledge-based approaches have issues:
  - e.g. top-10 closest vectors to the military sense of "company" in AutoExtend



AutoExtend
company_n^9
company
company_n^8
company_n^6
company_n^7
company_v^1
firm
business_n^1
firm_n^2
company_n^1

Open Problems and Future Work

3. Interpretability
   – The reason why things work or do not work is not obvious
     - E.g. avgSimC and maxSimC are based on implicit disambiguation that improves word similarity, but is not proven to disambiguate well
     - Many approaches are tuned to the task
   – Embeddings are difficult to interpret and debug

Open Problems and Future Work

4. Link the representations to rich semantic resources like WikiData and BabelNet
   – Enabling applications that can readily take advantage of huge amounts of multilinguality and information about concepts and entities
   – Improving the representation of low-frequency/isolated meanings

Open Problems and Future Work

5. Scaling semantic representations to sentences and documents
   – Sensitivity to word order
   – Combine vectors into syntactic-semantic structures
   – Requires disambiguation, semantic parsing, etc.
   – Compositionality

Open Problems and Future Work

6. Addressing multilinguality
   – a key trend in today's NLP research

- We are already able to perform POS tagging and dependency parsing in dozens of languages
  – Also mixing up languages

Open Problems and Future Work

- We can perform Word Sense Disambiguation and Entity Linking in hundreds of languages
  - Babelfy (Moro et al. 2014)
  - but with only a few sense vector representations
- Now: it is crucial that sense and concept representations are language-independent
- Enabling comparisons across languages
- Also useful in semantic parsing

Open Problems and Future Work

- Representations are most of the time evaluated in English
  - single words only
- It is important to evaluate sense representations in other languages and across languages
  - Check out the SemEval 2017 Task 2: multilingual and cross-lingual semantic word similarity (multilwords, entities, domain-specific, slang, etc.)

Open Problems and Future Work

7. Integrate sense representations into Neural Machine Translation

- Previous results in the 2000s working on semantically-enhanced SMT are not very encouraging
- However, many options have not been considered

Thank you!

Questions please!