Semantic Similarity Frontiers: From Concepts to Documents

David Jurgens
Stanford University
Mohammad Taher Pilehvar
Sapienza University of Rome

ERC grant 259234

Tutorial Objectives

- Make sense of current Semantic Similarity state of the art!
  - Formulate tasks and required resources
  - Standard and state-of-the-art algorithms
  - Current evaluation metrics

Tutorial Objectives

- Make sense of current Semantic Similarity state of the art!
  - Formulate tasks and required resources
  - Standard and state-of-the-art algorithms
  - Current evaluation metrics
- Provide practical knowledge
  - What open source tools and data are available
  - What are the current open problems

Tutorial Objectives

- Make sense of current Semantic Similarity state of the art!
  - Formulate tasks and required resources
  - Standard and state-of-the-art algorithms
  - Current evaluation metrics
- Provide practical knowledge
  - What open source tools and data are available
  - What are the current open problems
- Target audience: we assume no knowledge of any machine learning or lexical semantics
  - **Stop us to ask questions at any time!**

Tutorial *non*-Objectives

- Provide gory details of methodologies
  - We focus more on the landscape and knowing *which* methods matter
  - But feel free to ask questions on details if interested!
- Covering all work on a similarity task
  - Course materials provide an extended bibliography
  - We focus on the most exciting ideas (to us)

You should leave feeling comfortable knowing what papers to read next, why, and roughly what they're about!

Quick outline of the morning

- Foundations in Semantic Similarity
  - Concepts, Terminology, and Examples
- State of the Art Overviews
  - Similarity when comparing Concepts, Words, Phrases, Sentences, Paragraphs, *or* Documents
  - Cross-Level Semantic Similarity
- Open source Tools and Resources
- Current Challenges and Future Work

Quick outline of the morning

- Foundations in Semantic Similarity
  - Concepts, Terminology, and Examples
- State of the Art Overviews
  - Similarity when comparing Concepts, Words, Phrases, Sentences, Paragraphs, *or* Documents
  - Cross-Level Semantic Similarity
- Open source Tools and Resources
- Current Challenges and Future Work

Coffee Break happens in here!
10:30 - 11:00

Semantic Similarity's Key Question:

How similar are two linguistic items?

How similar are two sentences?

The boss fired the worker

The boss fired the employee

How similar are two sentences?

The boss fired the worker

The boss fired the employee

The supervisor let the employee go

How similar are two sentences?

The boss fired the worker	
The boss fired the employee	very similar
The supervisor let the employee go	very similar
The supervisor reprimanded the worker	somewhat similar
The boss promoted the worker	related

Don’t we already have solutions for semantic similarity?

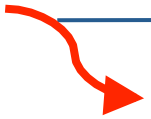
Lots of work on all types of text and concept input

Allison and Dix (1986)
Gusfield (1997)
Wise (1996)
Keselj et al. (2003)
50+ Approaches from SemEval
2012, 2013, 2014

Sussna (1993, 1997)
Wu and Palmer (1994)
Resnik (1995)
Jiang and Conrath (1997)
Lin (1998)
Hirst and St-Onge (1998)
Leacock and Chodorow (1998)
Patwardan (2003)
Banerjee and Pederson (2003)

Salton and McGill (1983)
Landauer et al. (1998)
Turney (2007)
Gabrilovich and Markovitch (2007)
Ramage et al. (2009)
Yeh et al. (2009)
Radinsky et al. (2011)

We refer to these as Linguistic Levels



Sentence Word Sense

Lots of work on all types of text and concept input

Allison and Dix (1986)
Gusfield (1997)
Wise (1996)
Keselj et al. (2003)
50+ Approaches from SemEval
2012, 2013, 2014

Not to mention word embeddings…

Sussna (1993, 1997)
Wu and Palmer (1994)
Resnik (1995)
Jiang and Conrath (1997)
Lin (1998)
Hirst and St-Onge (1998)
eacock and Chodorow (1998)
Patwardan (2003)
anerjee and Pederson (2003)

Gab 7)

Ramage et al. (2009)
Yeh et al. (2009)
Radinsky et al. (2011)

We refer to these as Linguistic Levels

Sentence Word Sense

Why do we have so many similarity methods?!

- New resources or machine learning methods become available
  - ~20 embeddings papers at EMNLP alone

Why do we have so many similarity methods?!

- New resources or machine learning methods become available
  - ~20 embeddings papers at EMNLP alone
- New datasets reveal weaknesses in previous methods
  - SOA is a moving target

Why do we have so many similarity methods?!

- New resources or machine learning methods become available
  - ~20 embeddings papers at EMNLP alone
- New datasets reveal weaknesses in previous methods
  - SOA is a moving target
- Need to adapt for new types of input or domains
  - Microtext, Biomedical, Multilingual

Why do we have so many similarity methods?!

- New resources or machine learning methods become available
  - ~20 embeddings papers at EMNLP alone
- New datasets reveal weaknesses in previous methods
  - SOA is a moving target
- Need to adapt for new types of input or domains
  - Microtext, Biomedical, Multilingual
- Application-specific similarity functions

Do we still need *more* methods?

- Semantic similarity itself is not an end-task, but rather a component
  - Applications can select the similarity method that yields the best performance.

- Performance on new benchmarks is still not satisfactory
  - Low hanging similarity fruit is solved, but *many* challenging cases still remain

Foundations

Semantic similarity can be defined on many linguistic levels

- Word senses (concepts)
- Words
- Phrases
- Sentences
- Paragraphs
- Documents

For the most part, different algorithms are used for each kind of item being compared.

Similarity is *graded*

car vs. automobile -> 1.0
car vs. vehicle -> 0.6
car vs. tire -> 0.2
car vs. street -> 0.1

Similarity has psychological quirks

- Nontransitive
  - Cuba vs. Jamaica
  - Cuba vs. China
  - Jamaica vs. China

- Asymmetric
  - North Korea vs. China
  - China vs. North Korea

These are ignored by nearly all approaches, but see Gawron (2014)

Similarity vs. Relatedness

Similarity is a specific type of relatedness

- **Similarity**: **synonyms** and **hyponyms/hyperonyms**, and **siblings** are highly similar
  - Doctor vs. surgeon, Bike vs. bicycle

- **Related**: **topically related** or based on any other semantic relation
  - Heart vs. surgeon, tyre vs. car

Relational similarity

- The **degree of correspondence** between two relations:
  - Linux – grep
  - Windows – findstr

  - France – paris
  - Italy - Rome

- SemEval-2012 Task 2: Measuring Degrees of Relational Similarity (Jurgens et al)

Desiderata for a Semantic Similarity Method

- Consistently interpretable similarity scores with explanations of why similar
- Works well for different types of text (news, web, social media, …)
- Applicable to multiple linguistic types (words, phrases, sentences, …)

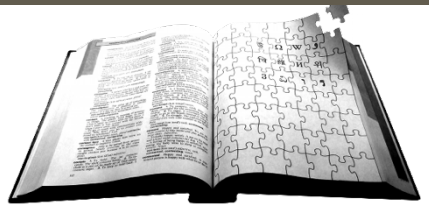
Typically, two main resources for measuring similarity

Massive corpora of text documents

Typically, two main resources for measuring similarity

Massive corpora of text documents

WordNet
A lexical database for English



WIKTIONARY
the free dictionary

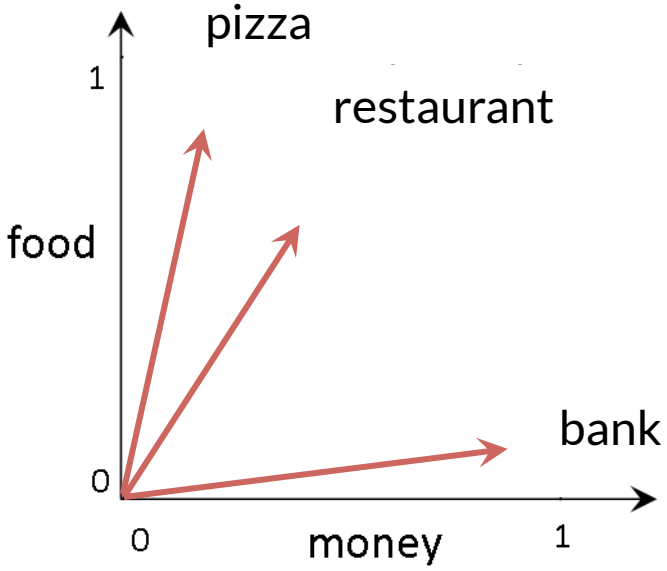
BabelNet

WIKIPEDIA
The Free Encyclopedia

Semantic resources and knowledge bases

Many methods represent semantics using a vector space model (VSM)

Vector spaces provide a machine-interpretable or mathematical format



Vector Space Models

- Simple representation based on linear algebra
- Easy comparison of different items based on a continuous scale of similarity
- Supported by studies in Cognitive science
- Flexible way of adjusting the degree of complication through setting the number of dimensions

Vector Space Models

Explicit

- Individual dimensions denote specific linguistic items, e.g., words
- Usually higher in dimension
- The vector is interpretable

Continuous

- Dimensions do not correspond to explicit concepts
- Usually lower in dimension

Vector Space Models

Vector comparison techniques

Kullback–Leibler (KL) divergence

$$D_{KL}\left(\mathcal{S}_1 \| \mathcal{S}_2\right) = \sum_{h \in H} log_e\left(\frac{\mathcal{S}_1^h}{\mathcal{S}_2^h}\right) \mathcal{S}_1^h$$

Jensen–Shannon (JS) divergence

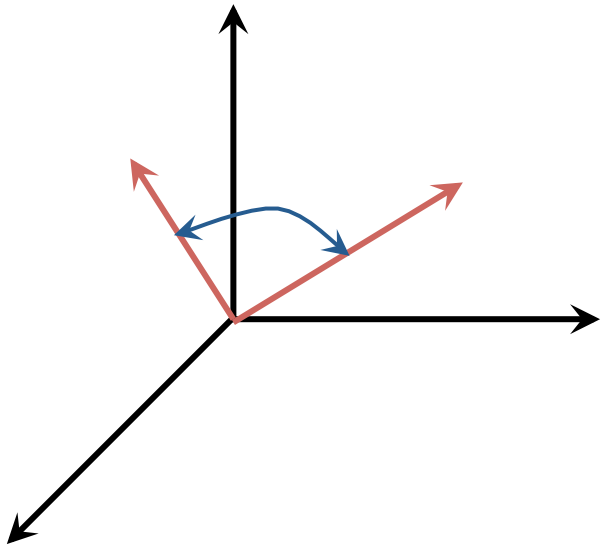
$$D_{JS}\left(\mathcal{S}_1, \mathcal{S}_2\right) = \frac{1}{2} D_{KL}\left(\mathcal{S}_1 \middle\| \frac{\mathcal{S}_1 + \mathcal{S}_2}{2}\right) + \frac{1}{2} D_{KL}\left(\mathcal{S}_2 \middle\| \frac{\mathcal{S}_1 + \mathcal{S}_2}{2}\right)$$

Vector Space Models

Vector comparison techniques

Cosine similarity

$$Sim_{Cos}(\mathcal{S}_1, \mathcal{S}_2) = \frac{\mathcal{S}_1 \cdot \mathcal{S}_2}{\|\mathcal{S}_1\| \ \|\mathcal{S}_2\|}$$



Vector Space Models

Vector comparison techniques

Tanimoto similarity (1957)

$$f(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

Vector Space Models

Rank-based Vector comparison techniques

Rank-Biased Overlap (RBO)

The set of overlapping dimensions between the top-*d* elements

$$RBO\,(\mathcal{S}_1, \mathcal{S}_2) = (1-p)\sum_{d=1}^{|H|} p^{d-1}\,\frac{|H_d|}{d}$$

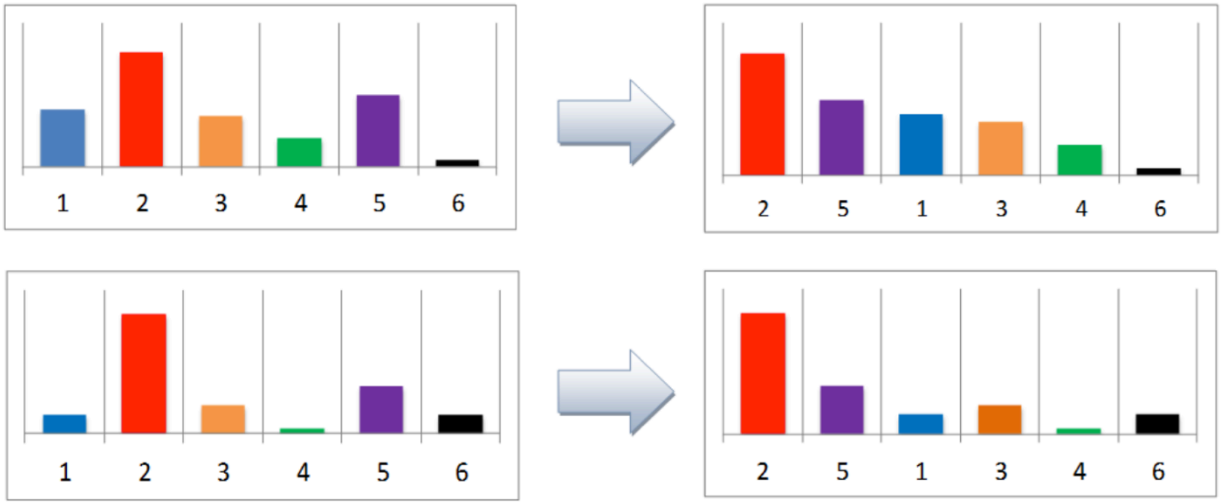
A parameter that determines the relative importance of the top elements.

Vector Space Models

Rank-based Vector comparison techniques

Weighted Overlap

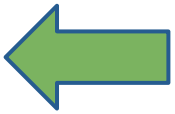
$$Sim_{WO}(\mathcal{S}_1, \mathcal{S}_2) = \frac{\sum_{h \in H} \left(r_h(\mathcal{S}_1) + r_h(\mathcal{S}_2)\right)^{-1}}{\sum_{i=1}^{|H|} (2i)^{-1}}$$



Semantic Similarity: State of the Art

Many approaches incorporate techniques from more specific linguistic levels

- Word senses (concepts)
- Words
- Phrases
- Sentences
- Paragraphs
- Documents



Start here and work our way to bigger ideas!

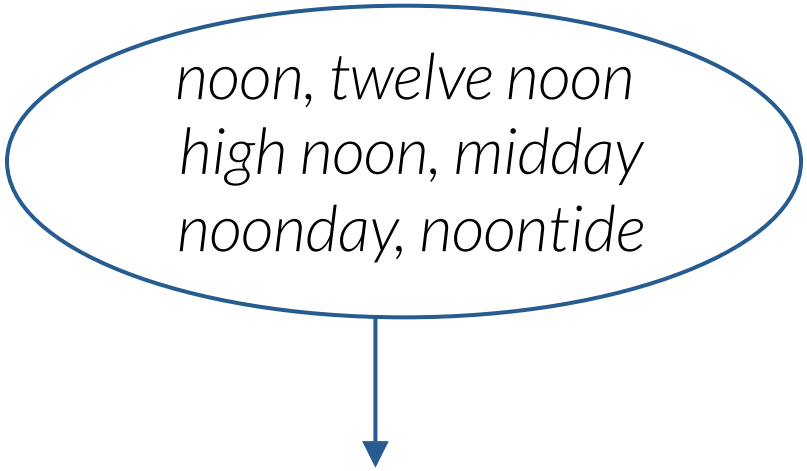
Semantic Similarity

between word senses

Concepts vs. senses

A WordNet synset (concept):

the middle of the day



(noon#n#1)

Applications - general

- Lowest (most fine-grained) level of semantic similarity: can be extended to applications that require higher levels of similarity

MT evaluation, paraphrases recognition, textual entailment, information retrieval, question answering, text summarization, lexical substitution or simplification, query expansion

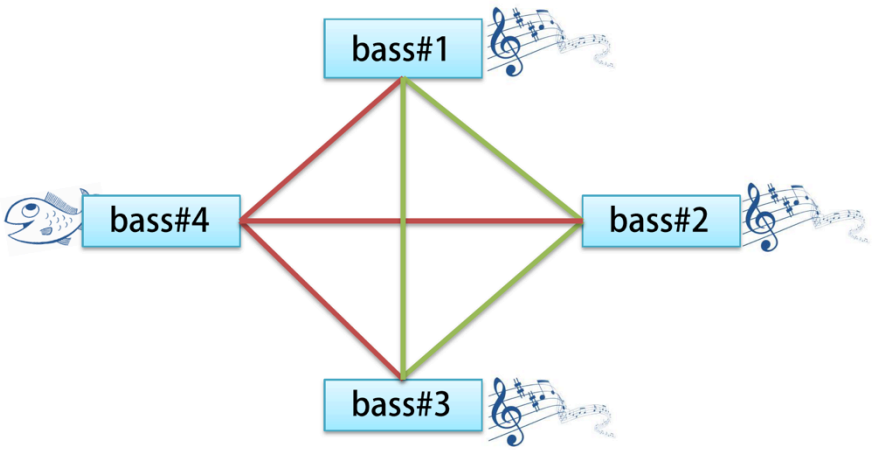
Applications - specific

Word Sense Disambiguation

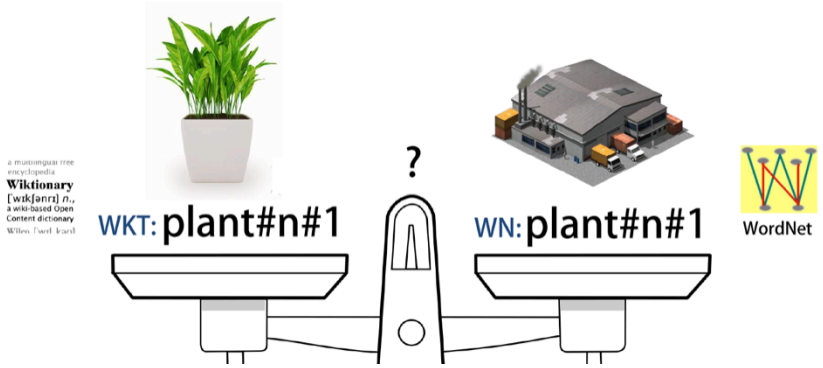
install the updated application

- software application?
- application for a job?
- practical usage?

Coarsening



Resource Alignment



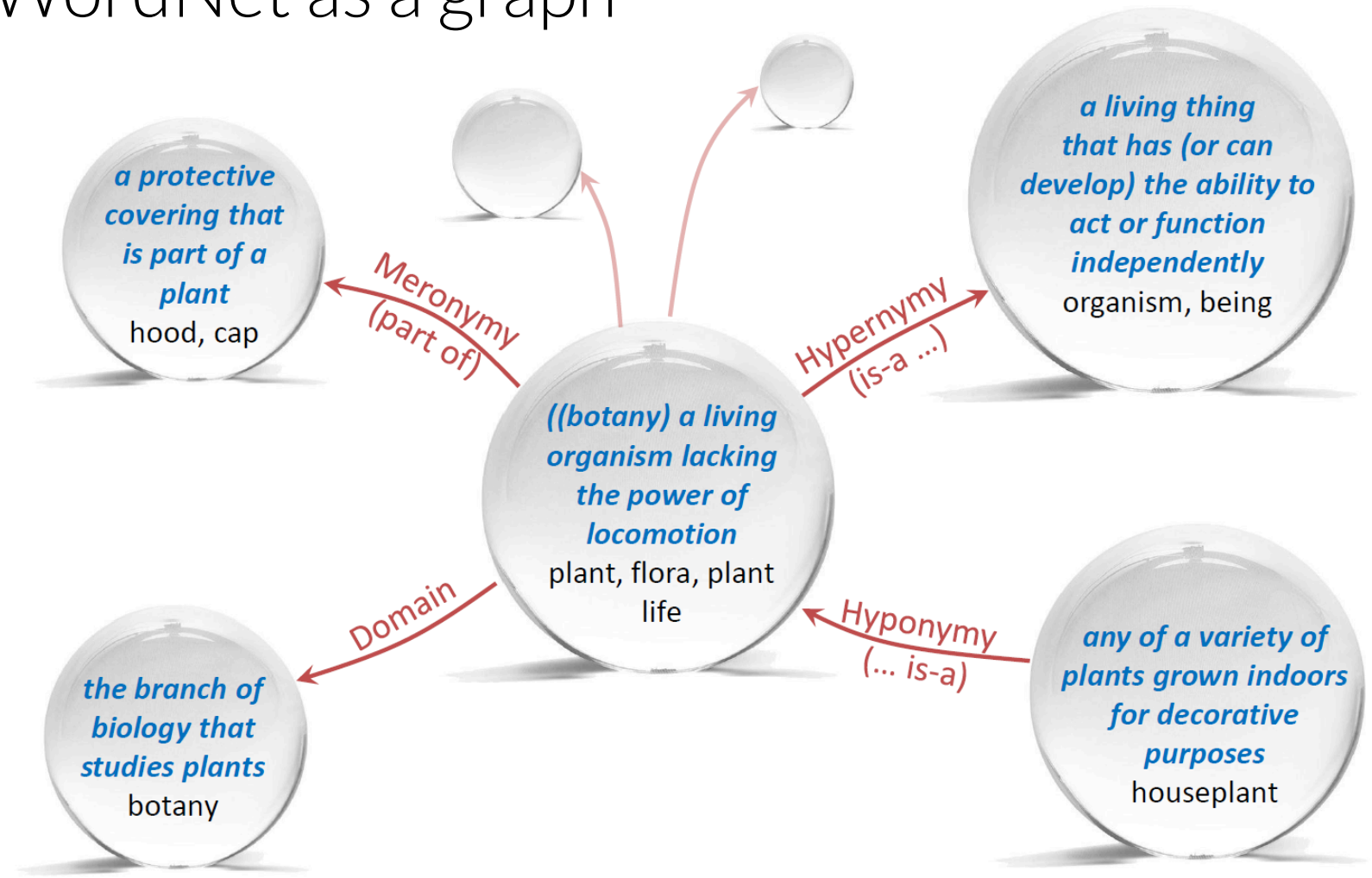
Sense Similarity Techniques

- Tied to sense inventories
  - Graph distance-based
    - WordNet-based
    - Thesauri-based
    - Dictionary-based
  - Explicit sense representation
    - Simple gloss-based
    - Random walk-based
    - Distributional
- Not tied to sense inventories

Sense Similarity Techniques

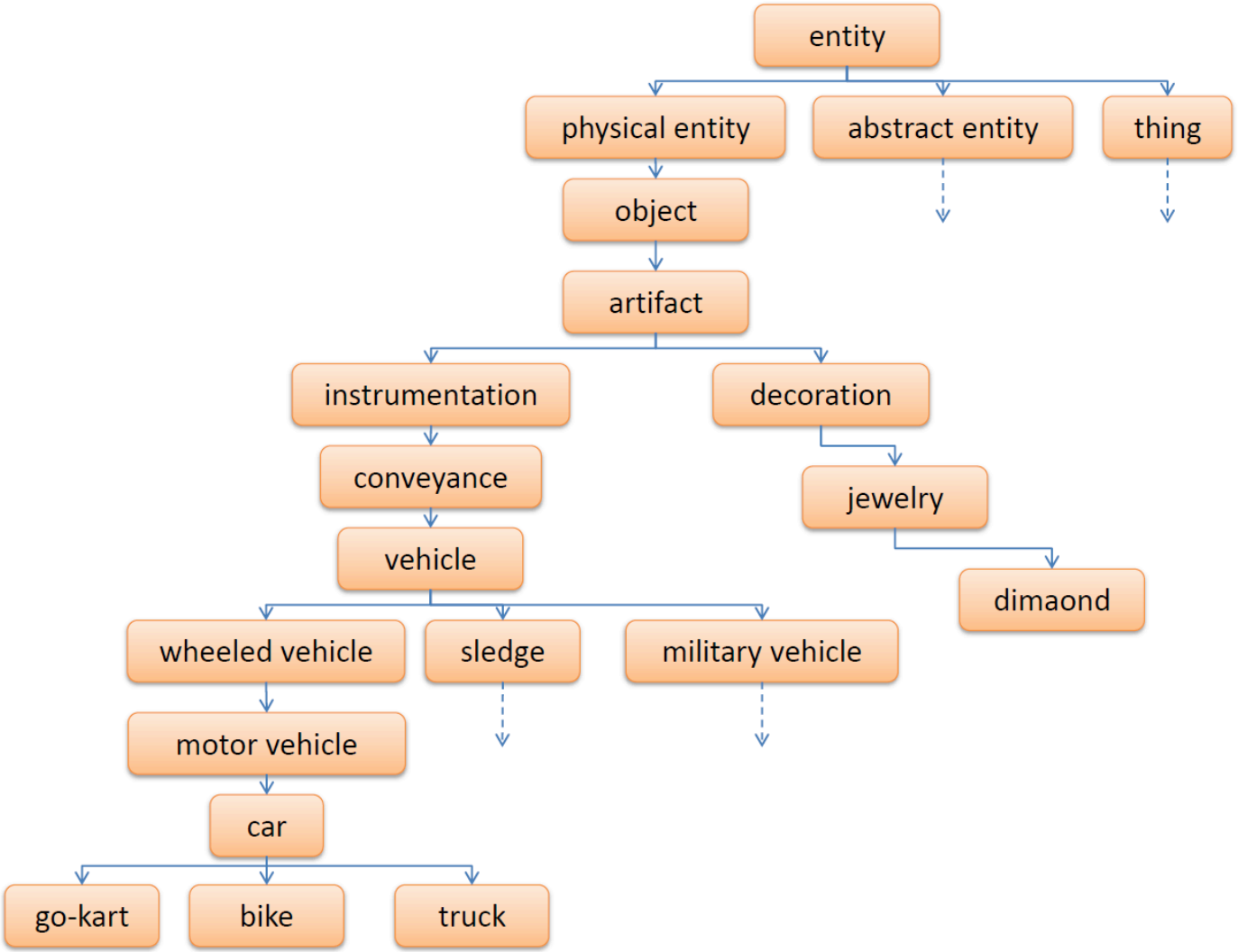
Tied to sense inventories: graph distance

WordNet as a graph



Sense Similarity Techniques

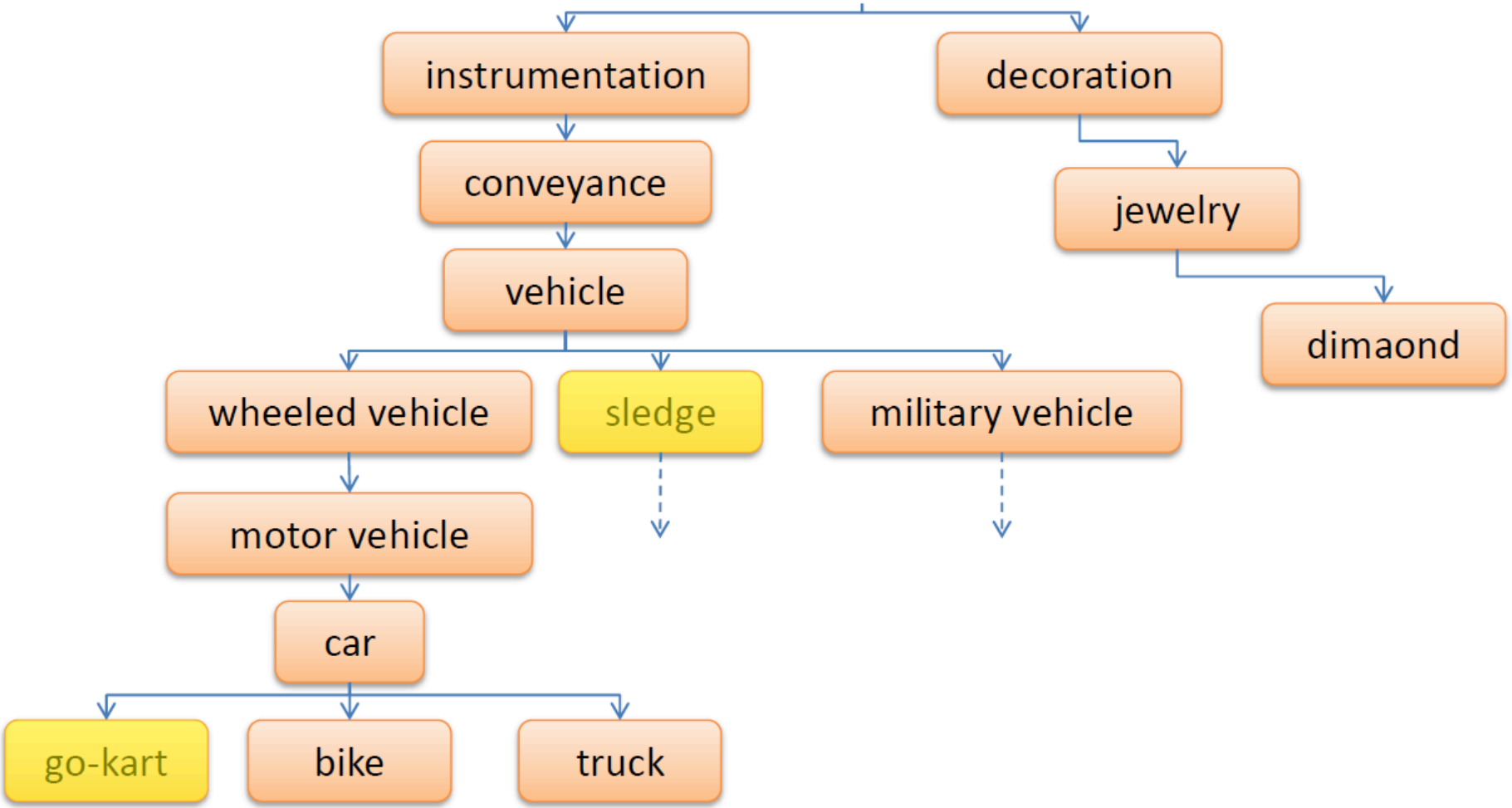
Tied to sense inventories: WordNet graph distance



Sense Similarity Techniques

Tied to sense inventories: WordNet graph distance

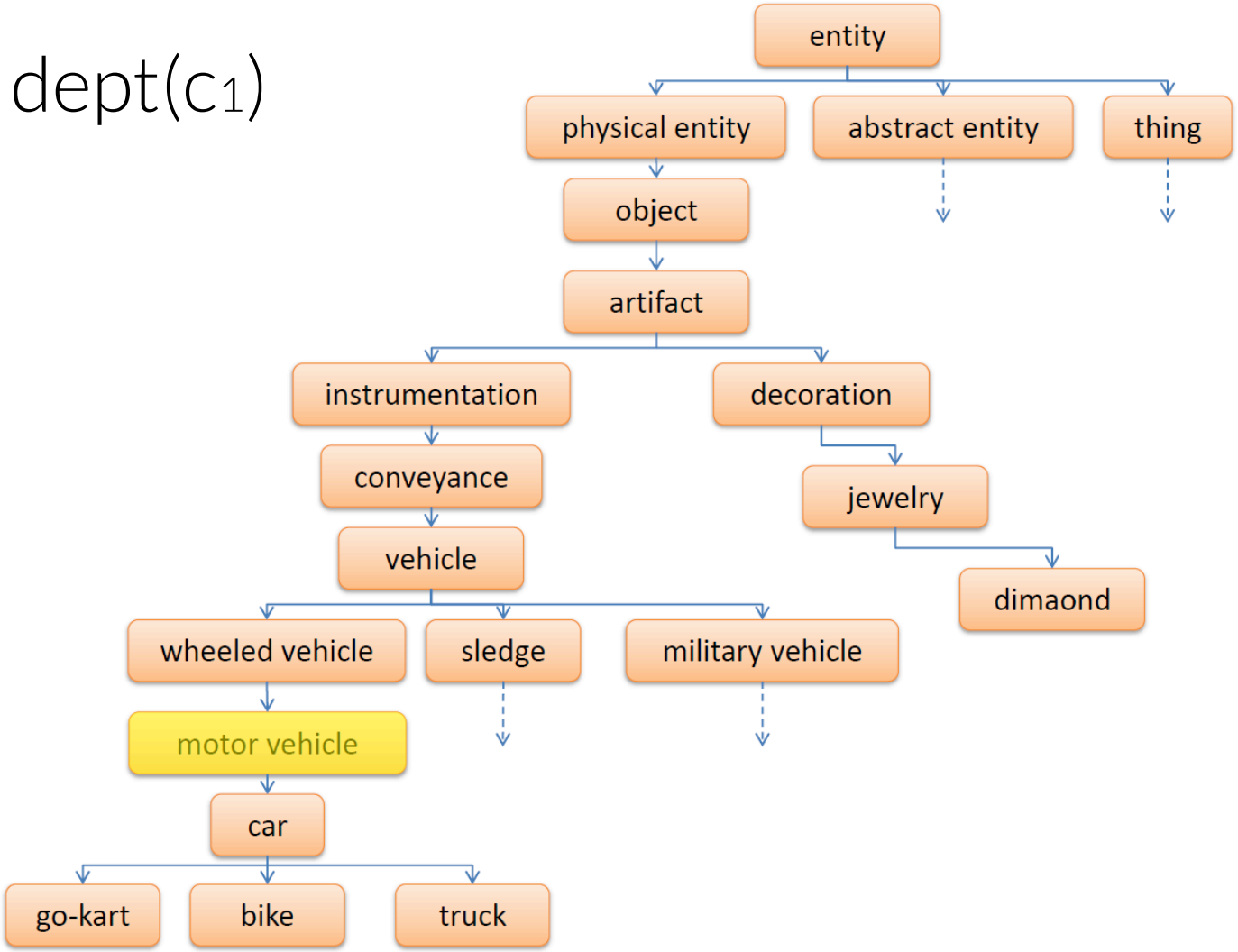
$len(c_1,c_2)$



Sense Similarity Techniques

Tied to sense inventories: WordNet graph distance

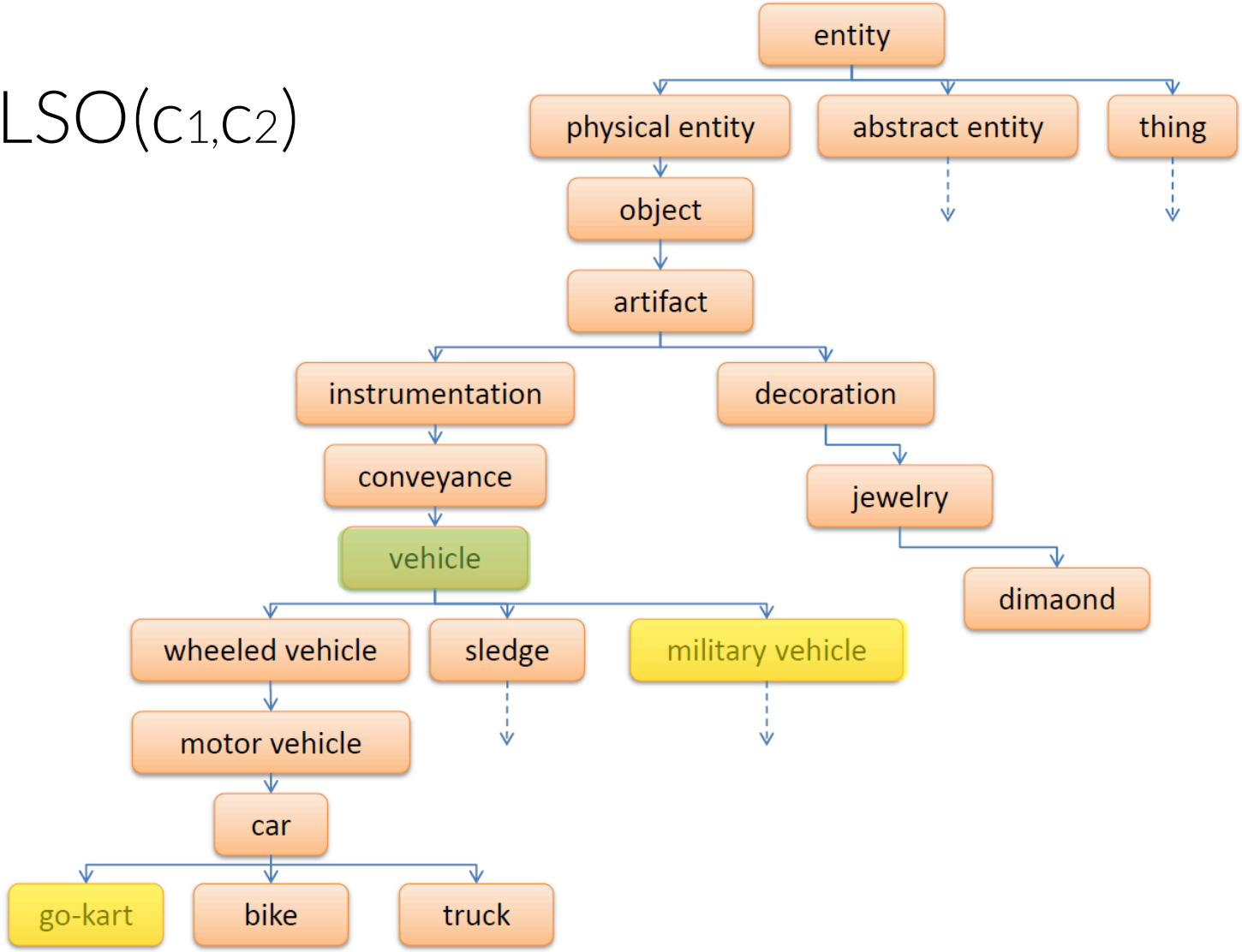
dept(c_1)



Sense Similarity Techniques

Tied to sense inventories: WordNet graph distance

LSO(c_1,c_2)



Sense Similarity Techniques

Tied to sense inventories: WordNet graph distance

Conventional WordNet-based techniques

Survey: Budanitsky and Hirst (2006)

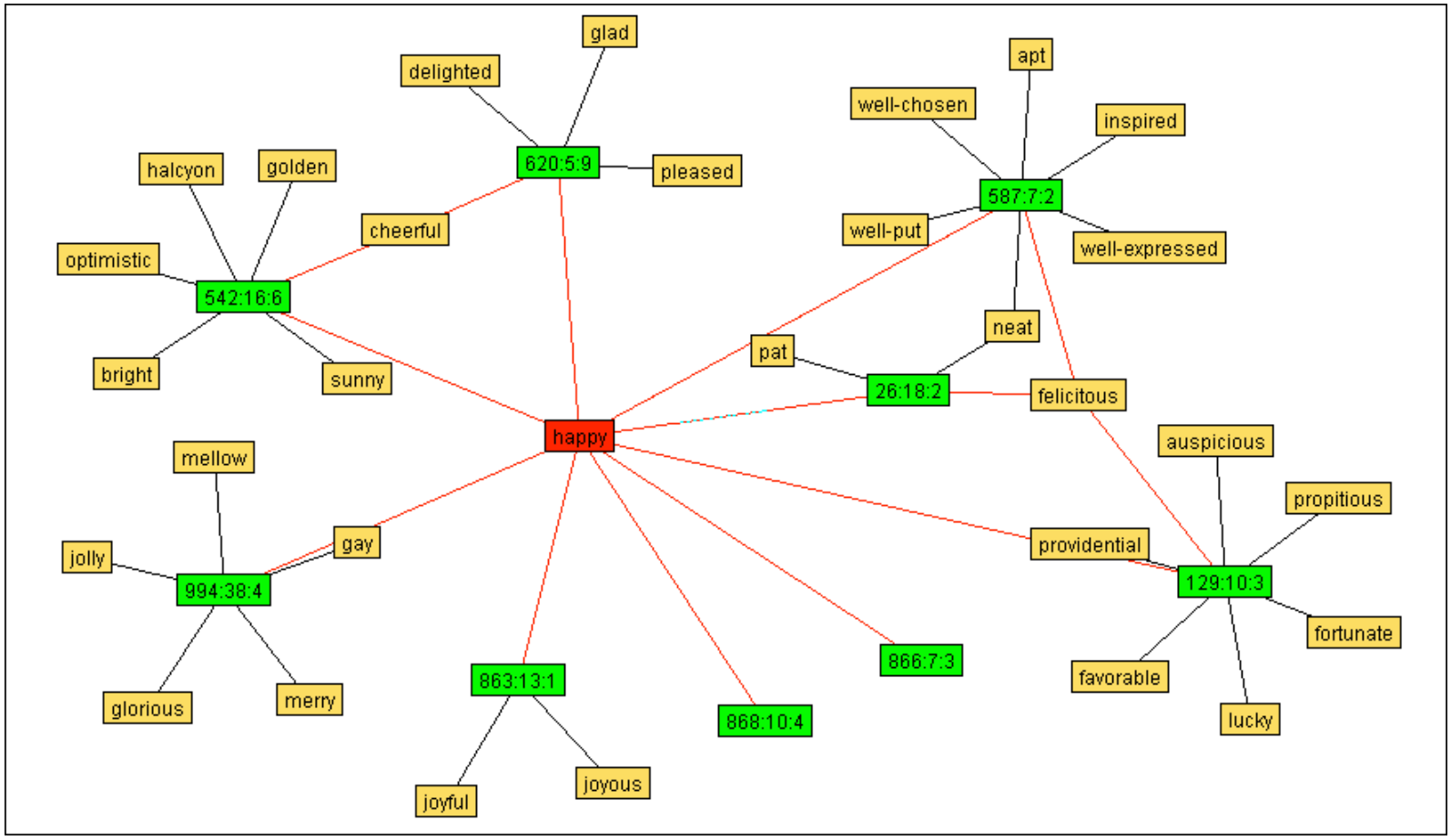
- WordNet structure only
  - Hirst and St-Onge (1998)
  - Sussna's Depth-relative Scaling (1993, 1997)
  - Wu and Palmer (1994)
  - Leacock and Chodorow's (1998)

- Combined with statistics from corpora
  - Jiang and Conrath's Measure (1997)
  - Resnik (1995)
  - Lin's Measure (1998)

Sense Similarity Techniques

Tied to sense inventories: Thesauri-based

Roget's thesaurus: Morris and Hirst (1991), Jarmasz and Szpakowicz (2003)



Sense Similarity Techniques

Tied to sense inventories: Dictionary-based

Longman Dictionary (LDOCE): Kozima and Furugori (1993), Kozima and Ito (1997)

- Constructs a semantic network from a subset of the dictionary, 2851 nodes, called Paradigme

- Computes similarity by spreading the activation in the network

Sense Similarity Techniques

Tied to sense inventories

Explicit semantic representation

Sense Similarity Techniques

Tied to sense inventories: Explicit semantic representation

Simple gloss-based: Exploiting WordNet's content

application#n#2 --

a verbal or written request for assistance or employment or admission to a school

application#n#4 --

a program that gives a computer instructions that provide the user with tools to accomplish a task

example:

Meerkat Mafia - Kashyap et al (2014)
@ SemEval-2014 Task-3: CLSS

Sense Similarity Techniques

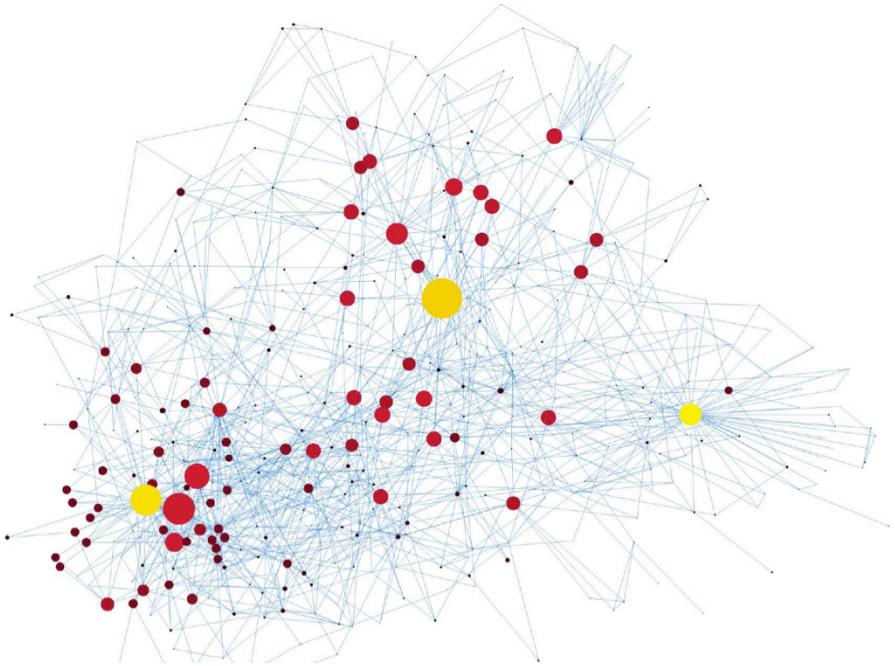
Tied to sense inventories: Explicit semantic representation

Random walks on semantic networks

The Personalized PageRank algorithm

Semantic similarity: Pilehvar et al (2013)

WSD: Agirre et al (CL 2014)



Sense Similarity Techniques

Tied to sense inventories: Explicit semantic representation

Distributional

SensEmbed - word2vec sense embeddings

Iacobacci et al (2015)

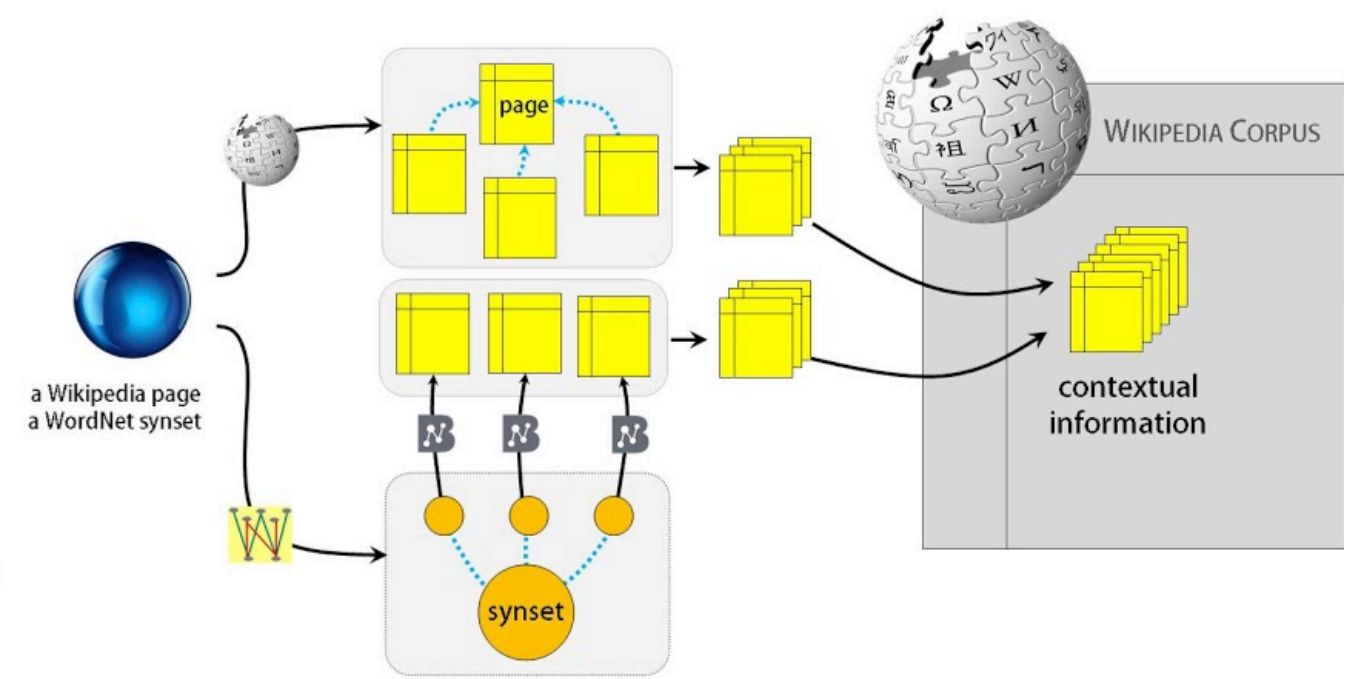
+

Sense Similarity Techniques

Tied to sense inventories: Explicit semantic representation

Distributional

NASARI and MUFFIN - Camacho-collados et al (2015)



Sense Similarity Techniques

Tied to sense inventories: Explicit semantic representation

Distributional

Chen et al (emnlp 2014)

Joint word sense representation and disambiguation

- Learn word representations (word2vec skip-gram)
- Use them for sense representation (average gloss)
- Automatically disambiguate large amounts of text
- Modify the objective of Skip-gram to learn sense representations

Sense Similarity Techniques

Tied to sense inventories: Explicit semantic representation

Distributional

Rothe and Schutze (acl 2015)

Extends word embeddings (word2vec) to embeddings of other data types: WordNet synsets and word senses

- Constructs an auto-encoder
- Learns these representations based on WordNet constraints (word/synset is the summation of its lexemes + WN relations)

Sense Similarity Techniques

Not Tied to sense inventories

Also called

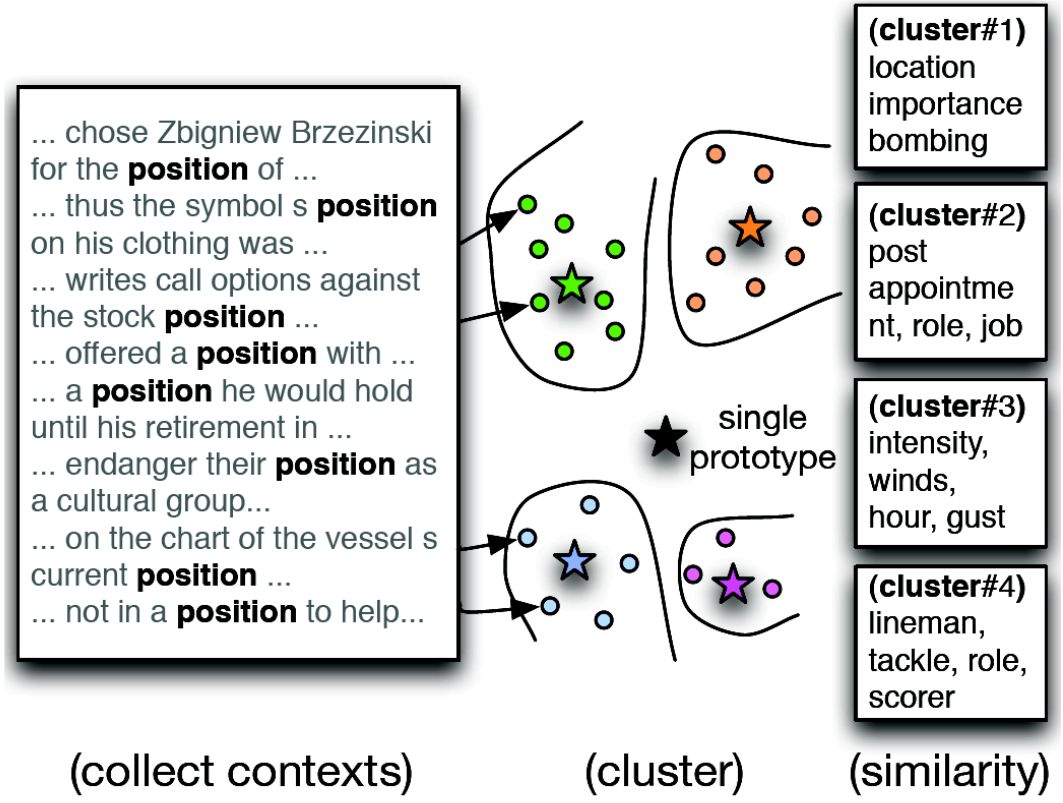
multi-prototype or **topic-based** representations

Usually based on clustering

Sense Similarity Techniques

Not Tied to sense inventories

Reisinger and Mooney (2010)



Sense Similarity Techniques

Not Tied to sense inventories

Reisinger and Mooney (2010)

Measuring similarity - isolated words:

$$\text{AvgSim}(w, w') \overset{\text{def}}{=} \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} d(\pi_k(w), \pi_j(w'))$$

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

Sense Similarity Techniques

Not Tied to sense inventories

Reisinger and Mooney (2010)

Measuring similarity - words in contexts:

$$\text{AvgSimC}(w, w') \stackrel{\text{def}}{=} \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} d_{c,w,k} d_{c',w',j} d(\pi_k(w), \pi_j(w'))$$

likelihood of the cluster given the context

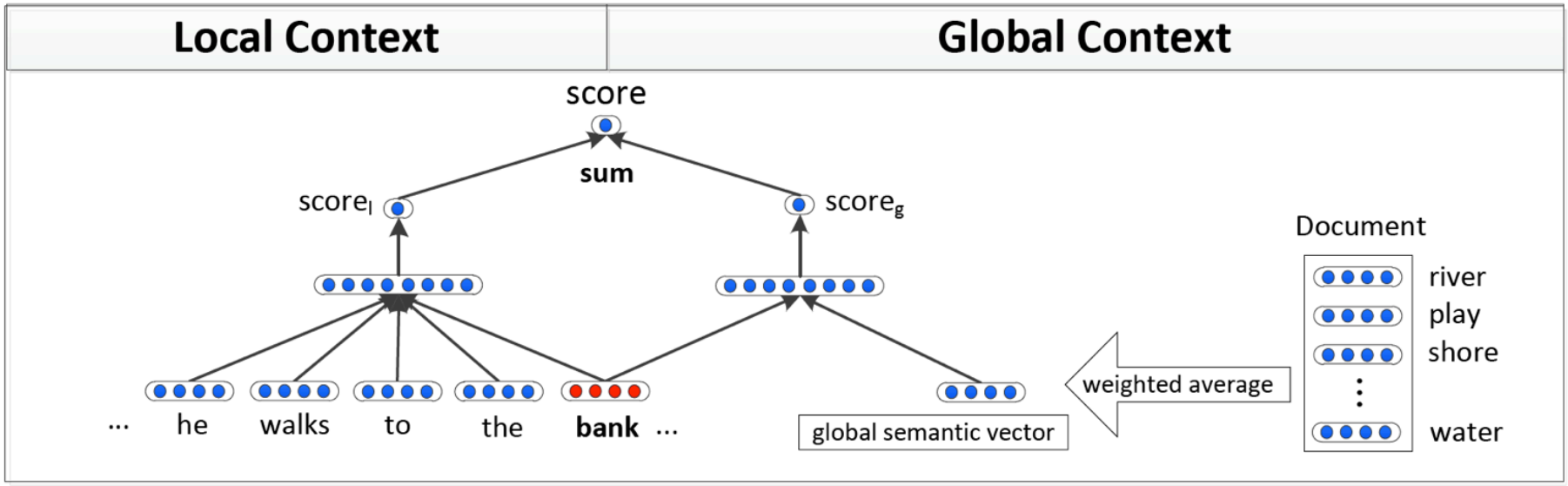
$$\text{MaxSimC}(w, w') \stackrel{\text{def}}{=} d(\hat{\pi}(w), \hat{\pi}(w'))$$

Sense Similarity Techniques

Not Tied to sense inventories

Huang et al (2012)

- Learns word embeddings with local and global objectives
- Then clusters the contexts of a word and learns multi-prototype representations



Sense Similarity Techniques

Not Tied to sense inventories

Neelakantan et al (emnlp 2014)

Multi-Sense Skip-gram (MSSG) model

(fixed number of senses)

Sense discrimination and learning embeddings are performed jointly

by disambiguating a word using current parameters

Non-parametric MSSG model

(varying number of senses per word)

Different in the sense discrimination phase

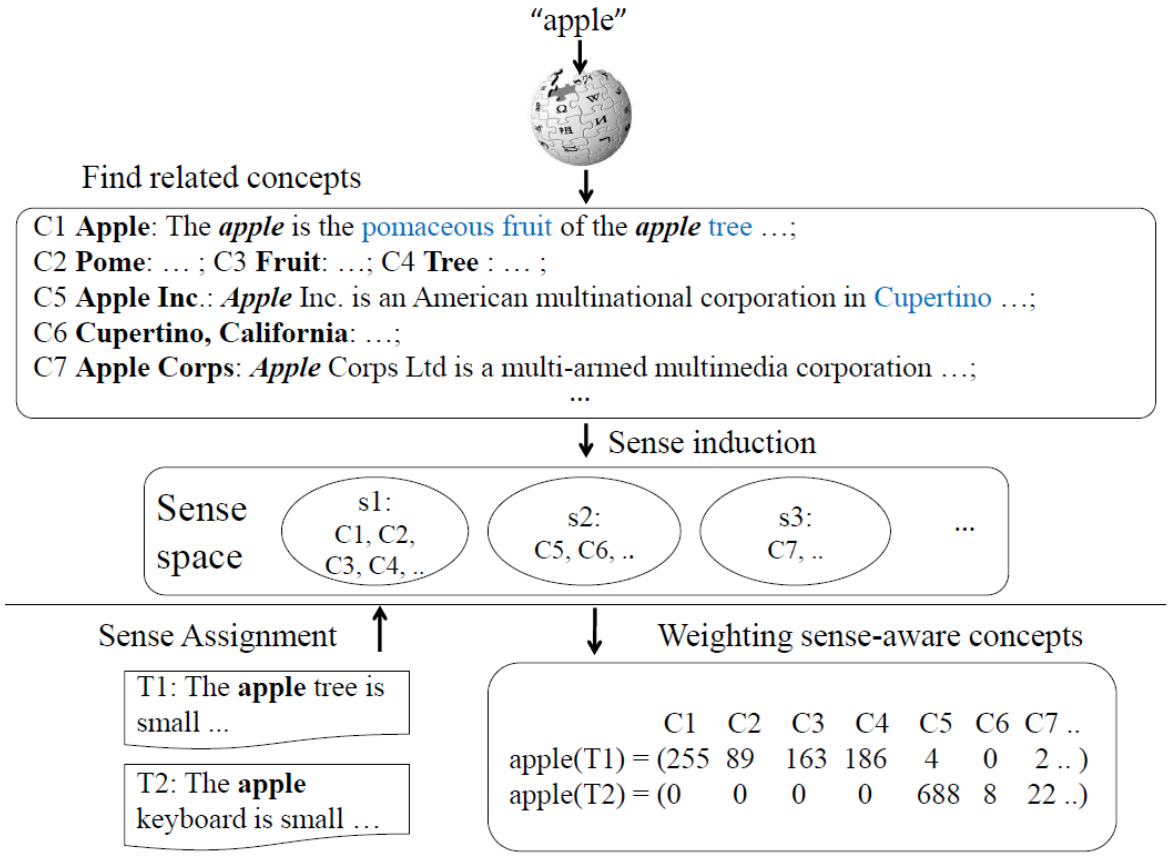
Online non-parametric clustering

Sense Similarity Techniques

Not Tied to sense inventories

SaSA - Sense-aware Semantic Analysis

Wu and Giles (AAAI 2015)



Sense Similarity Techniques

Not Tied to sense inventories

Topical Word Embeddings - Liu et al (AAAI 2015)

Different senses of a word can overlap
-> soft clustering

Uses LDA to learn representations for <word,topic> pairs

$$\mathcal{L}(D) = \frac{1}{M}\sum_{i=1}^{M}\sum_{-k\leq c\leq k, c\neq 0} \log \Pr(w_{i+c}|w_i) + \log \Pr(w_{i+c}|z_i).$$

Sense Similarity

Evaluation benchmarks

- Word similarity

and all other word-level applications

- Sense merging
- Word Sense Disambiguation
- Stanford's Contextual Word Similarities (SCWS)
- Cross Level Semantic Similarity
  (more details to follow)

Word Similarity

Word similarity is a lot like sense similarity

He went to the ATM to deposit the money.

She goes to the bank to withdraw cash.

Word similarity is a lot like sense similarity … except for ambiguity

He went to the ATM to deposit the money.

She goes to the bank to withdraw cash.

She goes to the shore near the silt deposit.

Word similarity is a lot like sense similarity … except for ambiguity

He went to the ATM to deposit the money.

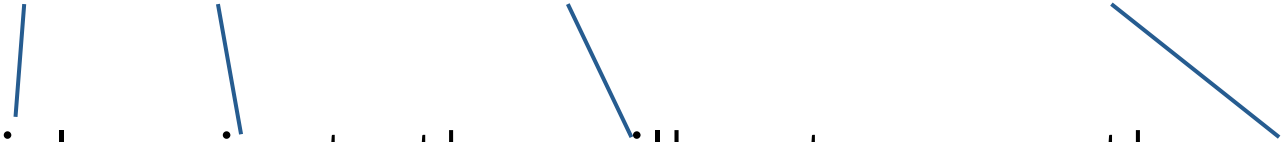
She goes to the bank to withdraw cash.

She goes to the shore near the silt deposit.

Most approaches measure similarity completely out of context.

Word similarity lets you easily build to larger linguistic level's similarities

The boy sailed the boat over the ocean.



The girl navigate the sailboat across the sea.

Many applications benefit from having word representations that encode similarity or having effective word similarity functions.

- Text classification (Baker and McCallum, 1998)
- Document classification (Sebastinani et al, 2002)
- Question answering (Tellex et al, 2003)
- IR (Sanderson, 1994), Manning et al (2008)
- Textual entailment (Baroni, 2014 - SICK)
- Named entity recognition (Turian et al, 2010, Passos et al, 2014)
- Dependency parsing (Bansal et al, 2014)
- Chunking (Turian et al, 2010, Dhillon and Ungar, 2011)
- Paraphrase detection (Socher et al, 2011)

Ideal references for comparing impact of new approaches

Most approaches evaluate on similarity benchmarks, rather than tasks

Numeric Word-Pair Similarity Tests

- Rubenstein & Goodenough, 1965 (RG)
- WordSim-353 (Finkelstein et al., 2001)
- Rare Words (Luong et al., 2013)
- MEN (Bruni et al., 2012)
- Radinsky et al., (2010)

Word Choice Tests

- TOEFL, ESL, Reader's Digest

TOEFL Synonymy recognition

enormous?

- ☐ appropriate
- ☐ unique
- ☑ tremendous
- ☐ decided

RG-65 judgement correlation

autograph	shore	0.06
coast	forest	0.85
midday	noon	3.94

Stanford Rare Word (RW) judgement correlation

dispossess	deprive	6.83
entrapping	capture	8.00
ruralist	advocate	0.67
acoustical	remedy	0.14
quieten	hush	9.38

What if we know nothing (about the words)?

You shall know a word by the company it keeps

-- Firth (1957)

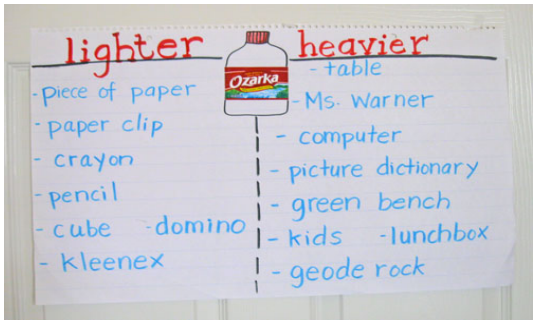
Learning semantic representations from text

1) Corpus

Learning semantic representations from text

1) Corpus

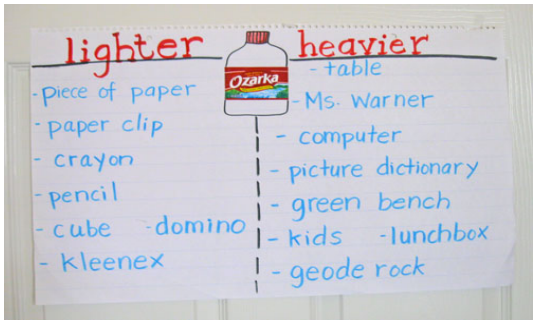
2) Preprocessing



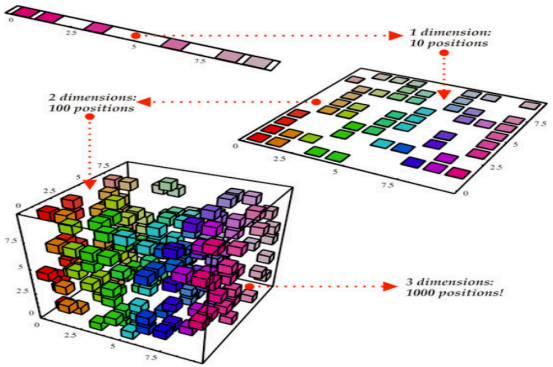
Learning semantic representations from text

1) Corpus

2) Preprocessing



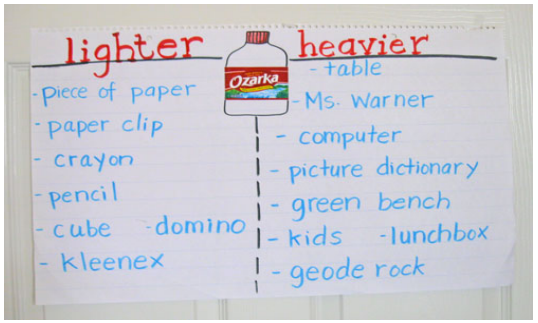
3) Dimensionality Reduction



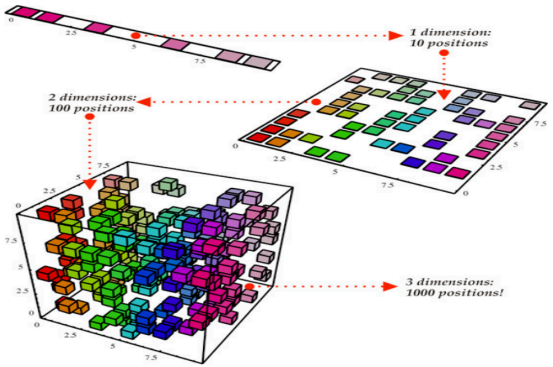
Learning semantic representations from text

1) Corpus

2) Preprocessing

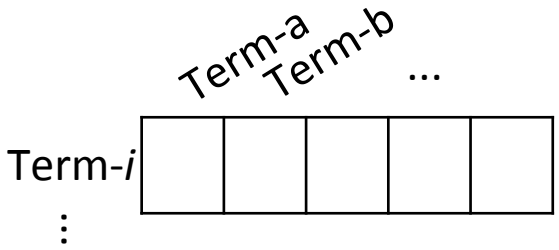


3) Dimensionality Reduction



4) Post Processing

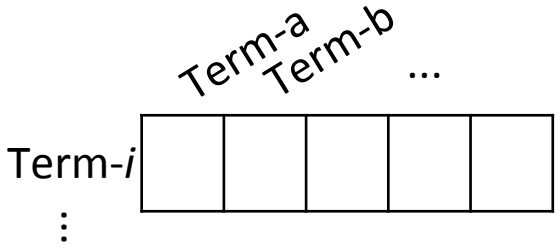
Three Typical Setups: Term-Term, Term-Context or Term-Document Matrix



Cells record the number of times...

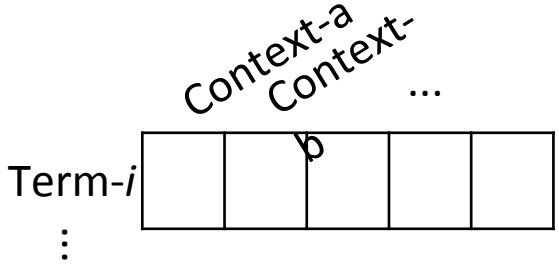
term *j* occurs in the context window of term *i*.

Three Typical Setups: Term-Term, Term-Context or Term-Document Matrix



Cells record the number of times...

term *j* occurs in the context window of term *i*.

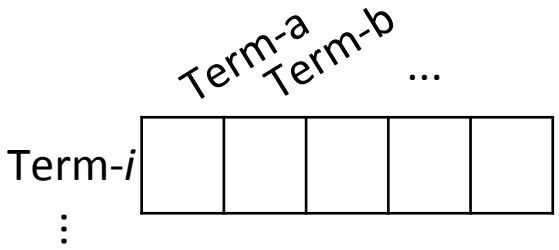


term *i* occurs in a context window

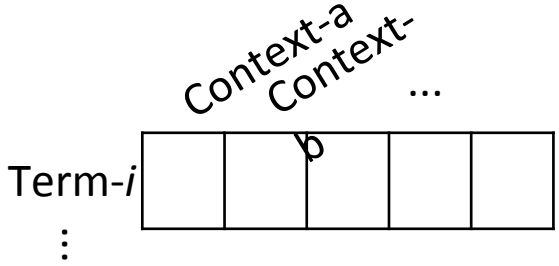
- w_{-2}, w_{-1}, w, w_1, w_2
- or analogously, with dependencies

Three Typical Setups: Term-Term, Term-Context or Term-Document Matrix

Cells record the number of times...

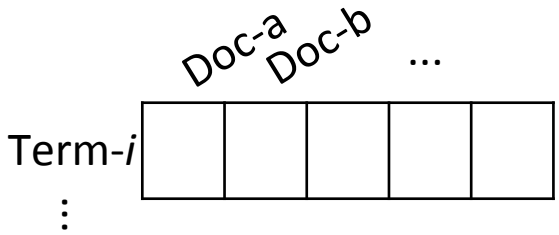


term *j* occurs in the context window of term *i*.



term *i* occurs in a context window

- w_{-2}, w_{-1}, w, w_1, w_2
- or analogously, with dependencies

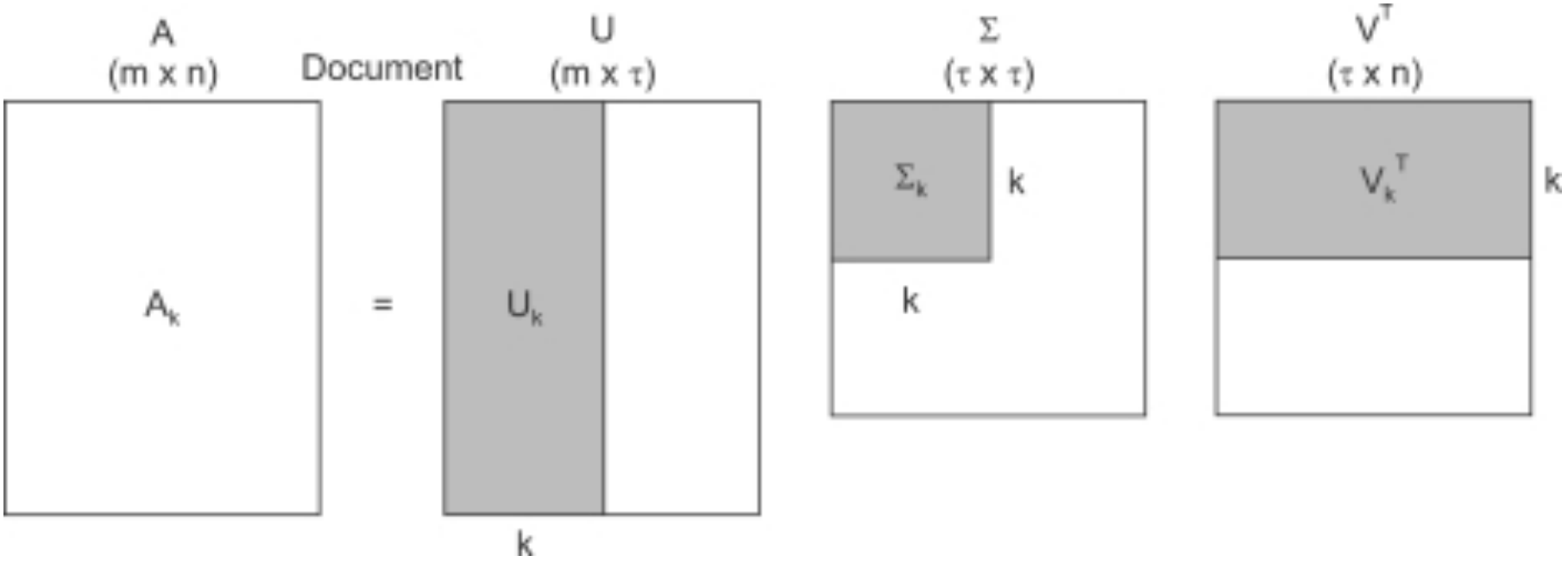


term *i* occurs in document *j*.

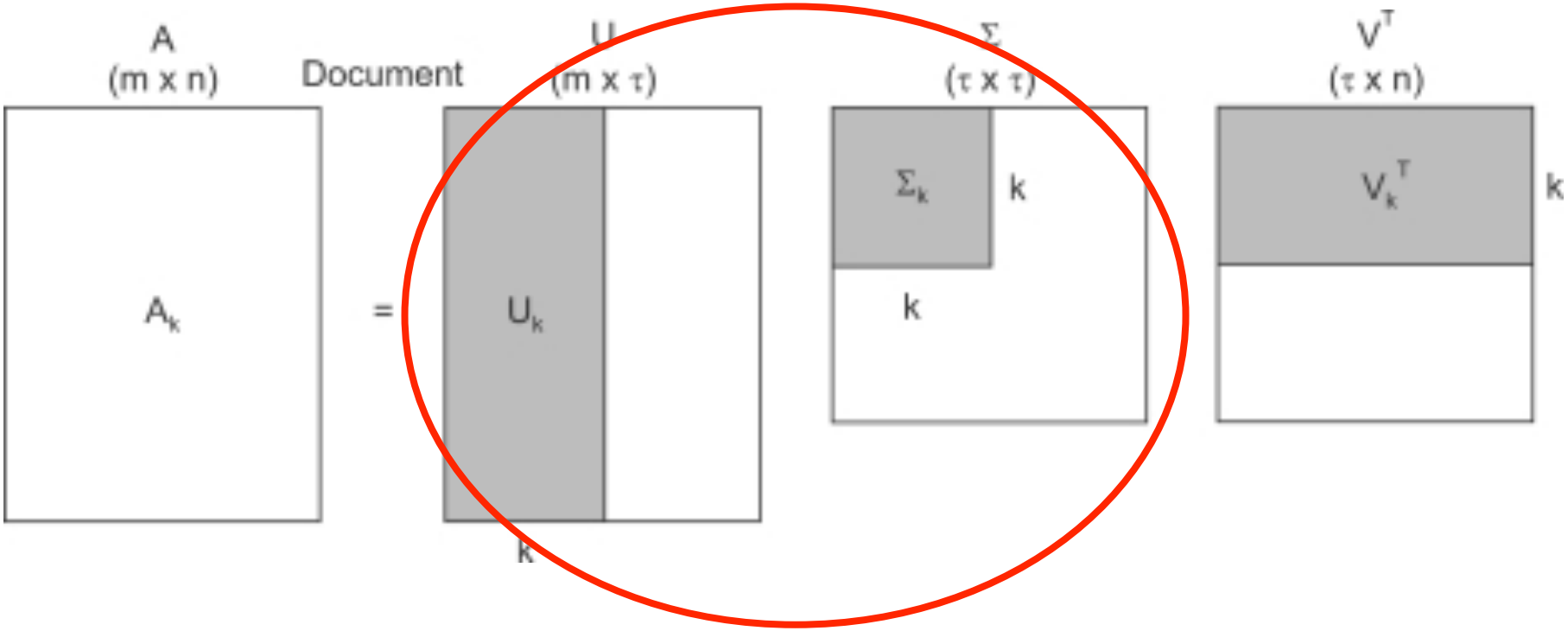
Raw word co-occurrence is rarely satisfactory as a representation

- All words are treated as equally informative
  - the, big, metallic, biophosphorescence
- Vector length is proportional to vocabulary size
  - Eventually issues with computation and space
- Infrequent words have overly-sparse vectors

Standard Approach: Reduce the dimensionality using the Singular Value Decomposition (SVD)



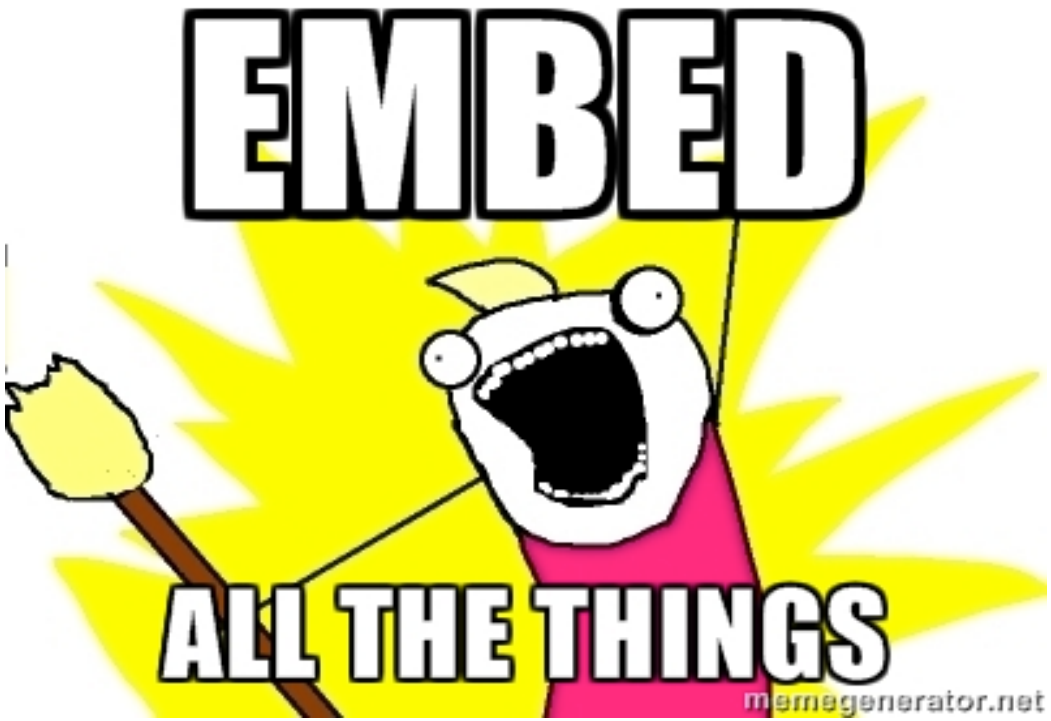
Standard Approach: Reduce the dimensionality using the Singular Value Decomposition (SVD)



Typically, U * Σ is used as the vector space.

State of the Art: Reduce dimensionality with Neural Embeddings (word2vec)

also known as



word2vec

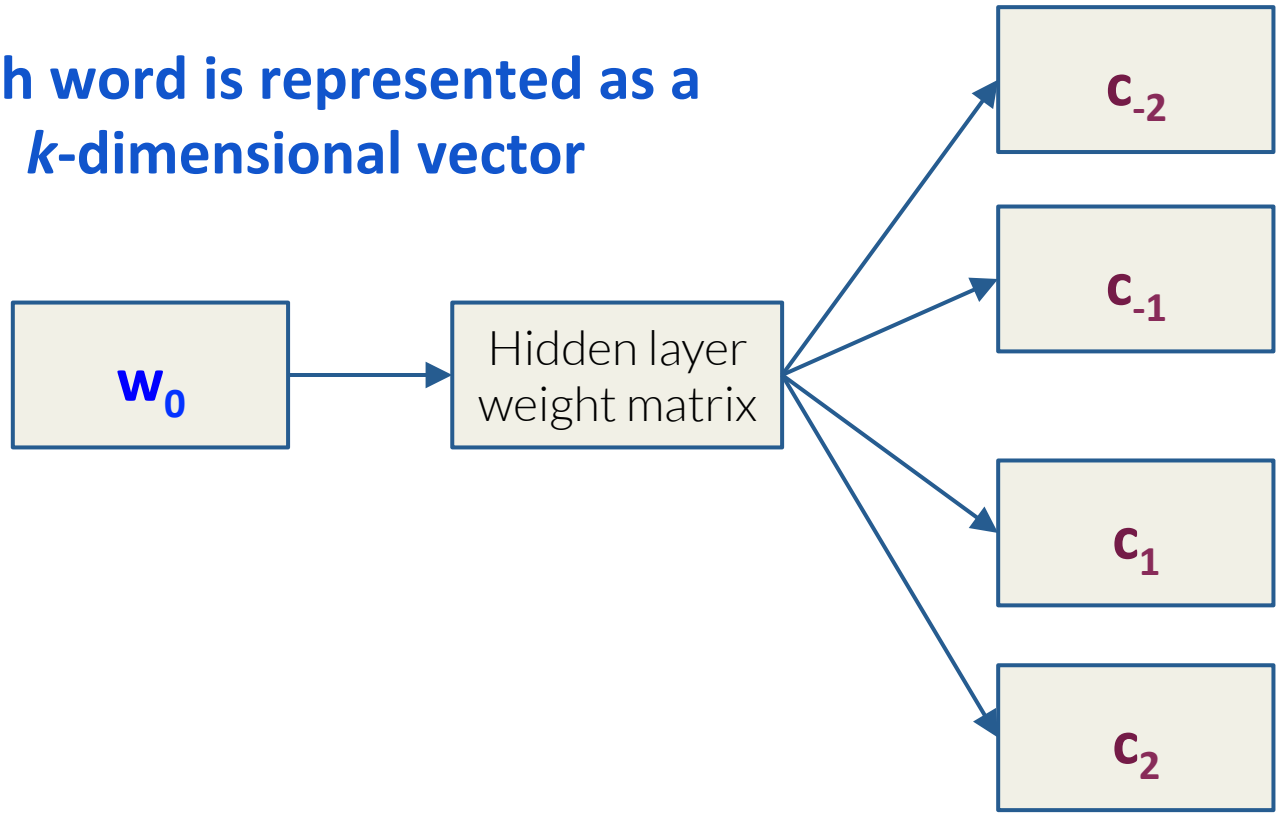
More a software system than an algorithm

- Training methods
  - Negative Sampling
  - Hierarchical Softmax
- Context representations
  - Continuous Bag of Words (CBoW)
  - Skip grams

(Mikolov et al., 2013a,b,c)

word2vec: a neural look

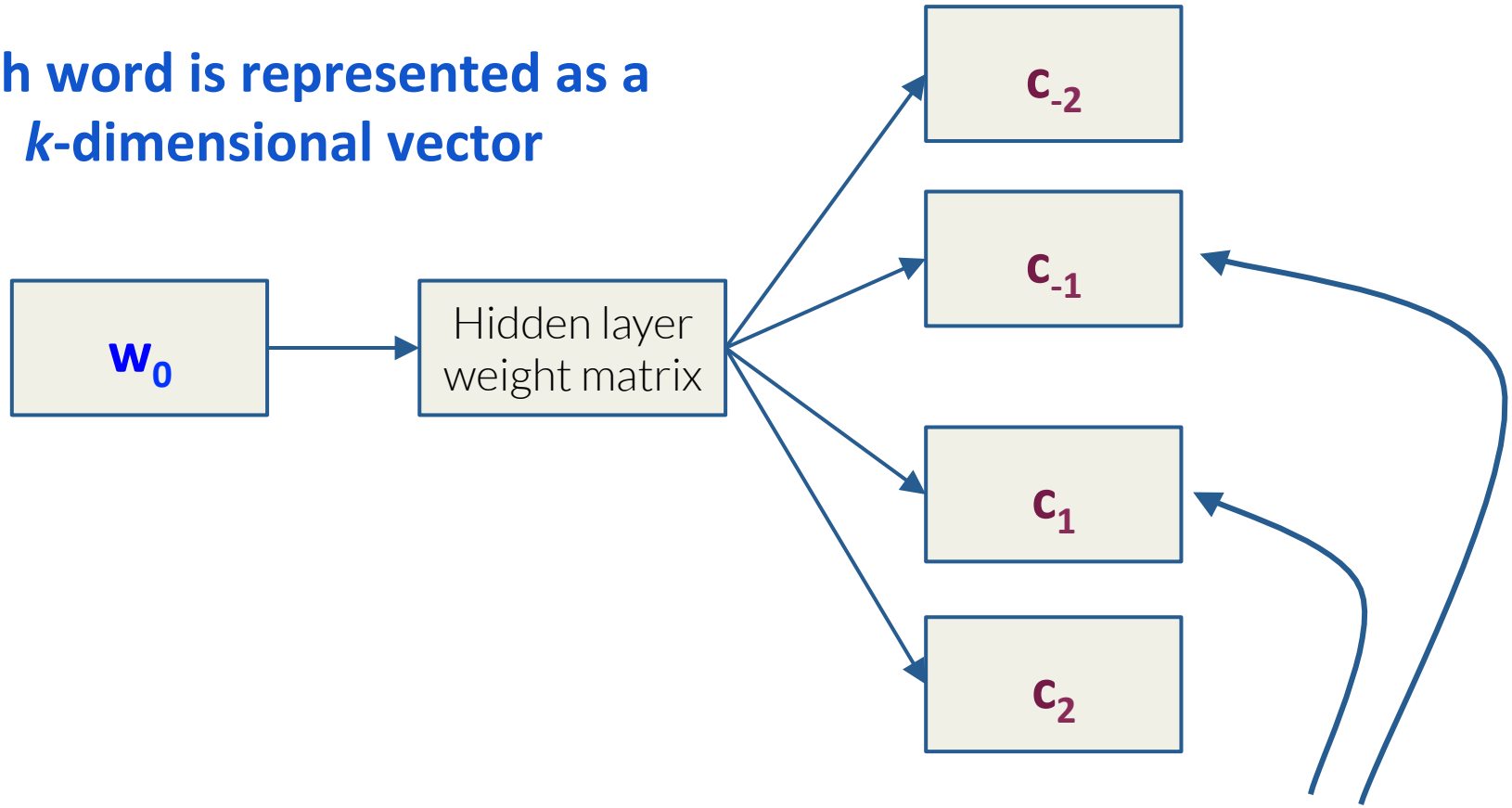
Each word is represented as a *k*-dimensional vector



(Mikolov et al., 2013a,b,c)

word2vec: a neural look

Each word is represented as a *k*-dimensional vector

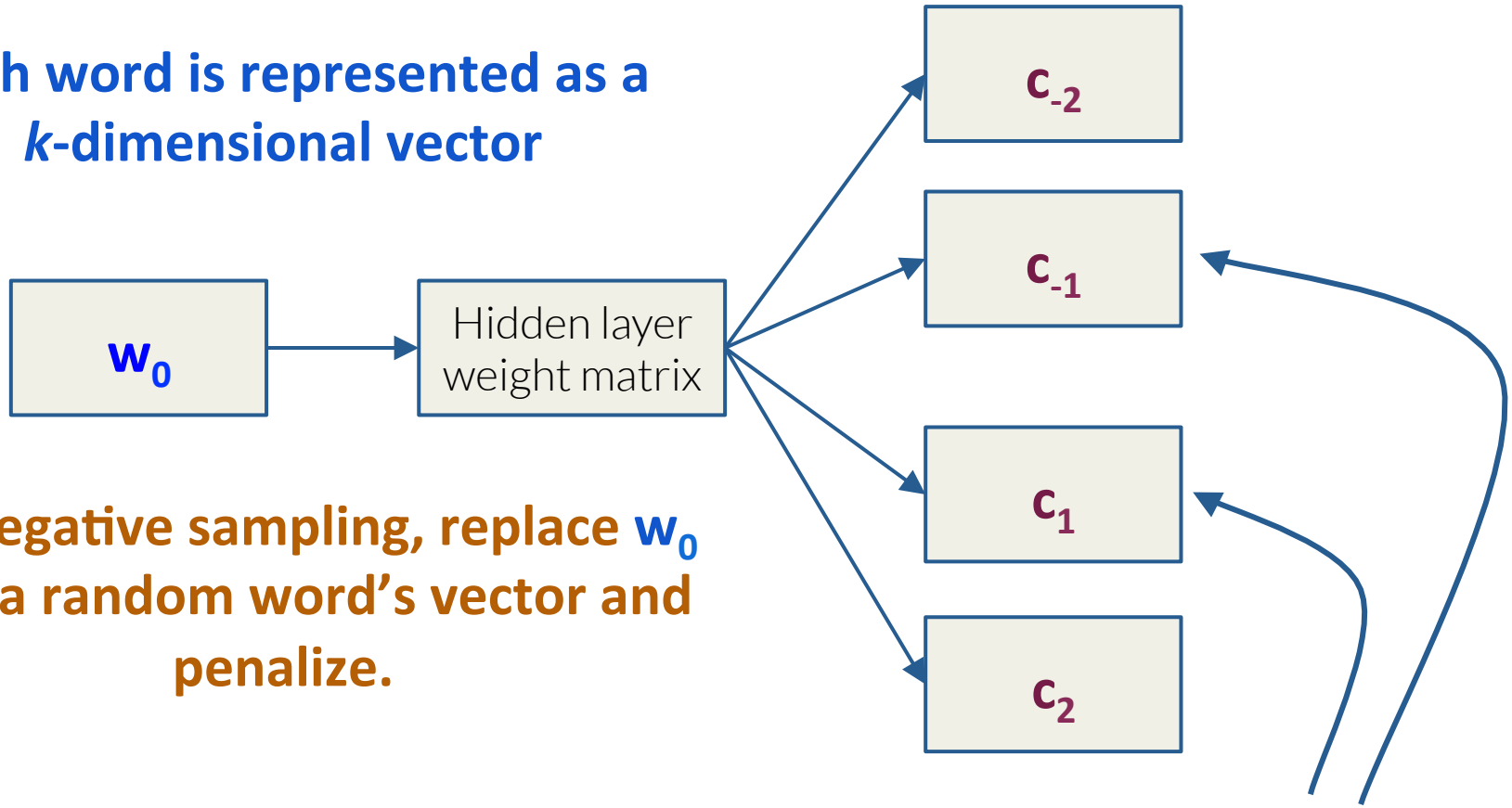


The system is trained to predict the representations for context words before and after

(Mikolov et al., 2013a,b,c)

word2vec: a neural look

Each word is represented as a *k*-dimensional vector



For negative sampling, replace w_0 with a random word's vector and penalize.

The system is trained to predict the representations for context words before and after

(Mikolov et al., 2013a,b,c)

word2vec ≅ implicitly factorizing PMI-weighted word-context matrix

Key Implication: word2vec is building upon existing techniques by using a new decomposition

(Levy and Goldberg, 2014)

Huge gains from using embeddings!

	RG	WordSim	MEN	TOEFL
PMI+SVD	.70	.70	.72	.76
word2vec	.83	.78	.80	.86

Performance improvement over SVD-based methods is consistent across many tasks*

(Baroni et al., 2014)

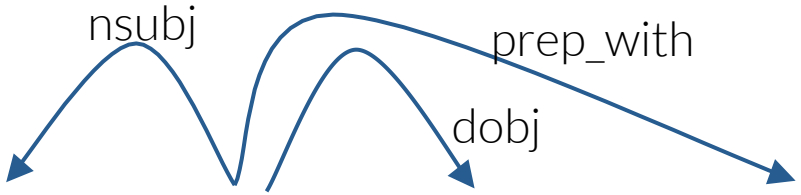
Could we get better performance with syntactic contexts?

Australian scientist discovers star with telescope

(Levy and Goldberg, 2014)

Could we get better performance with syntactic contexts?

Australian scientist discovers star with telescope



Australian scientist discovers star with telescope

(Levy and Goldberg, 2014)

Dependency-based embeddings capture functional information

Target Word	BoW5	BoW2	Deps
batman	nightwing	superman	superman
	aquaman	superboy	superboy
	catwoman	aquaman	supergirl
	superman	catwoman	catwoman
	manhunter	batgirl	aquaman
hogwarts	dumbledore	evernight	sunnydale
	hallows	sunnydale	collinwood
	half-blood	garderobe	calarts
	malfoy	blandings	greendale
	snape	collinwood	millfield
turing	nondeterministic	non-deterministic	pauling
	non-deterministic	finite-state	hotelling
	computability	nondeterministic	heting
	deterministic	buchi	lessing
	finite-state	primality	hamming

No quantitative results on standard benchmarks

(Levy and Goldberg, 2014)

Glove: capture the ratio of co-occurrence probabilities

word2vec: $\vec{w} \cdot \vec{c}^{T} = \text{pmi}(w, c) - \log k$

GloVe: $\vec{w} \cdot \vec{c}^{T} \cdot b_w \cdot b_c = \log(\#(w, c))$

Key insight: the context vector provides insight into so a word representation is $w + c$

(Mikolov et al., 2013; Pennington et al., 2014; Levy and Goldberg, 2015)

GloVe had initially impressive performance at word similarity

	MC	RG	SCWS	Rare Words
SVD	.727	.751	.565	.370
word2vec	.652	.697	.581	.372
GloVe	.727	.778	.529	.381

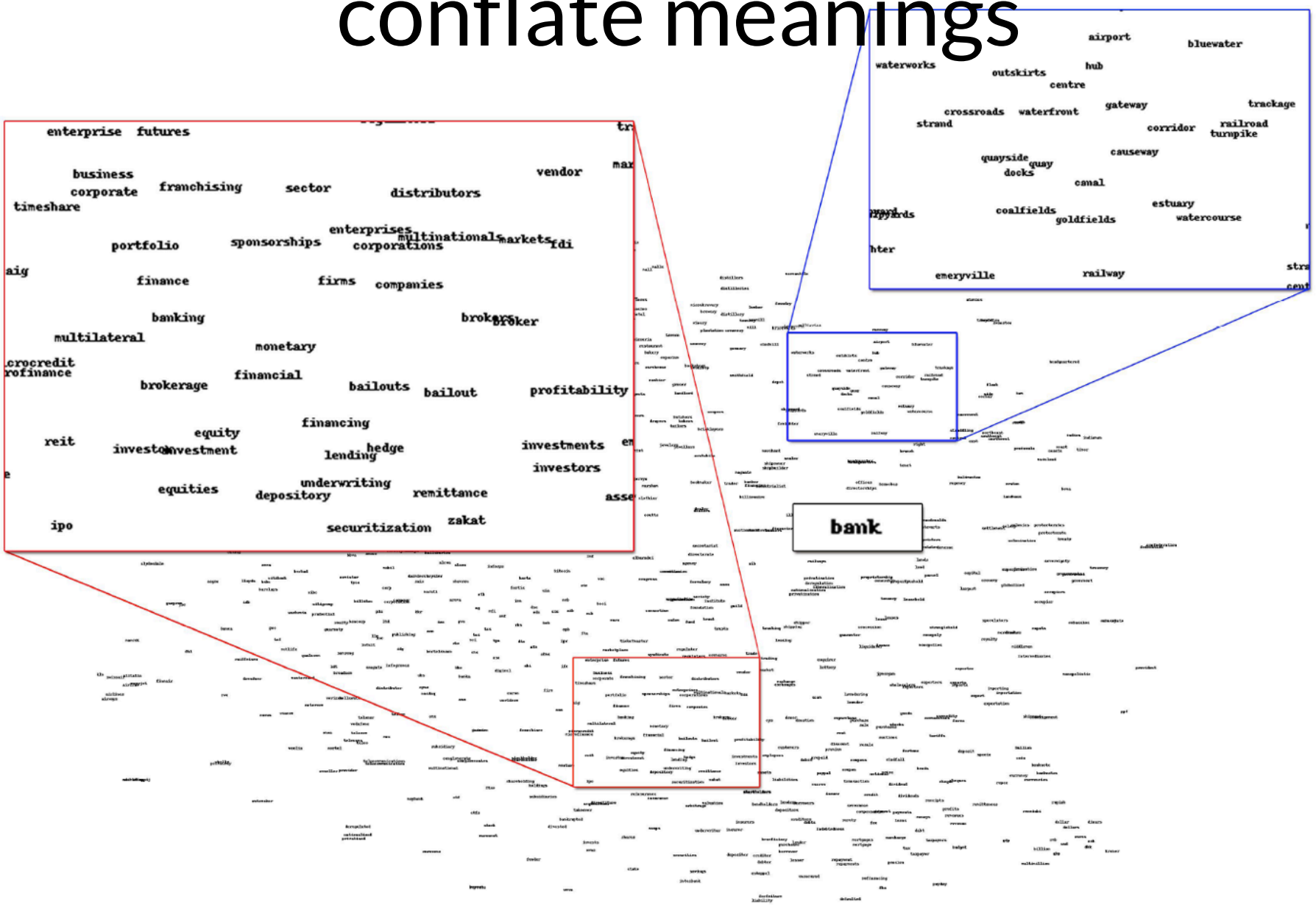
(Pennington et al., 2014)

However under equivalent tuning, word2vec performs better

	Word Sim	MEN	Rad. et al. (2011)	Rare Words	SimLex
PPMI	.755	.745	.686	.462	.393
PMI+SVD	.793	.778	.666	.514	.432
word2vec	.793	.774	.693	.470	.438
GloVe	.725	.729	.632	.403	.398

(Levy and Goldberg, 2015)

Regular embeddings still conflate meanings



Incorporating senses* seems to improve performance

	SCWS	RG	MEN	SimLex
word2vec	.657	.694	.707	.311
Gaussian Embeddings (Vilnis and McCallum, 2015)		.710	.713	.322
TWI (Liu et al. 2015)	.681			

But results vary based on test setup

	SCWS	RG	MEN	SimLex
word2vec	.657	.694	.707	.311
Gaussian Embeddings (Vilnis and McCallum, 2015)		.710	.713	.322
TWI (Liu et al. 2015)	.681			

	SCWS	WordSim	MEN	SimLex
PMI+SVD		.793	.778	.432
word2vec	.581	.793	.774	.438

Many other sense-based embeddings never evaluate on similarity

(Pennington et al., 2014; Levy and Goldberg, 2015)

Results suggest that more dimensions in word vectors can compensate for conflating meanings

	NER	Semantic Relatedness	Sentiment
word-embeddings (50 dims)	.852	.748	.747
sense-embeddings (50 dims)	.854	.762	.750
word-embeddings (100 dims)	.867	.770	.763

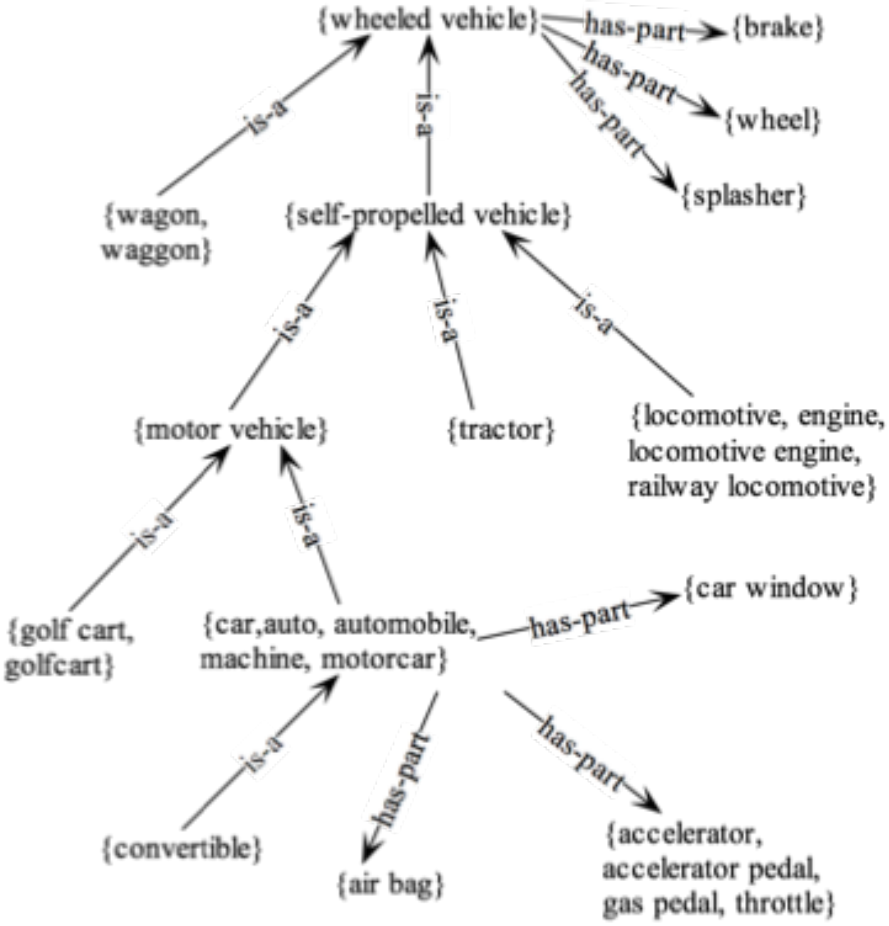
(Li and Jurafsky, 2015)

Learning-Approach Recap

- Nothing magic in the representation
  - similar to SVD with PMI-weighted matrix
- word2vec state of the art for most use cases
  - But dependency-based relations may be useful in some circumstances
  - Also, one of the fastest to train
- Sense-aware representations have a yet to show a clear benefit

What if we already know something about the words?

The structure of WordNet, Wikipedia, and other knowledge bases can be used to measure word similarity



Great for when you need a similarity value

Not as great when you need a representation to use, unless you create one

Wikipedia links create a knowledge graph with edges between related pages

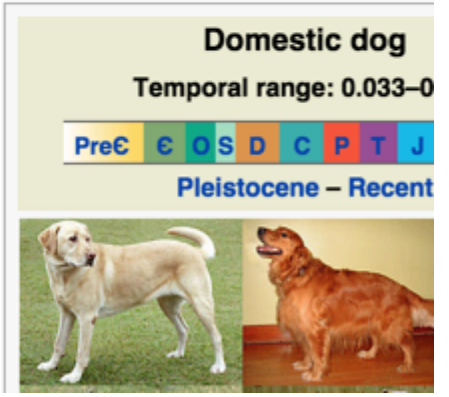
Dog

From Wikipedia, the free encyclopedia

This article is about the domestic dog. For related species known as "dogs", see Canidae. For other uses, see Dog (disambiguation). "Doggie" redirects here. For the Danish artist, see Doggie (artist).

The **domestic dog** (*Canis lupus familiaris* or *Canis familiaris*) is a domesticated canid which has been selectively bred for millennia for various behaviors, sensory capabilities, and physical attributes.[2]

Although initially thought to have originated as a manmade variant of an extant canid species (variously supposed as being the dhole,[3] golden jackal,[4] or gray wolf[5]), extensive genetic studies undertaken during the 2010s indicate that dogs diverged from other wolf-like canids in Eurasia 40,000 years ago.[6] Being the oldest domesticated animals, their long association with people has allowed dogs to be uniquely attuned to human behavior,[7] as well as thrive on a starch-rich diet which would be inadequate for other canid species.[8]



Ideal for path-based measures of similarity and for random walks!

WikiRelate: Apply WordNet measures on Wikipedia's graph

Best results with Leacock & Codorow's method:
-log(path_length($page_1$, $page_2$) / max_depth)

	RG	MC	WordSim-353
L&C (Wikipedia)	.41	.54	.48

(Leacock and Chodorow, 1998; Strube and Ponzetto, 2006)

WikiRelate: Apply WordNet measures on Wikipedia's graph

Best results with Leacock & Codorow's method:
-log(path_length($page_1$, $page_2$) / max_depth)

	RG	MC	WordSim-353
L&C (Wikipedia)	.41	.54	.48
L&C (WordNet)	.82	.86	.34

Large amount of noise in Wikipedia's graph creates issues for similarity-specific calculations. I.e,. difficult to tell edges and nodes are important.

(Leacock and Chodorow, 1998; Strube and Ponzetto, 2006)

Idea: Identify important pages in Wikipedia using Personalized PageRank

- Given a page *p*, find all wiki-linked pages to *p* and initialize the PPR vector to these pages
  - Optionally prune (a) pages with spaces in the name and (b) pages account for fewer than x% of the links
- Run PPR and compare vectors

(Agirre et al., 2009)

Idea: Identify important pages in Wikipedia using Personalized PageRank

- Given a page *p*, find all wiki-linked pages to *p* and initialize the PPR vector to these pages
  - Optionally prune (a) pages with spaces in the name and (b) pages account for fewer than x% of the links

- Run PPR and compare vectors

	MC	WordSim-353
PPR	.60	.45
WikiRelate	.54	.48

(Agirre et al., 2009)

Idea: Identify important pages in Wikipedia using Personalized PageRank

- Given a page *p*, find all wiki-linked pages to *p* and initialize the PPR vector to these pages
  - Optionally prune (a) pages with spaces in the name and (b) pages account for fewer than x% of the links

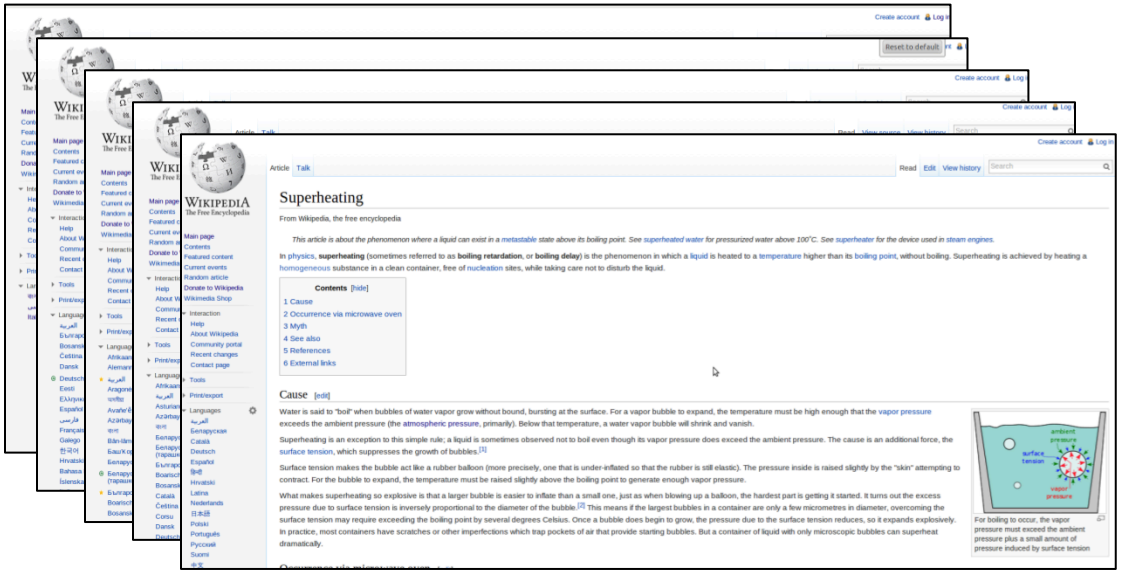
- Run PPR and compare vectors

	MC	WordSim-353
PPR	.60	.45
WikiRelate	.54	.48
ESA	.72	.75

(Agirre et al., 2009; Gabrilovich and Markovitch, 2007)

(Still) State of the Art for Wikipedia: Explicit Semantic Analysis

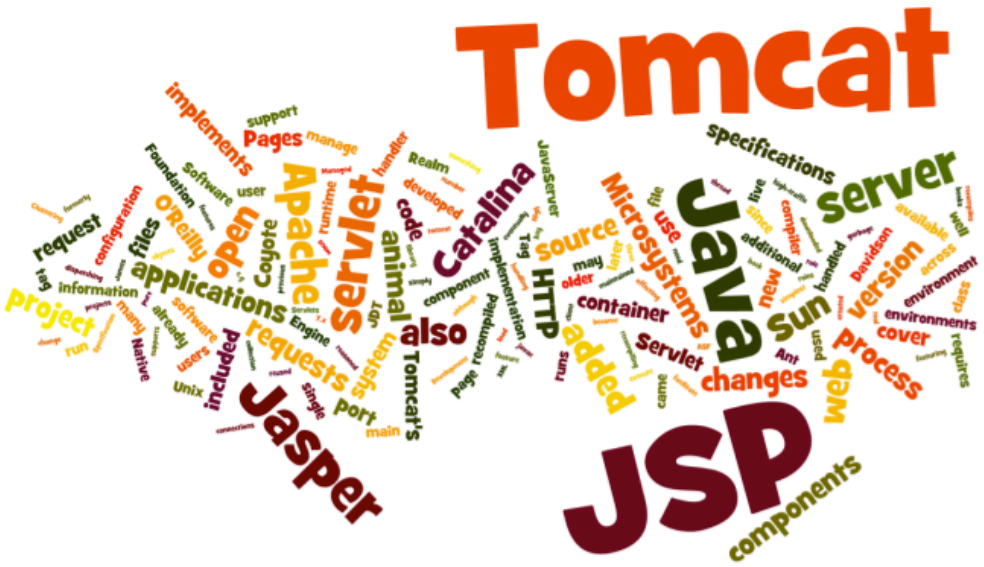
Consider each Wikipedia article as a concept



$\{c_1, c_2, c_3, \ldots c_N\}$ where N is the number of articles in Wikipedia

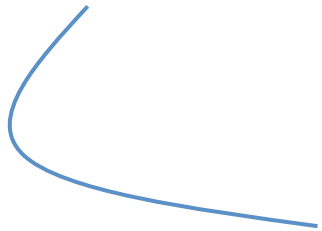
Gabrilovich and Markovitch (2007)

WIKIPEDIA *The Free Encyclopedia* articles for Tomcat



Explicit Semantic Analysis (ESA)

For a given word (e.g., equipment) calculate an inverted index entry to all the *N* documents:



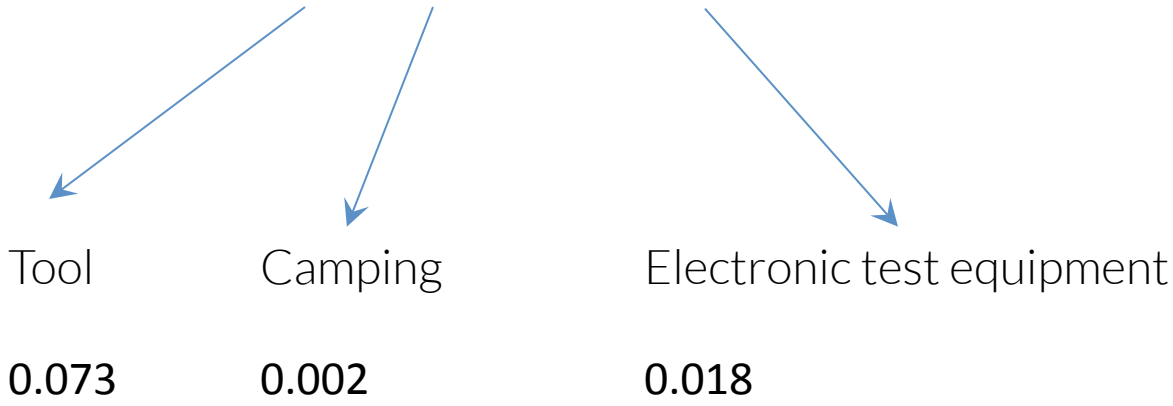
$$T[i,j] = tf(t_i, d_j) \cdot \log \frac{n}{df_i},$$

$$tf(t_i, d_j) = \begin{cases} 1 + \log count(t_i, d_j), & if\ count(t_i, d_j) > 0 \\ 0, & otherwise \end{cases}$$

Explicit Semantic Analysis (ESA)

For a given word (e.g., equipment) calculate an inverted index entry to all the *N* documents:

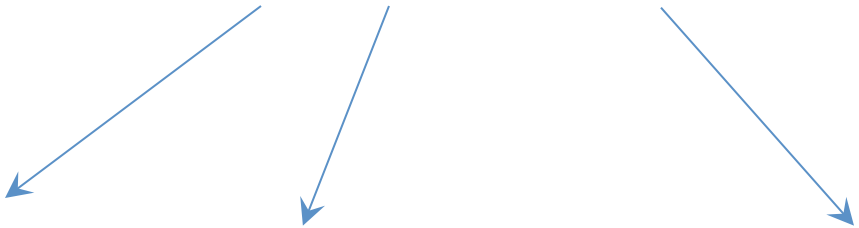
$$\{c_1, c_2, c_3, \ldots c_N\}$$



Explicit Semantic Analysis (ESA)

For a given word (e.g., equipment) calculate an inverted index entry to all the *N* documents:

$$\{c_1, c_2, c_3, \ldots c_N\}$$



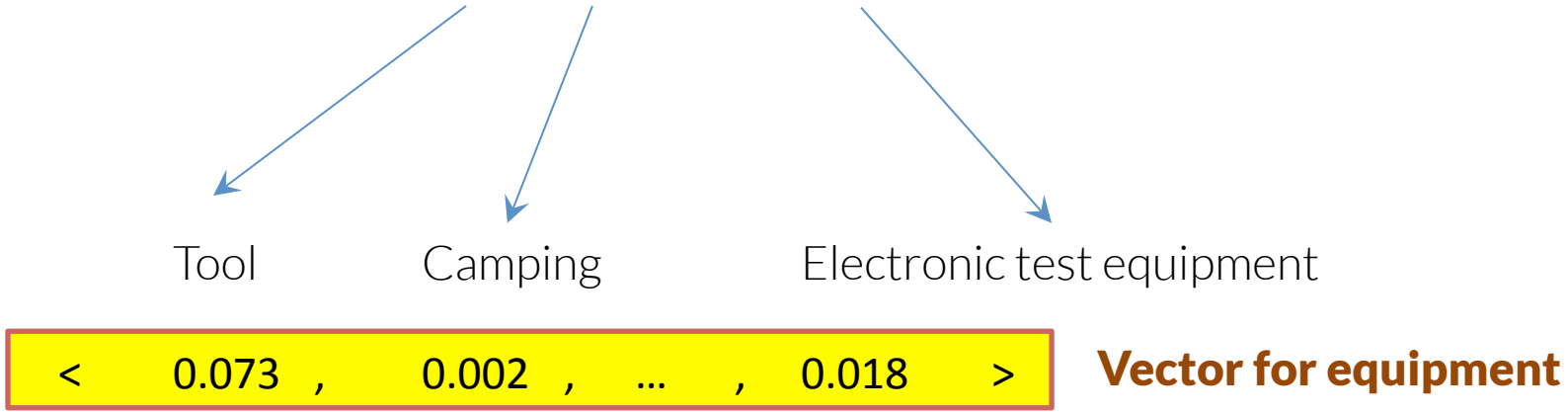
Tool Camping Electronic test equipment

< 0.073 , 0.002 , ... , 0.018 >

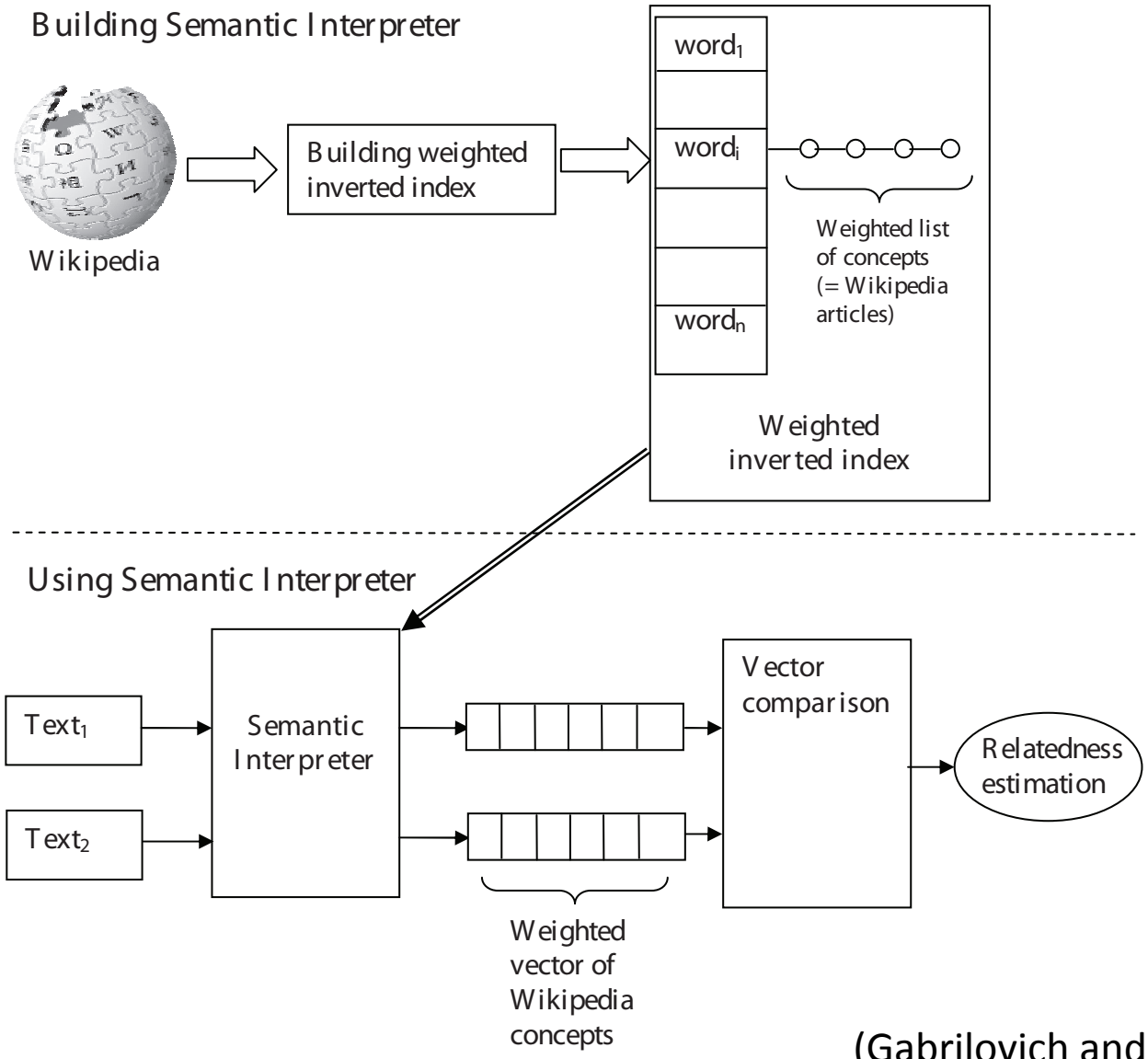
Explicit Semantic Analysis (ESA)

For a given word (e.g., equipment) calculate an inverted index entry to all the *N* documents:

$$\{c_1, c_2, c_3, \ldots c_N\}$$



ESA pipeline



(Gabrilovich and Markovitch, 2007)

ESA (example)

#	Input: *"equipment"*	Input: *"investor"*
1	Tool	Investment
2	Digital Equipment Corporation	Angel investor
3	Military technology and equipment	Stock trader
4	Camping	Mutual fund
5	Engineering vehicle	Margin (finance)
6	Weapon	Modern portfolio theory
7	Original equipment manufacturer	Equity investment
8	French Army	Exchange-traded fund
9	Electronic test equipment	Hedge fund
10	Distance Measuring Equipment	Ponzi scheme

(Gabrilovich and Markovitch, 2007)

ESA: text modeling

- Centroid of the vectors representing the individual words

- Can be weighted:

$$\sum_{w_i \in T} v_i k_j$$

TFIDF weight of word i in the text

Inverted index for word i

Wiktionary provides links with more semantic structure

English [edit]

Alternative forms [edit]

- darg, dawg (*dialectal*); doggie, doggy (*childish*)

Noun [edit]

dog (*plural* **dogs**)

1. A mammal, *Canis lupus familiaris*, that has been domesticated for thousands of years, of highly variable appearance due to human breeding. [quotations ▼]

   *The **dog** barked all night long.*

2. A male dog, wolf or fox, as opposed to a bitch (often attributive). [quotations ▼]
3. (*derogatory*) A dull, unattractive girl or woman.

   *She's a real **dog**.*

4. (*slang*) A man (derived from definition 2).

   *You lucky **dog**! He's a sly **dog**.*

Wiktionary provides links with more semantic structure

English [edit]

Alternative forms [edit]

- darg, dawg (*dialectal*); doggie, doggy (*childish*)

Noun [edit]

dog (*plural* **dogs**)

1. A mammal, *Canis lupus familiaris*, that has been domesticated for thousands of years, of highly variable appearance due to human breeding. [quotations ▼]
   *The **dog** barked all night long.*
2. A male dog, wolf or fox, as opposed to a bitch (often attributive). [quotations ▼]
3. (*derogatory*) A dull, unattractive girl or woman.

Synonyms [edit]

- (*animal*): taxonomic names: *Canis familiaris*, *Canis domesticus*, *Canis familiarus domesticus*, *Canis canis*, *Canis aegyptius*, *Canis familiarus aegyptius*, *Canis melitaeus*, *Canis familiarus melitaeus*, *Canis molossus*, *Canis familiarus molossus*, *Canis saultor*, *Canis familiaris saultor*
- (*animal*): domestic dog, hound, canine; *see also* Wikisaurus:dog
- (*male*): stud, sire
- (*man*): bloke (*British*), chap (*British*), dude, fellow, guy, man; *see also* Wikisaurus:man
- (*morally reprehensible person*): cad, bounder, blackguard, fool, hound, heel, scoundrel
- (*mechanical device*): click, detent, pawl
- (*metal support for logs*): andiron, firedog, dogiron

Coordinate terms [edit]

- (*male adult dog*): bitch, pup, puppy

Hyponyms [edit]

- (*animal*): Afghan hound, bloodhound, chihuahua, coonhound, dachshund, deerhound, foxhound, gazehound, German shepherd, greyhound, hound, Irish Wolfhound, Norwegian Elkhound, otterhound, pointer, poodle, retriever, Russian Wolfhound, scenthound, setter, sheepdog, shepherd, sighthound, spaniel, staghound, terrier, wolfhound

Hypernyms [edit]

- (*animal*): canid

Ideal for path-based measures of similarity and for random walks!

Random Walks are still useful if you use a semantically structured resource

	RG
ADW w/ Wiktionary (Pilehvar and Navigli, 2015)	.920
ADW w/ WordNet (Pilehvar et al. 2013)	.868
PPR w/ WordNet (Hughes and Ramage, 2007)	.838
PPR w/ WordNet (Agirre et al., 2009)	.830
ESA (Gabrilovich and Markovitch, 2007)	.749
WikiRelate (Strube and Ponzetto, 2006)	r = 0.53

Word vectors don't need to be learned either!

Idea: create binary vectors of whether a word satisfies a set properties from knowledge bases

- WordNet: is hypernym of *x*
- FrameNet: evokes frame *x*
- Sentiment: evokes emotion or sentiment
- ~172K features total

Optionally compress vectors using an SVD

(Faruqui and Dyer, 2015)

Word vectors don't need to be distributional either!

	RG	SimLex	WordSim-353
word2vec	.728	.436	.656
GloVe	.766	.369	.605
LSA	.770	.496	.673
Ling (full)	.778	.566	.446
Ling (with SVD)	.670	.576	.454

Significant gains in similarity just by encoding knowledge bases in a vector format

(Faruqui and Dyer, 2015)

Word vectors don't need to be distributional either!

	RG	SimLex	WordSim-353
word2vec	.728	.436	.656
GloVe	.766	.369	.605
LSA	.770	.496	.673
Ling (full)	.778	.566	.446
Ling (with SVD)	.670	.576	.454
ADW	.868		

There may still be better ways to encode knowledge though

(Pilehvar et al., 2013; Faruqui and Dyer, 2015)

What if we knew something but still wanted to learn?

Idea: modify vectors learning (or representations) to match desired properties of knowledge bases

Impose constraints such as

- Sim(word, synonym) > Sim(word, antonym)
- Similarity is greater when concepts are more categorically related (e.g., using hypernyms)

Constraints could be added during learning or could be used to retrofit already-learned vectors

(Iacobacci et al., 2015; Liu et al., 2015; Faruqui et al, 2015)

Idea: modify vectors learning (or representations) to match desired properties of knowledge bases

	Where is knowledge added?	RG	TOEFL	WordSim-353
word2vec	N/A	.728	83.75	.709
Li et al., (2015)	Learning		87.5	.727
Faruqui et al., (2015)	Representation	.778	100	.700
Iacobacci et al., (2015)	Similarity Func.	.871		.779

Significant opportunities to add knowledge at different stages, with the ability to tune the representation or how it is used for a specific task

(Iacobacci et al., 2015; Liu et al. 2015; Faruqui et al, 2015)

Phrase similarity

Compositionality

Moving from words to phrases, sentences, and larger pieces of texts

How would we compare...

“the usual morning cup of joe”
“drip coffee with freshly-ground arabica beans”

How would we compare...

“the usual morning cup of joe”
“drip coffee with freshly-ground arabica beans”

“must do our utmost”
“must make every effort”

How would we compare...

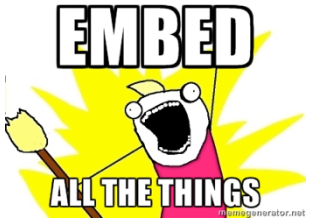
"the usual morning cup of joe"
"drip coffee with freshly-ground arabica beans"

"must do our utmost"
"must make every effort"

Measuring the similarity of the phrases requires understanding each item as a whole.

We need compositionality!

Learn phrase representations directly during embedding!



Directly learns word2vec representations for phrases

- First detects phrases in the training corpus by using a simple frequency-based approach
- Treating these phrases as single tokens, obtains phrase-specific representations

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

Incapable of handling syntactic dependencies or related phrasal constructions

Mikolov et al (2013)

Initial idea: compose from existing word representations

Combining individual words' vectors

Simple average: $p_i = u_i + v_i$

Weighted average: $p_i = \alpha u_i + \beta v_i$

Including one or more distributional neighbors: $\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum \mathbf{n}$

Multiplicative: $p_i = u_i \cdot v_i$

Combined multiplication and addition: $p_i = \alpha u_i + \beta v_i + \gamma u_i v_i$

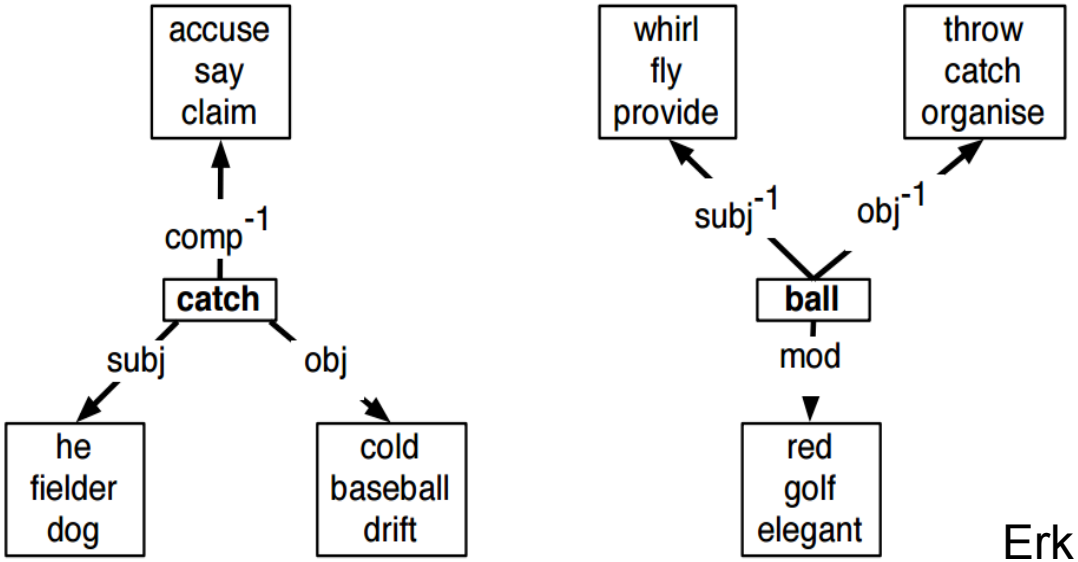
Better at distinguishing high and low semantic similarity

Mitchell and Lapata (2008)

Combine words by taking syntax into account

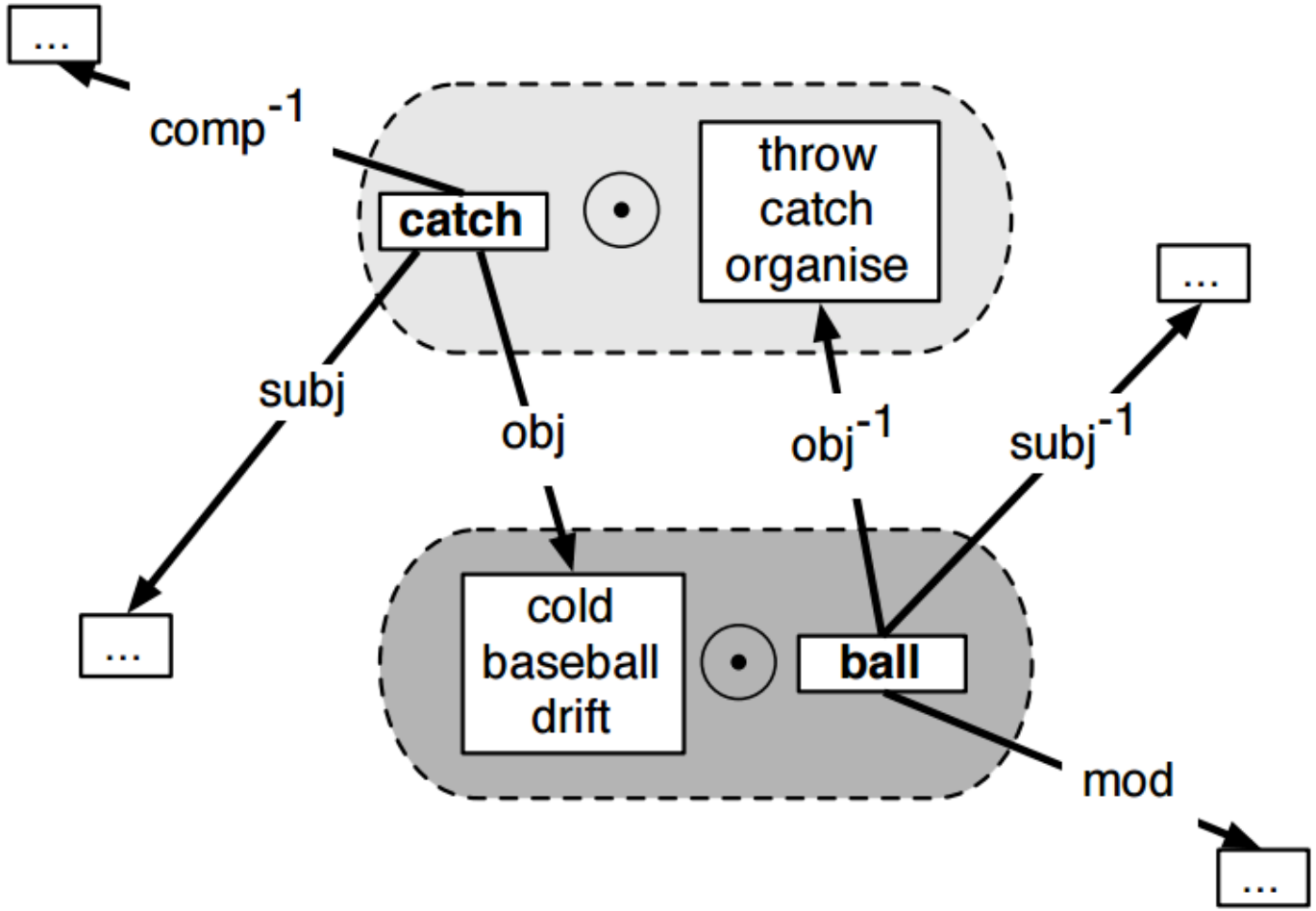
- Integrates lexical information with selectional preferences

- Computes the meaning of a word a in the context of the word b (disambiguates the meaning of a word in the context of another)



Erk and Pado (2008)

Combine words by taking syntax into account



Erk and Pado (2008)

Moving beyond element-wise composition

Simple average: $\mathbf{z}_i = \mathbf{u}_j + \mathbf{v}_k$

Adjectives as matrices: $\mathbf{z}_i = \mathbf{U}_j\mathbf{v}_k$

- Learn each adjective's $\mathbf{U}$ by comparing vectors when adjective is and isn't present.

(Baroni and Zamparelli, 2010; Guevara, 2010)

Moving beyond element-wise composition

Simple average: $\mathbf{z}_i = \mathbf{u}_j + \mathbf{v}_k$

Adjectives as matrices: $\mathbf{z}_i = \mathbf{U}_j\mathbf{v}_k$

- Learn each adjective's $\mathbf{U}$ by comparing vectors when adjective is and isn't present.

Composition as matrices: $\mathbf{z}_i = \mathbf{A}\mathbf{u}_j + \mathbf{B}\mathbf{v}_k$

- Estimating A and B is a regression problem with multiple dependent variables. Use a dictionary to find training pairs (u, v, z)!

(Baroni and Zamparelli, 2010; Guevara, 2010; Zanzotto et al., 2010)

Moving beyond element-wise composition

Simple average: $\mathbf{z}_i = \mathbf{u}_j + \mathbf{v}_k$

Adjectives as matrices: $\mathbf{z}_i = \mathbf{U}_j\mathbf{v}_k$

- Learn each adjective's $\mathbf{U}$ by comparing vectors when adjective is and isn't present.

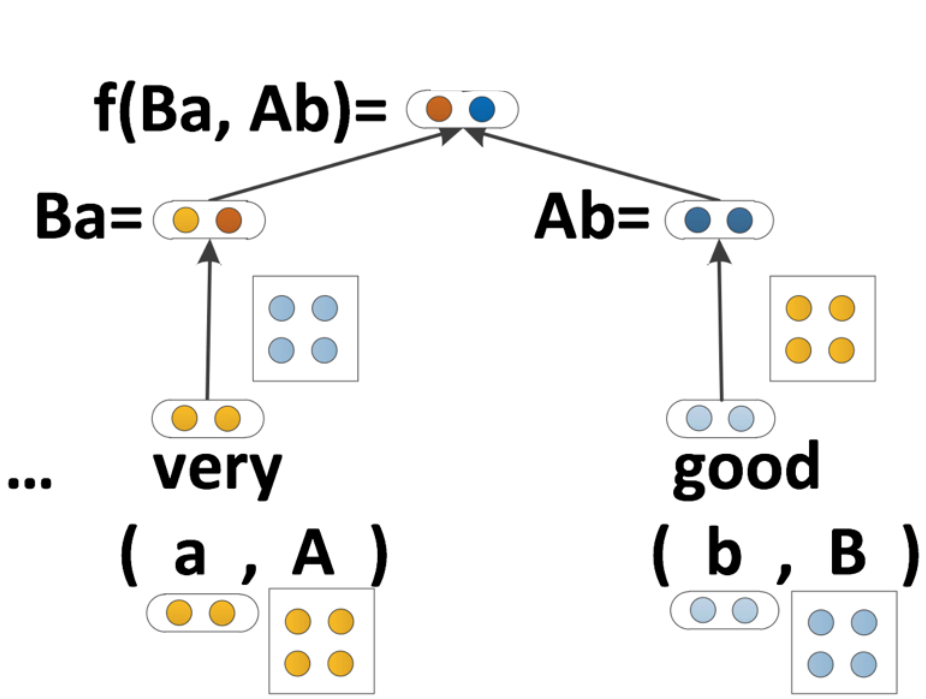
Key insight: composition is decoupled from word type!

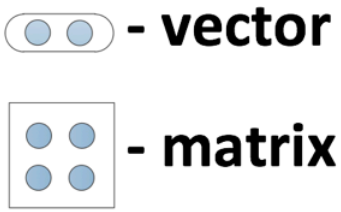
Composition as matrices: $\mathbf{z}_i = \mathbf{A}\mathbf{u}_j + \mathbf{B}\mathbf{v}_k$

- Estimating A and B is a regression problem with multiple dependent variables. Use a dictionary to find training pairs (u, v, z)!

(Baroni and Zamparelli, 2010; Guevara, 2010; Zanzotto et al., 2010)

Compose with a recursive neural net

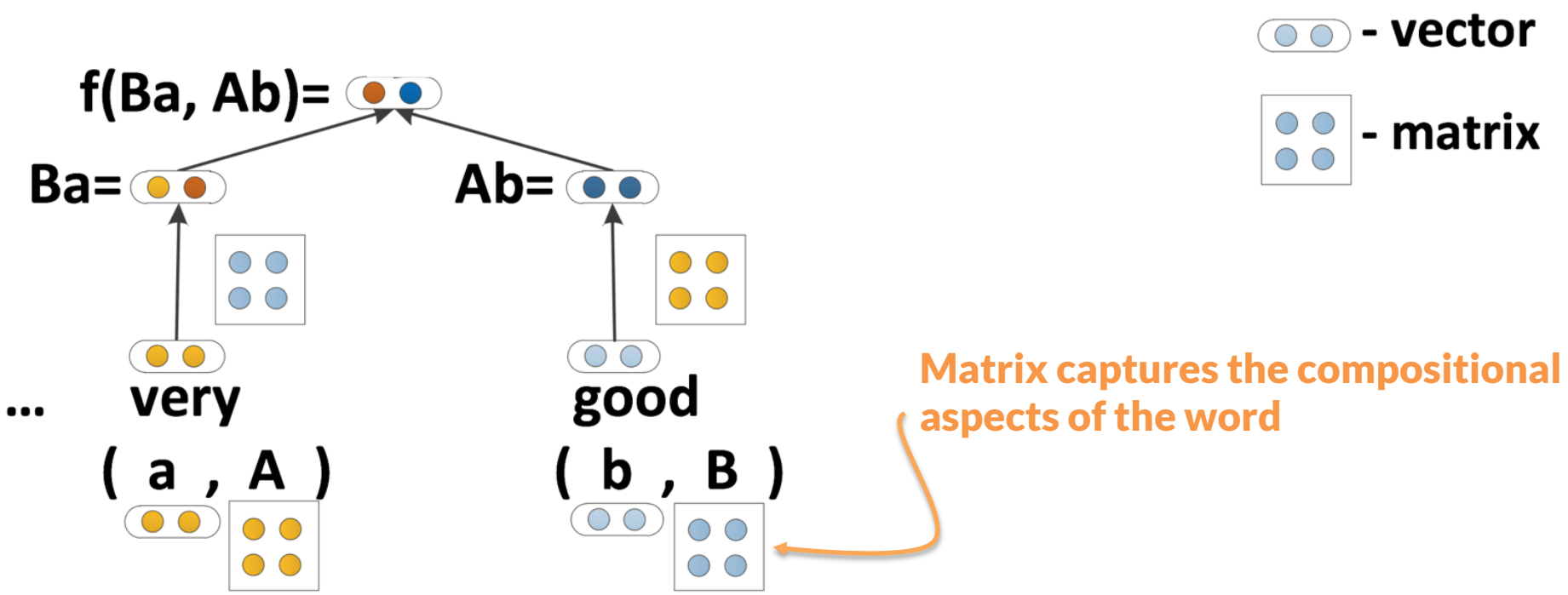




Note: Requires data be parsable.

(Socher et al., 2012)

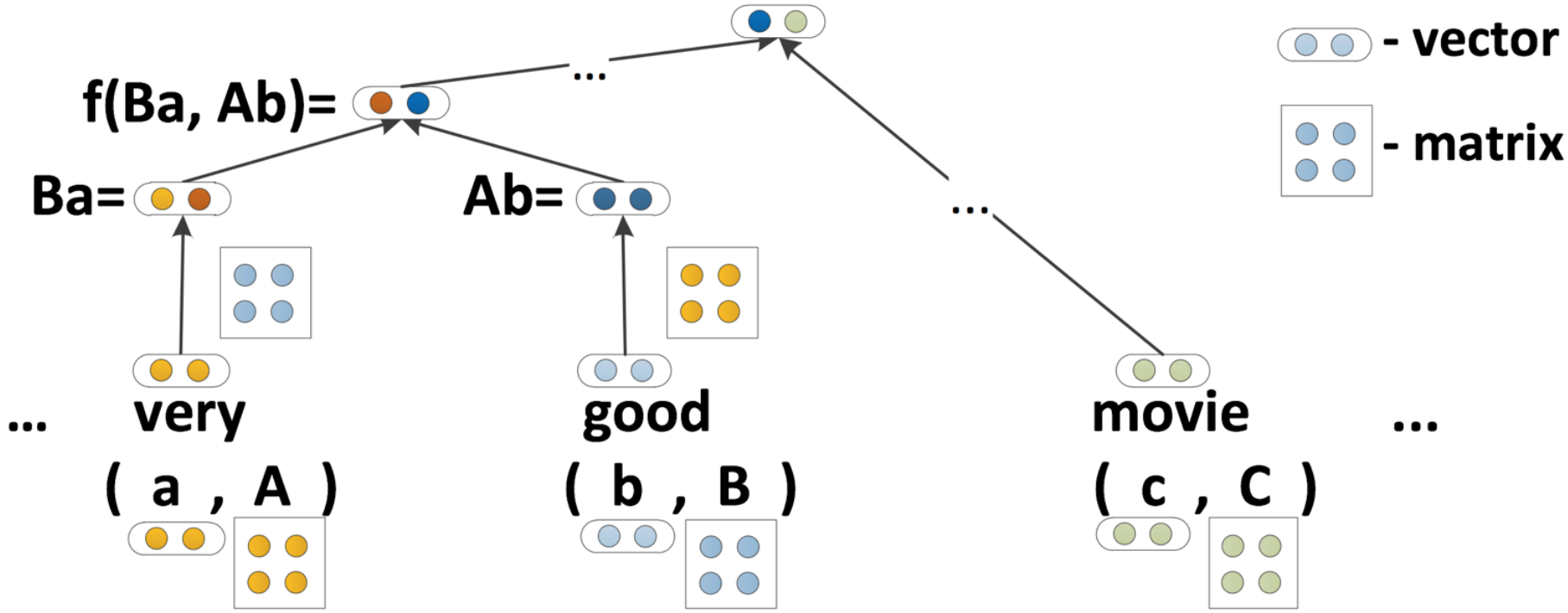
Compose with a recursive neural net



Note: Requires data be parsable.

(Socher et al., 2012)

Compose with a recursive neural net



Not ideal performance in compositionality-specific tasks (Blacoe and Lapata, 2012; Hashimoto et al., 2014) -- partly because the model isn't trained for compositionality!

(Socher et al., 2012)

Idea: Design an RNN with a cost function based on good paraphrase

- Create a paraphrase ranking corpus from PPDB (Ganitkevitch et al., 2013)
- Modify the RNN from Socher et al. (2014) so that the loss function penalizes similar representations of bad paraphrase examples
- Initialize with word2vec, but tune the vectors

(Wieting et al., 2015)

Idea: Design an RNN with a cost function based on good paraphrase

		M&L Bigrams	M&L Paraphrase	Annotated PPDB
word2vec	additive	.39	.36	.20
paragram	additive	.42	.46	.32
paragram	RNN	.47	.52	.40
Hashimoto et al. (2014)		.47	.41	-
Mitchell and Lapata (2010)		.44	-	-

A supervised RNN provides significant benefits over representing phrases using vector addition.

(Wieting et al., 2015)

Sentence Similarity

Sentence similarity is one of the most active areas

Many applications benefit:

- Paraphrasing
- Textual entailment
- Machine translation
- Question Answering

Easy to build models using combinations of string similarity and word-semantics similarity!

Semantic Textual Similarity

- **2012** (A pilot): 35 teams 88 runs
- **2013** (+typed): 34 teams 89 runs
- **2014** (Multilingual):
  English 15 teams 38 runs
  Spanish 9 teams 22 runs
- **2015** (+Pilot on Interpretability):
  English 29 teams 74 runs
  Spanish 7 teams 16 runs
  Interpretable STS 7 teams 29 runs
- **2016** (Interpretable STS)

Semantic Textual Similarity

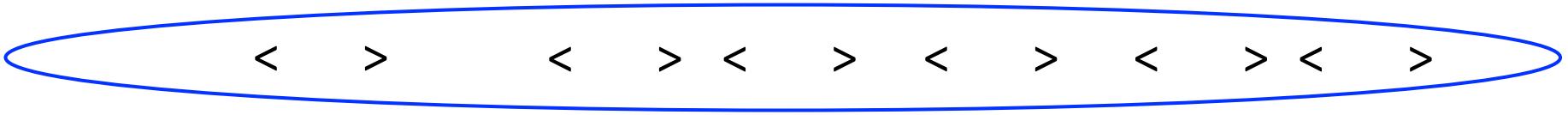
year	dataset	pairs	source
2012	MSRpar	1500	newswire
2012	MSRvid	1500	videos
2012	OnWN	750	glosses
2012	SMTnews	750	MT eval.
2012	SMTeuroparl	750	MT eval.
2013	HDL	750	newswire
2013	FNWN	189	glosses
2013	OnWN	561	glosses
2013	SMT	750	MT eval.
2014	HDL	750	newswire headlines
2014	OnWN	750	glosses
2014	Deft-forum	450	forum posts
2014	Deft-news	300	news summary
2014	Images	750	image descriptions
2014	Tweet-news	750	tweet-news pairs

IAA statistics:

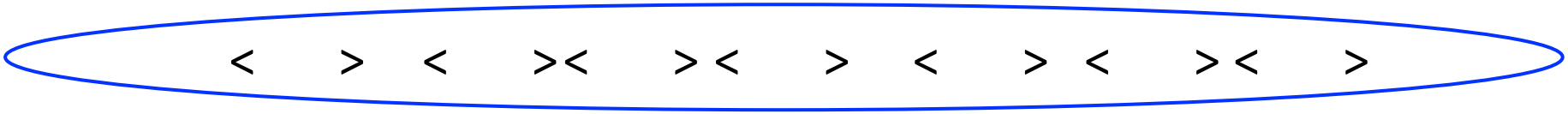
- HDL: 79.4%
- OnWN: 67.2%
- Deft-forum: 58.6%
- Deft-news: 70.7%
- Images: 83.6%
- Tweets-news: 74.4%

Sentence Similarity Techniques

Basic idea: Average vectors of the words in a sentence



Indonesia passenger plane wreckage located in remote Papua



Indonesia Plane Debris Found in Remote Papua Area

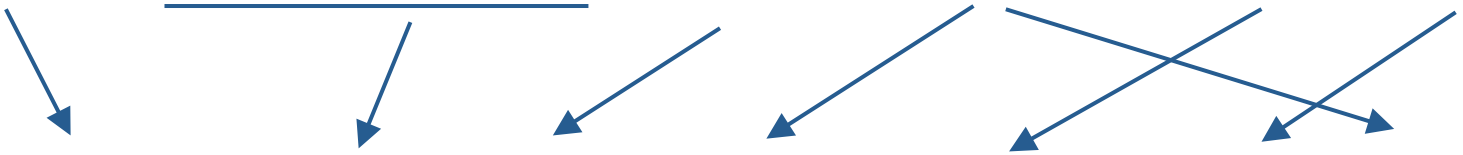
Sentence Similarity Techniques

Alignment

Aggregate the similarities of the closest pairs of words: Corley and Mihalcea (2005)

$$\text{sim}(T_1, T_2) = \frac{\sum_{w \in T_1} \text{maxSim}(w, T_2)\ \text{idf}(w)}{\sum_{w \in T_1} idf(w)}$$

Indonesia passenger plane wreckage located in remote Papua



Indonesia Plane Debris Found in Remote Papua Area

Sentence Similarity Techniques

Simple string-based similarity

- **Substring overlap**
  He is talking on a phone
  He talks on a telephone

Sentence Similarity Techniques

Simple string-based similarity

He is talking on a phone
He talks on a telephone

- **N-gram overlap** (character and **word**)

<begin> He is
He is talking
is talking on
talking on phone
on phone <end>

<begin> He talks
talks on a
on a telephone
A telephone <end>

Sentence Similarity Techniques

Simple string-based similarity

He is talking on a phone
He talks on a telephone

- **N-gram overlap** (**character** and word)

<b>He-	<b>He-
He-	He-
e-i	e-t
-is	-ta
is-	tal
s-t	alk
…	…

Sentence Similarity Techniques

Usually feature-based regression models

e.g., UKP (best system in STS-12)

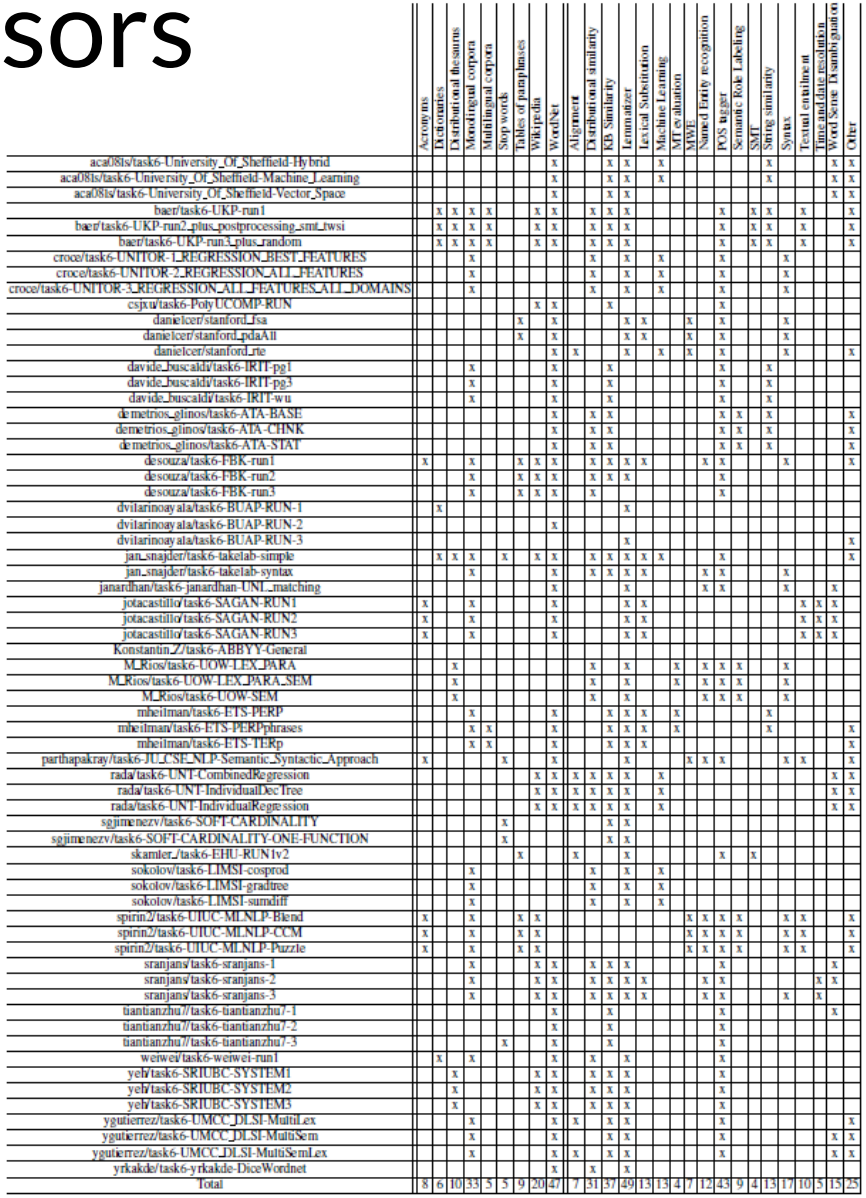
String-based similarity: character n-gram, GST, etc.

Semantic similarity: WordNet-based approaches, ESA, etc.

Other features: POS n-gram, SMT, etc.

Most STS systems are multi-feature regressors

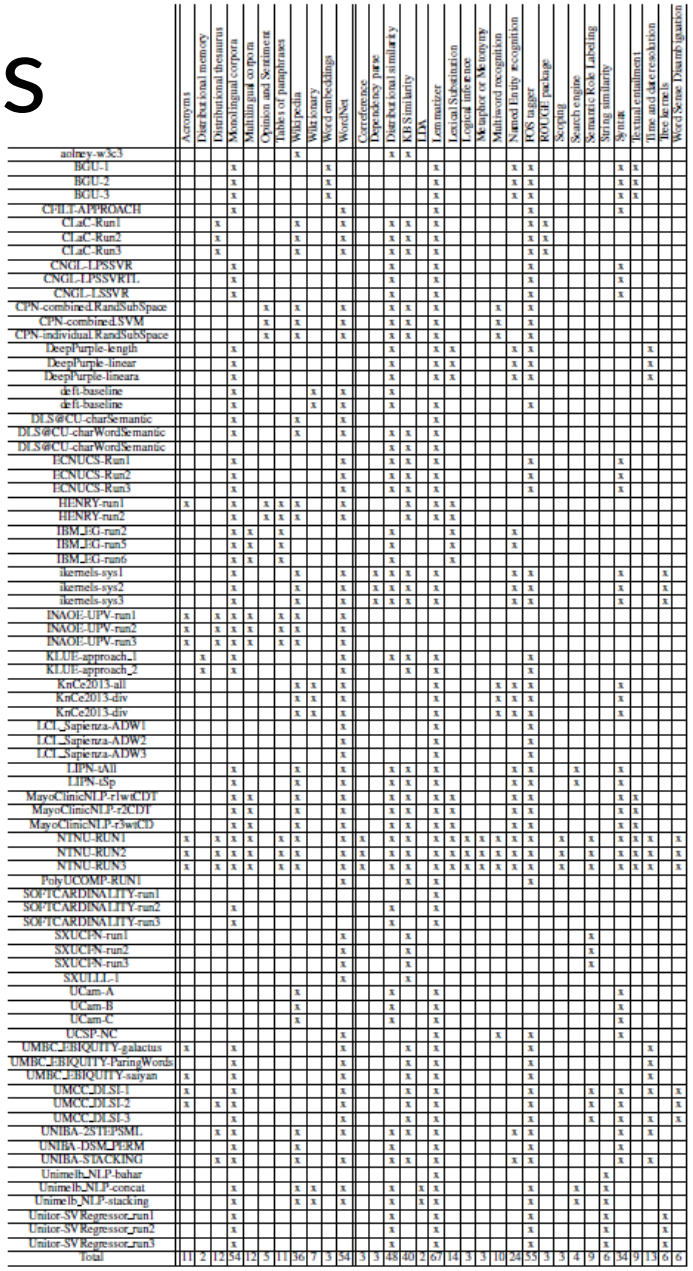
- STS-2012
- Resources and tools used by the systems (from the Task's paper)



	Acronyms	Dictionaries	Distributional thesaurus	Monolingual corpora	Multilingual corpora	Stopwords	Tables of paraphrases	Wikipedia	WordNet	Alignment	Distributional similarity	IDF	Lemmatizer	Lexical Substitution	Machine Learning	MT evaluation	MWE	Named Entity recognition	POS tagger	Semantic Role Labeling	SMT	String similarity	Syntax	Textual entailment	Time and date resolution	Word Sense Disambiguation	Other
aca08ls/task6-University_Of_Sheffield-Hybrid									X			X	X		X							X				X	X
aca08ls/task6-University_Of_Sheffield-Machine_Learning									X			X	X		X							X				X	X
aca08ls/task6-University_Of_Sheffield-Vector_Space									X			X	X													X	X
baer/task6-UKP-run1		X	X	X	X			X	X			X	X	X					X		X	X		X			X
baer/task6-UKP-run2_plus_postprocessing_smt_twsi		X	X	X	X			X	X			X	X	X					X		X	X		X			X
baer/task6-UKP-run3_plus_random		X	X	X	X			X	X			X	X	X					X		X	X		X			X
croce/task6-UNITOR-1_REGRESSION_BEST_FEATURES				X								X		X		X			X				X				
croce/task6-UNITOR-2_REGRESSION_ALL_FEATURES				X								X		X		X			X				X				
croce/task6-UNITOR-3_REGRESSION_ALL_FEATURES_ALL_DOMAINS				X								X		X		X			X				X				
csjxu/task6-PolyUCOMP-RUN							X	X					X						X								
danielcer/stanford_lsa							X		X					X	X		X		X				X				
danielcer/stanford_pdaAll							X		X					X	X		X		X				X				
danielcer/stanford_rte									X	X				X		X	X		X				X				X
davide_buscaldi/task6-IRIT-pg1				X					X				X						X			X					
davide_buscaldi/task6-IRIT-pg3				X					X				X						X			X					
davide_buscaldi/task6-IRIT-wu				X					X				X						X			X					
demetrios_glinos/task6-ATA-BASE									X		X	X							X	X		X					X
demetrios_glinos/task6-ATA-CHNK									X		X	X							X	X		X					X
demetrios_glinos/task6-ATA-STAT									X		X	X							X	X		X					X
desouza/task6-FBK-run1	X			X			X	X	X		X	X	X	X				X	X				X				X
desouza/task6-FBK-run2				X			X	X	X		X	X	X						X								
desouza/task6-FBK-run3				X			X	X	X		X								X								
dvilarinoayala/task6-BUAP-RUN-1		X											X														
dvilarinoayala/task6-BUAP-RUN-2									X																		
dvilarinoayala/task6-BUAP-RUN-3													X														X
jan_snajder/task6-takelab-simple		X	X	X		X		X	X		X	X	X	X	X				X								X
jan_snajder/task6-takelab-syntax				X					X		X	X	X	X				X	X				X				
janardhan/task6-janardhan-UNL_matching									X				X					X	X				X			X	
jotacastillo/task6-SAGAN-RUN1	X			X					X				X	X										X	X	X	
jotacastillo/task6-SAGAN-RUN2	X			X					X				X	X										X	X	X	
jotacastillo/task6-SAGAN-RUN3	X			X					X				X	X										X	X	X	
Konstantin_Z/task6-ABBYY-General																											
M_Rios/task6-UOW-LEX_PARA			X								X		X		X		X	X	X			X					
M_Rios/task6-UOW-LEX_PARA_SEM			X								X		X		X		X	X	X			X					
M_Rios/task6-UOW-SEM			X								X		X				X	X	X			X					
mheilman/task6-ETS-PERP				X					X			X	X	X							X						
mheilman/task6-ETS-PERPphrases				X	X				X			X	X	X		X					X						X
mheilman/task6-ETS-TERp				X	X				X			X	X	X													X
parthapakray/task6-JU_CSE_NLP-Semantic_Syntactic_Approach	X					X			X				X				X	X	X			X	X				X
rada/task6-UNT-CombinedRegression								X	X	X	X	X	X		X											X	X
rada/task6-UNT-IndividualDecTree								X	X	X	X	X	X		X											X	X
rada/task6-UNT-IndividualRegression								X	X	X	X	X	X		X											X	X
sgjimenezv/task6-SOFT-CARDINALITY						X							X	X													
sgjimenezv/task6-SOFT-CARDINALITY-ONE-FUNCTION						X							X	X													
skamler_/task6-EHU-RUN1v2							X			X				X					X		X						
sokolov/task6-LIMSI-cosprod				X							X		X		X												
sokolov/task6-LIMSI-gradtree				X							X		X		X												
sokolov/task6-LIMSI-sumdiff				X							X		X		X												
spirin2/task6-UIUC-MLNLP-Blend	X			X			X	X									X	X	X	X		X	X				X
spirin2/task6-UIUC-MLNLP-CCM	X			X			X	X									X	X	X	X		X	X				X
spirin2/task6-UIUC-MLNLP-Puzzle	X			X			X	X									X	X	X	X		X	X				X
sranjans/task6-sranjans-1				X				X	X		X	X							X							X	
sranjans/task6-sranjans-2				X				X	X		X	X	X	X				X	X						X	X	
sranjans/task6-sranjans-3				X				X	X		X	X	X	X				X	X			X		X			
tiantianzhu7/task6-tiantianzhu7-1									X			X							X							X	
tiantianzhu7/task6-tiantianzhu7-2									X			X							X								
tiantianzhu7/task6-tiantianzhu7-3						X			X			X							X								
weiwei/task6-weiwei-run1		X		X					X			X		X					X								
yeh/task6-SRIUBC-SYSTEM1			X					X	X			X	X	X					X								
yeh/task6-SRIUBC-SYSTEM2			X					X	X			X	X	X					X								
yeh/task6-SRIUBC-SYSTEM3			X					X	X			X	X	X					X								
ygutierrez/task6-UMCC_DLSI-MultiLex				X					X	X			X	X					X								X
ygutierrez/task6-UMCC_DLSI-MultiSem				X					X				X	X					X							X	X
ygutierrez/task6-UMCC_DLSI-MultiSemLex				X					X	X			X	X					X							X	X
yrkakde/task6-yrkakde-DiceWordnet									X		X		X														
Total	8	6	10	33	5	5	9	20	47	7	31	37	49	13	13	4	7	12	43	9	4	13	17	10	5	15	25

Most STS systems are multi-feature regressors

- STS-2013
- Resources and tools used by the systems (from the Task's paper)



	Acronyms	Distributional memory	Distributional thesaurus	Monolingual corpora	Multilingual corpora	Opinion and Sentiment	Tables of paraphrases	Wikipedia	Wiktionary	Word embeddings	WordNet	Coreference	Dependency parse	Distributional similarity	IDS Similarity	LDA	Lemmatizer	Lexical Substitution	Logical inference	Metaphor or Metonymy	Multiword recognition	Named Entity recognition	POS tagger	ROUGE package	Scoping	Search engine	Semantic Role Labeling	String similarity	Syntax	Textual entailment	Time and date resolution	Tree kernels	Word Sense Disambiguation
aolney-w3c3									x					x	x																	
BGU-1				x						x					x							x	x					x	x			
BGU-2				x						x					x							x	x					x	x			
BGU-3				x						x					x							x	x					x	x			
CFILT-APPROACH				x					x						x								x					x				
CLaC-Run1		x						x			x			x	x								x	x								
CLaC-Run2		x						x			x			x	x								x	x								
CLaC-Run3		x						x			x			x	x								x	x								
CNGL-LPSSVR				x											x								x					x				
CNGL-LPSSVRTL				x											x								x					x				
CNGL-LSSVR				x											x								x					x				
CPN-combinedLRandSubSpace						x		x			x				x						x											
CPN-combinedLSVM						x		x			x				x						x		x									
CPN-individualRandSubSpace						x		x			x				x						x		x									
DeepPurple-length				x											x							x	x						x			
DeepPurple-linear				x											x							x	x						x			
DeepPurple-linearа				x											x							x	x						x			
deft-baseline				x					x		x				x																	
deft-baseline				x					x		x				x								x									
DLS@CU-charSemantic				x							x																					
DLS@CU-charWordSemantic				x					x		x				x																	
DLS@CU-charWordSemantic															x																	
ECNUCS-Run1				x							x				x								x					x				
ECNUCS-Run2				x							x				x								x					x				
ECNUCS-Run3				x							x				x								x					x				
HENRY-run1	x			x		x	x	x			x					x	x															
HENRY-run2				x		x	x	x			x					x	x															
IBM_EG-run2				x	x			x							x						x											
IBM_EG-run5				x	x			x							x						x											
IBM_EG-run6				x	x			x																								
ikernels-sys1				x					x		x		x	x	x							x	x					x			x	
ikernels-sys2				x					x		x		x	x	x							x	x					x			x	
ikernels-sys3				x					x		x		x	x	x							x	x					x			x	
INAOE-UPV-run1	x		x	x	x		x	x			x																					
INAOE-UPV-run2	x		x	x	x		x	x			x																					
INAOE-UPV-run3	x		x	x	x		x	x			x																					
KLUE-approach_1		x		x							x			x	x								x									
KLUE-approach_2		x		x							x				x								x									
KnCe2013-all								x	x		x										x	x	x					x				
KnCe2013-div								x	x		x										x	x	x					x				
KnCe2013-div								x	x		x										x	x	x					x				
LCT_Sapienza-ADW1											x												x									
LCT_Sapienza-ADW2											x												x									
LCT_Sapienza-ADW3											x												x									
LIPN-tAll				x					x		x			x	x							x	x			x		x				
LIPN-tSp				x					x		x				x							x	x			x		x				
MayoClinicNLP-r1wtCDT				x	x				x		x			x	x							x	x					x	x			
MayoClinicNLP-r2CDT				x	x				x		x			x	x							x	x					x	x			
MayoClinicNLP-r3wtCD				x	x				x		x			x	x							x	x					x	x			
NTNU-RUN1	x		x	x	x			x	x		x	x	x	x	x	x	x	x	x	x	x	x	x			x		x	x	x		x
NTNU-RUN2	x		x	x	x			x	x		x	x	x	x	x	x	x	x	x	x	x	x	x			x		x	x	x		x
NTNU-RUN3	x		x	x	x			x	x		x	x	x	x	x	x	x	x	x	x	x	x	x			x		x	x	x		x
PolyUCOMP-RUN1											x				x								x									
SOFTCARDINALITY-run1															x																	
SOFTCARDINALITY-run2				x										x																		
SOFTCARDINALITY-run3				x										x																		
SXUCFN-run1											x				x											x						
SXUCFN-run2											x				x											x						
SXUCFN-run3											x				x											x						
SXULLL-1											x				x																	
UCam-A								x						x		x												x				
UCam-B								x						x		x												x				
UCam-C								x						x		x												x				
UCSP-NC											x					x												x				
UMBC_EBIQUITY-galactus	x			x							x				x								x						x			
UMBC_EBIQUITY-ParingWords				x							x				x								x						x			
UMBC_EBIQUITY-saiyan	x			x							x				x								x						x			
UMCC_DLSI-1	x			x							x				x								x			x	x	x	x			x
UMCC_DLSI-2	x		x	x							x				x								x			x	x		x			x
UMCC_DLSI-3				x							x				x								x			x	x	x	x			x
UNIBA-2STEPSMT			x	x					x		x			x	x						x	x					x	x				
UNIBA-DSM_PERM				x					x						x								x					x				
UNIBA-STACKING			x	x					x		x			x	x						x	x					x	x				
Unimelb_NLP-bahar															x											x						
Unimelb_NLP-concat				x				x	x		x			x	x							x			x		x					
Unimelb_NLP-stacking				x				x	x		x			x	x							x			x		x					
Unitor-SVRegressor_run1				x										x								x					x			x		
Unitor-SVRegressor_run2				x										x								x					x			x		
Unitor-SVRegressor_run3				x										x								x					x			x		
Total	11	2	12	54	12	5	11	36	7	3	54	3	3	48	40	2	67	14	3	3	10	24	55	3	3	4	9	6	34	9	13	6	6

Most STS systems are multi-feature regressors

- STS-2013 Resources and tools used by the systems
  - WordNet
  - Monolingual corpora
  - Wikipedia
  - Dictionaries
  - Multilingual corpora
  - Opinion and sentiment analysis
  - Lists and tables of paraphrases

Sentence Similarity Techniques

Soft cardinality

Jimenez et al (2010)

Uses only surface text information, a stop-word remover, and a stemmer

ranked 3rd in STS-12

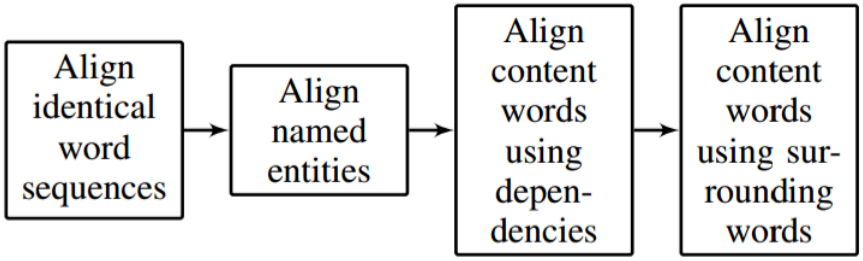
$$SIM(A,B) = \frac{|A \cap B| + bias}{\alpha \max(|A|,|B|) + (1-\alpha)\min(|A|,|B|)}$$

Sentence Similarity Techniques

Monolingual alignment

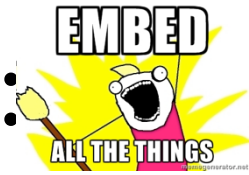
Sultan et al (2014): best system in STS-14 and -15

DLS@CU



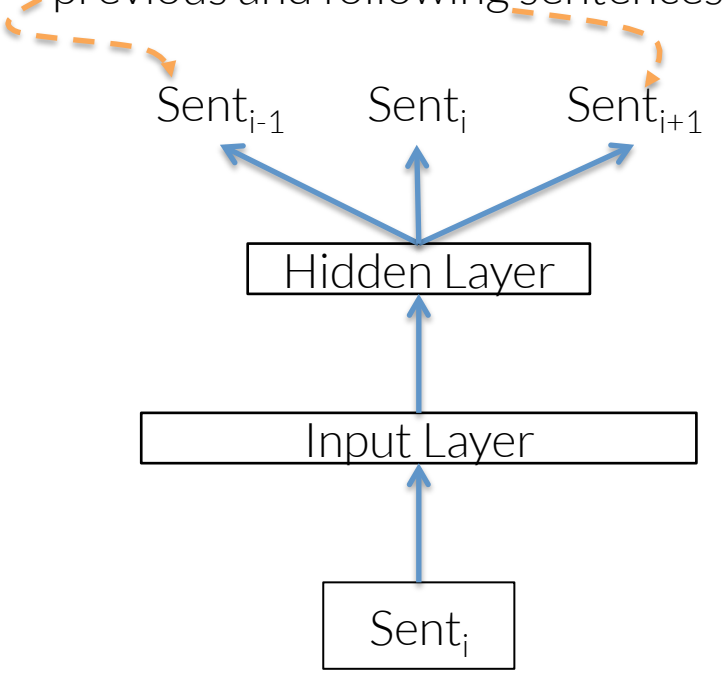
The similarity score is computed as a function of the proportions of aligned content words in the two input sentences.

Sentence Similarity Techniques: Skip-thought vectors



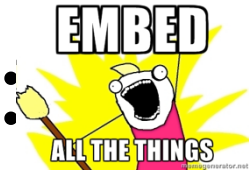
Embedding a sentence with unsupervised training

Given a sentence, predict the previous and following sentences



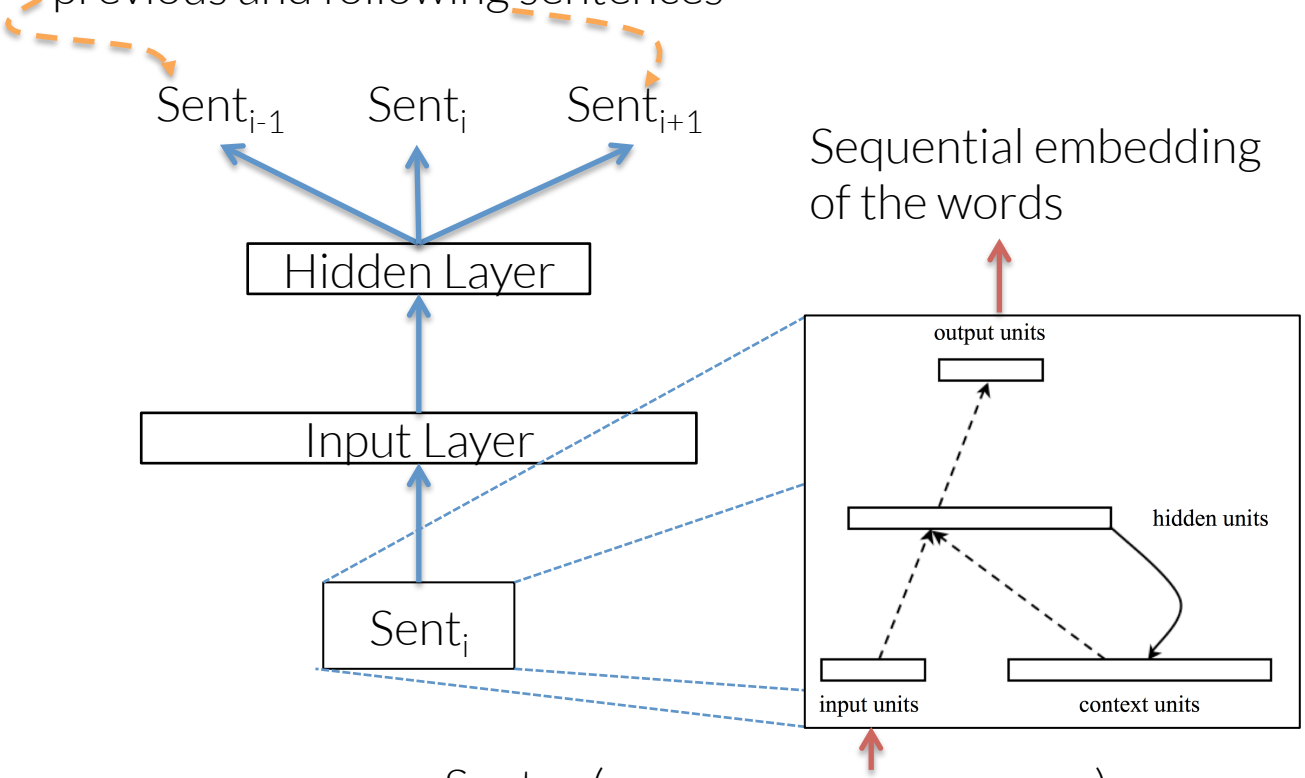
Kiros et al. (2015)

Sentence Similarity Techniques: Skip-thought vectors



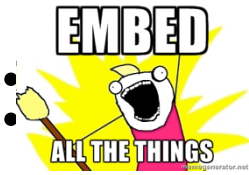
Embedding a sentence with unsupervised training

Given a sentence, predict the previous and following sentences

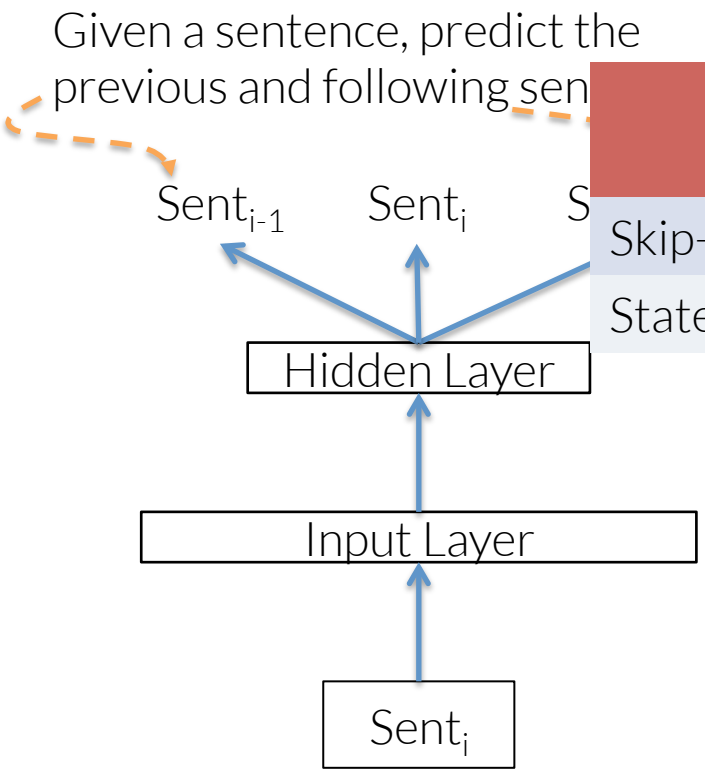


Kiros et al. (2015)

Sentence Similarity Techniques: Skip-thought vectors



Embedding a sentence with unsupervised training



	MSR Paraphrase Detection (MSE)	SICK Semantic Relatedness (F1)
Skip-Thought Vecs	0.2561	83.0
State of the Art	0.2532	84.1

Not state of the art, but high performance on a wide variety of tasks

Kiros et al. (2015)

Coffee Break

30 minutes

Paragraph Similarity

Paragraphs represent large thematic, topical units -- more than just a sequence of sentence

The Lisbon region is the wealthiest region in Portugal and it is well above the European Union's GDP per capita average – it produces 45% of the Portuguese GDP. Lisbon's economy is based primarily on the tertiary sector. Most of the headquarters of multinationals operating in Portugal are concentrated in the Grande Lisboa Subregion, specially in the Oeiras municipality. The Lisbon Metropolitan Area is heavily industrialized, especially the south bank of the Tagus river (Rio Tejo).

Little evaluation directly on paragraph similarity

- Often used as the unit of text for applications
  - Plagiarism detection
  - Summarization
  - Essay grading
  - Scientific abstracts
  - Document chunking
  - Document alignment

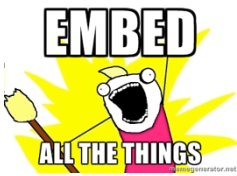
Simplest Idea: Model paragraphs as a bag of words (BoW)

Paragraph BoW representations run into all the same issues as with words

- huge dimensionality makes them cumbersome
- ignores word semantics

Paragraphs also include word ordering and sentence ordering

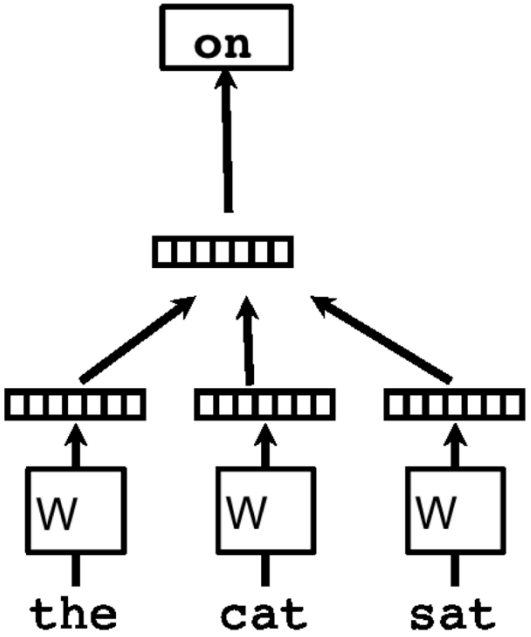
- The topic sentence can matter!



Current state of the art: doc2vec

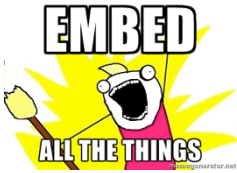
Tackles two problems with bag-of-word and topic modeling approaches:

- They lose the ordering of the words
- They ignore semantics of the words



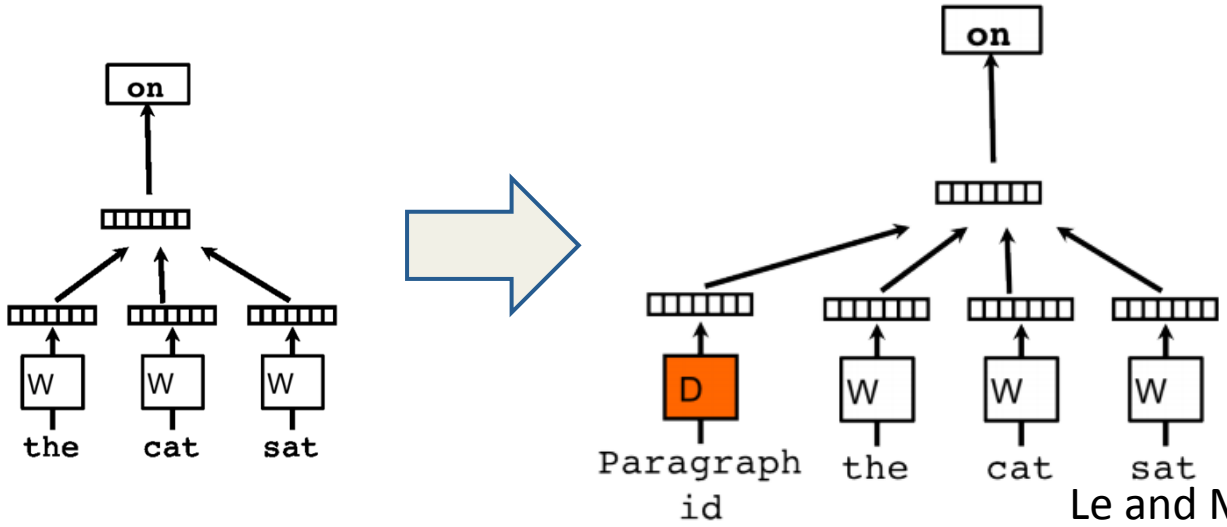
Base model is a prediction task to predict the next word in a sequence

Le and Mikolov (2014)



Current state of the art: doc2vec

- Incorporate paragraph structure explicitly by adding a paragraph vector to the predictive model
  - Every paragraph is mapped to a unique vector
  - A paragraph is thought of as another word that remembers what is missing from the current context



Le and Mikolov (2014)

Document Similarity

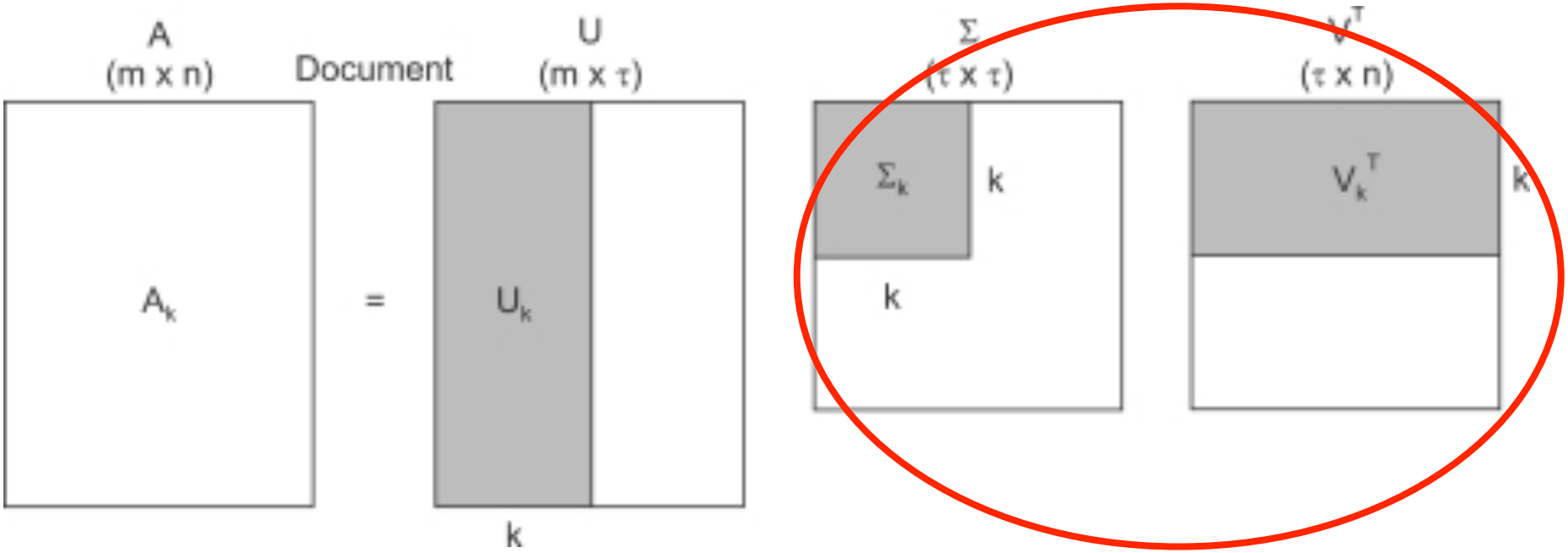
Simplest Approach: Term-Document Matrix

$$
\begin{array}{c|cccc}
 & T_1 & T_2 & \cdots & T_t \\
D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\
D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\
\vdots & \vdots & \vdots & & \vdots \\
\vdots & \vdots & \vdots & & \vdots \\
D_n & w_{1n} & w_{2n} & \cdots & w_{tn}
\end{array}
$$

With sufficiently large documents, document vectors are robust representations

Early document similarity techniques used vector space models

Latent Semantic Indexing (LSI, aka LSA) developed by Deerwester (1988) to address already-discussed issues with VSMs.



Topic Modeling: Viewing document contents as a mixture of topics

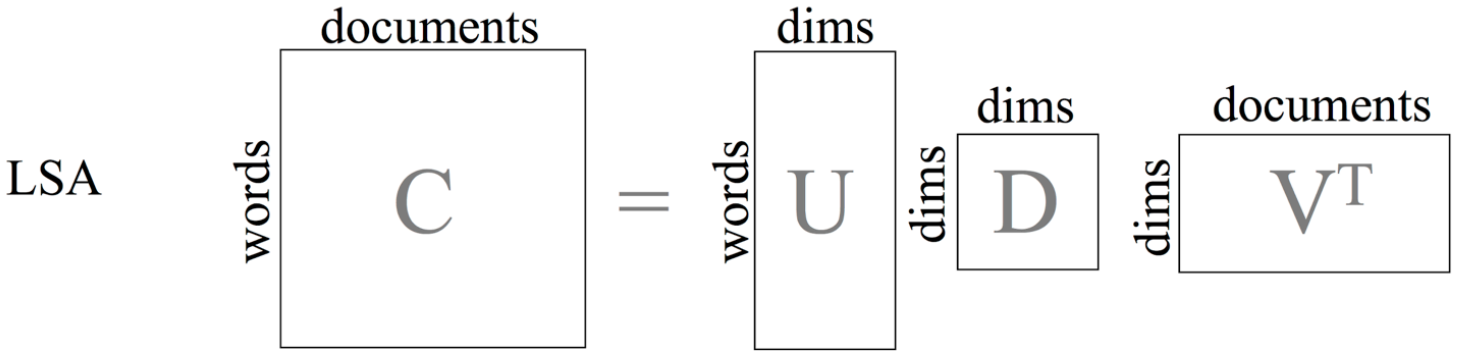


Image credit: Steyvers and Griffiths (2007)

Topic Modeling: Viewing document contents as a mixture of topics

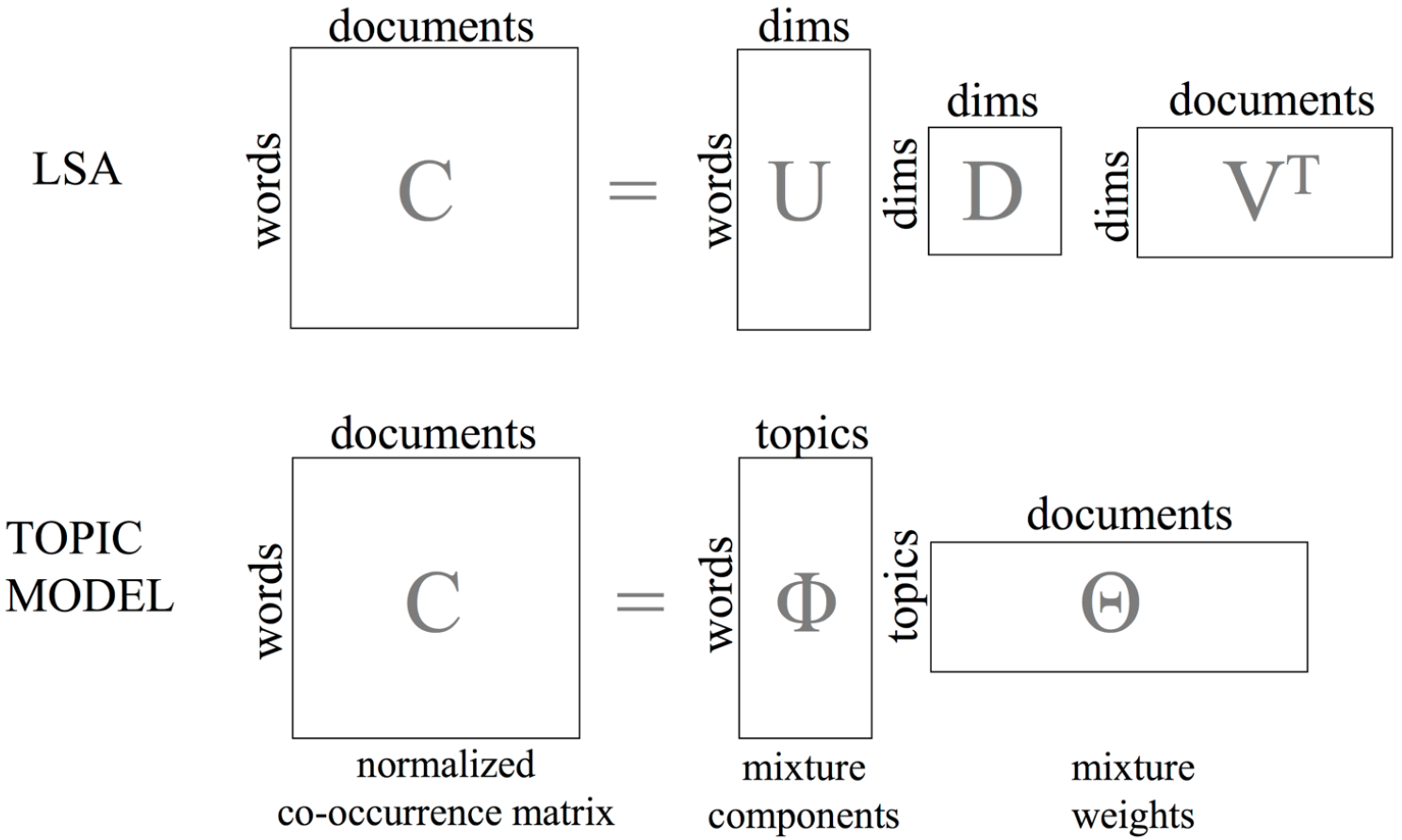


Image credit: Steyvers and Griffiths (2007)

Topic Modeling: Viewing document contents as a mixture of topics

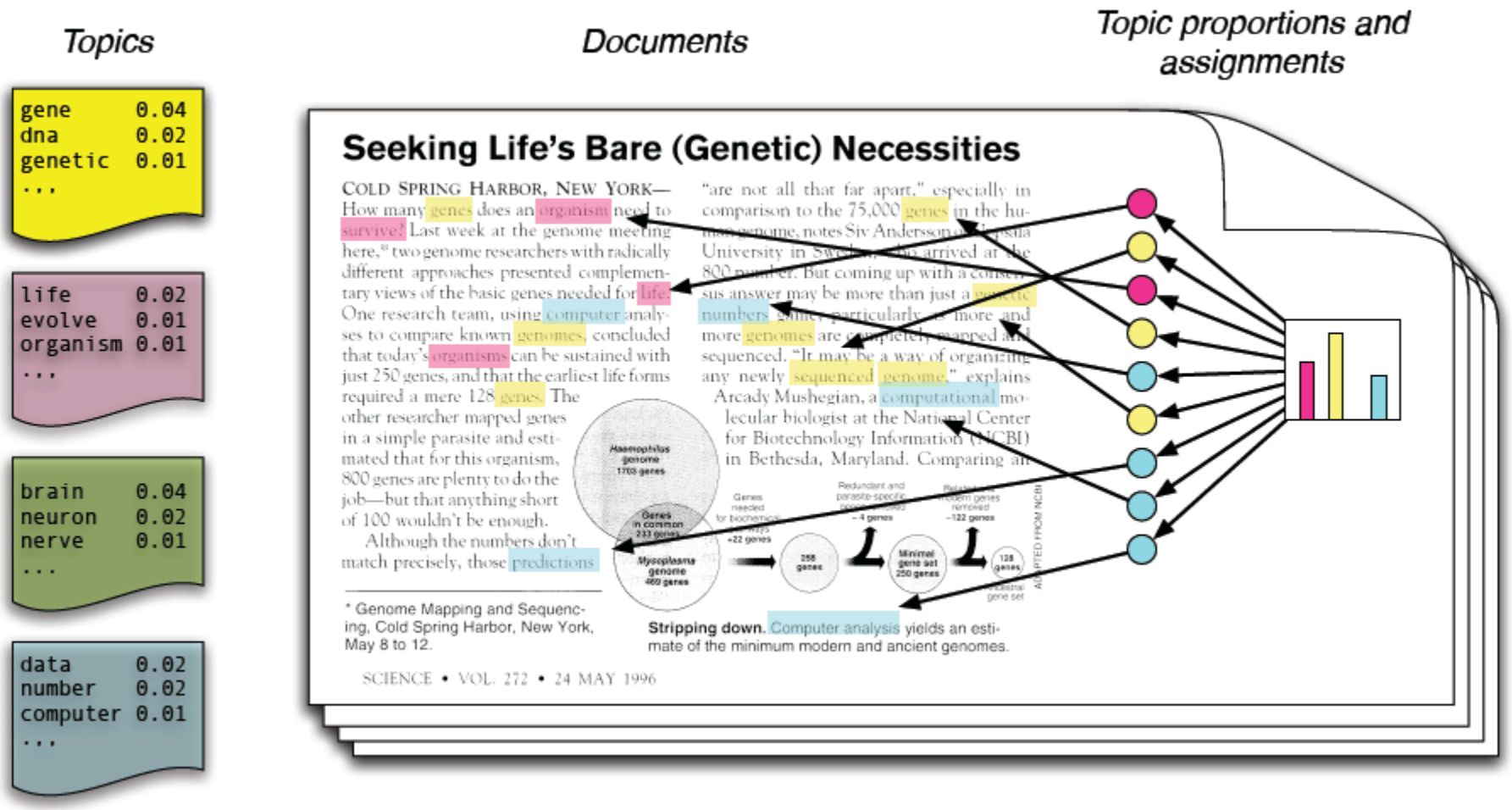


Image credit: Blei (2012)

Document Similarity Techniques

Latent Dirichlet Allocation

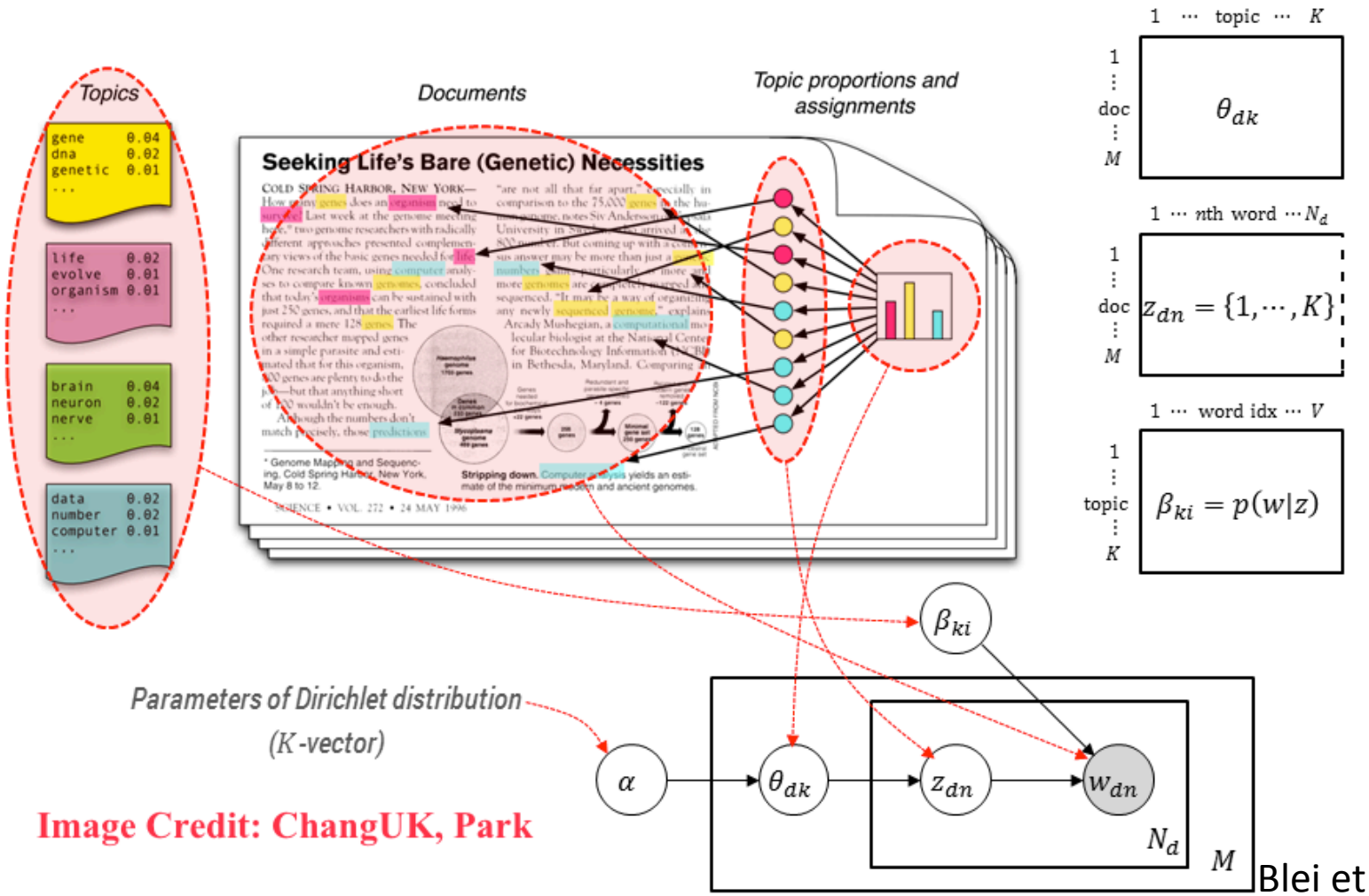


Image Credit: ChangUK, Park

Blei et al (2003)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

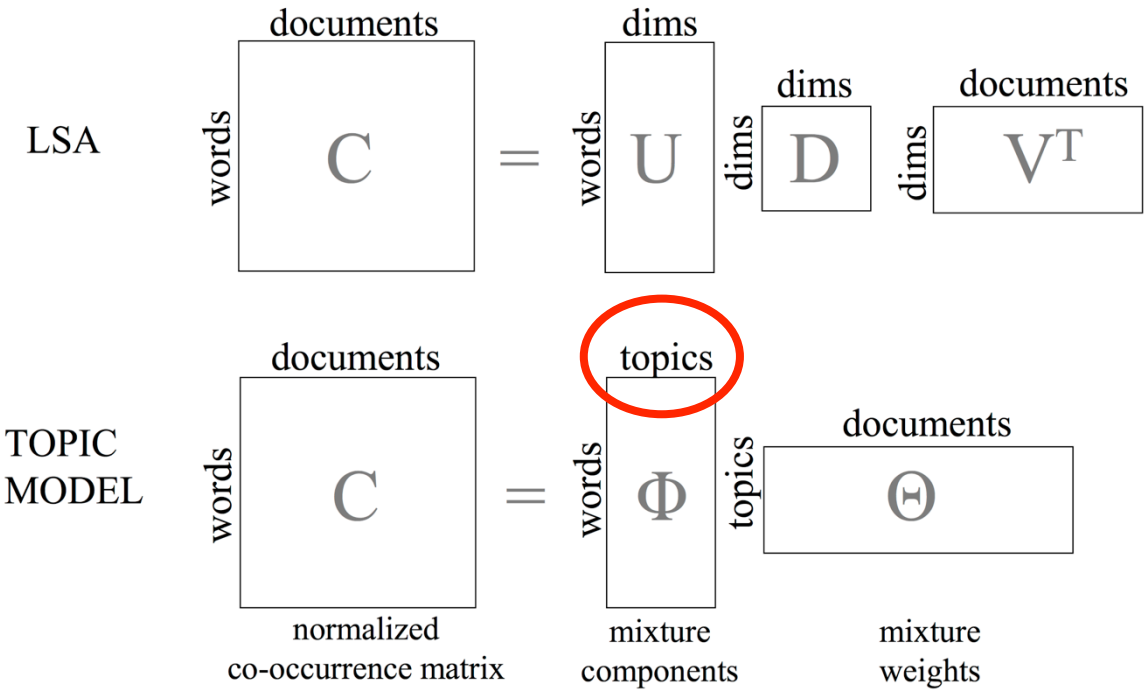
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Blei et al (2003)

Key points for using topic distributions as document representations

- Selecting the number of topics
- Identify relationships between topics
- Moving beyond token-topic assignments

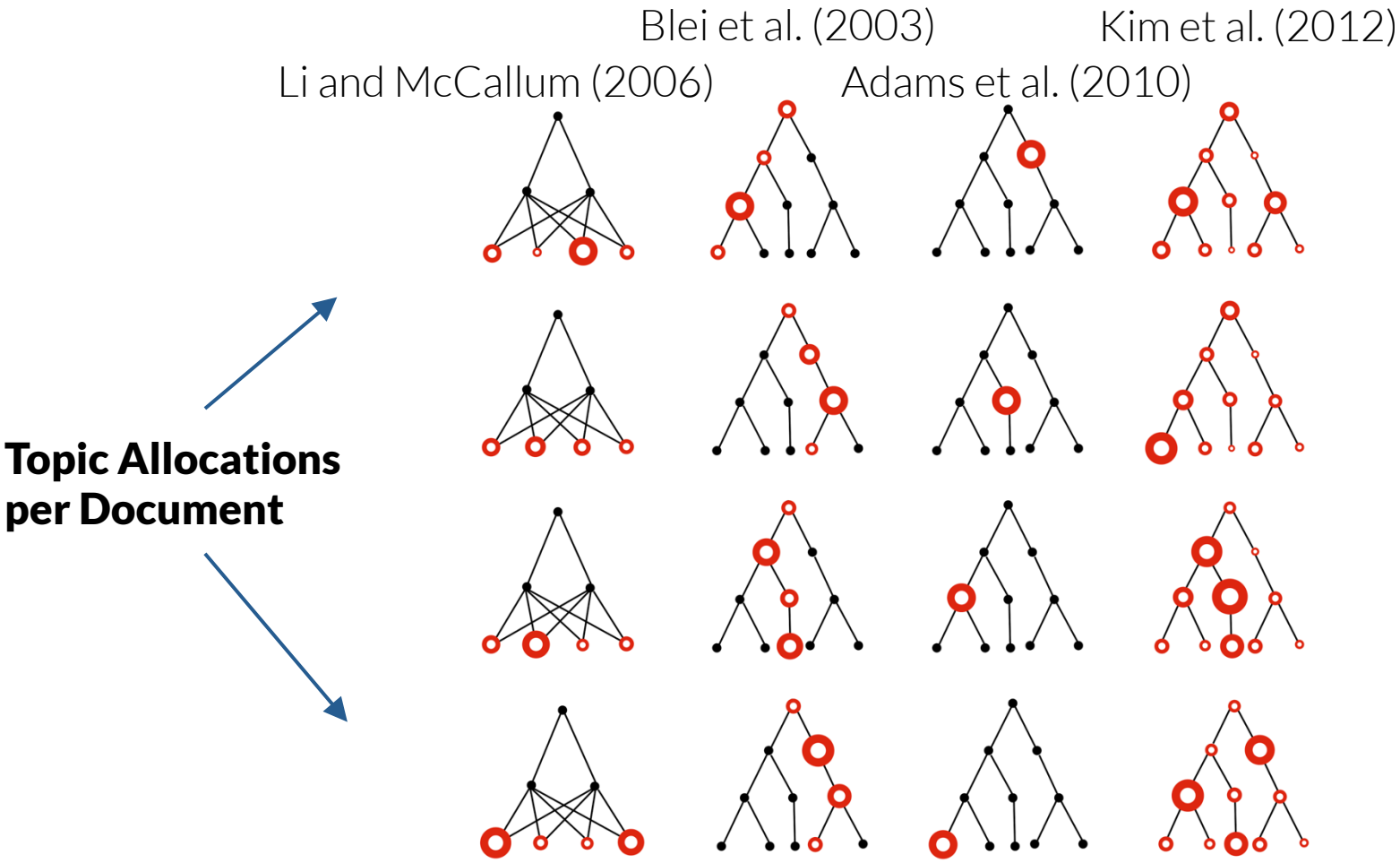
How many topics should you use?



Let a Hierarchical Dirichlet Process (HDP) model decide for you.

(Teh et al., 2005; Fourtassi and Dupoux, 2013)

Introducing structure into the topics



Hierarchical topic organizations can potentially yield more informative document representations

Image credit: Kim et al. (2012)

Incorporating Multi-Word Expressions into topics

- Pre-process the corpus to glob MWEs together, e.g., “white house” -> white_house
  - Not feasible for domain-specific MWEs

Incorporating Multi-Word Expressions into topics

- Pre-process the corpus to glob MWEs together, e.g., “white house” -> white_house
  - Not feasible for domain-specific MWEs
- Learn the MWEs on the fly by looking at topic-assignment sequences
  - TurboTopics (Blei and Lafferty, 2009)

TurboTopics example phrases

Huffington Post

movie	barack obama	marriage	hillary clinton
the film	obamas	state	campaign
hollywood	campaign	in california	bill clinton
director	sen barack obama	gay	shes
first	democratic	decision	the clinton
character	the illinois senator	court	hillarys
documentary	michelle	law	president
theater	recent	supreme court	sen clinton
best	speech	couples	mark penn
sex and the city	choice	ruling	politics
hbo	sen clinton	rights	sexism
scene	david axelrod	equality	the first
to make	president	legal	her campaign
release	camp	to marry	supporters
screen	the huffington post	married	made
actor	endorsed	samesex couples	fight
made	seen	states	called
stars	attacks	gay marriage	mrs clinton
indiana jones	political	sexual orientation	political
seen	gave	the california supreme court	hillary rodham clinton

Physics arXiv

mass	model	lattice qcd	phase transitions
star formation	point	mass	model
stars	monte carlo simulations	dirac operator	symmetry
masses	fixed point	chiral perturbation theory	point
black hole	results	operators	quantum
stellar	lattice	quarks	systems
star	scaling	limit	phase transition
black holes	numerical	theta	phase diagram
massive	ising model	quark	system
msun	two	mev	order
solar masses	we study the	simulations	field
stellar mass	models	lattice spacing	order parameter
black hole mass	quantum monte carlo	chiral symmetry breaking	critical
the stellar	interactions	results	two
young	numerical simulations	effects	transitions in
the mass	simulation	small	models
times	dimensions	baryon	different
myr	analytical	in the continuum limit	symmetry breaking
imf	phase	physical	first order
supermassive black holes	spin glass	quenched	phenomena

n-gram topics

film	obama	california	clinton
movie	barack	marriage	hillary
films	obamas	gay	clintons
movies	sen	court	campaign
hollywood	campaign	state	bill
documentary	senator	couples	shes
director	democratic	supreme	president
jones	illinois	decision	hillarys
screen	president	married	supporters
character	presidential	samesex	penn
cannes	recent	rights	politics
festival	political	marry	sexism
city	speech	law	political
theater	huffington	ruling	rodham
star	politics	states	democratic
hbo	michelle	equality	first
scene	voters	legal	say
actor	supporters	lesbian	sen
played	candidacy	equal	mrs
indiana	choice	appeals	presidency

mass	carlo	lattice	phase
black	monte	qcd	transitions
star	simulations	chiral	phases
stellar	point	theory	transition
stars	model	mass	quantum
masses	results	quark	critical
hole	fixed	finite	symmetry
massive	critical	quenched	field
formation	study	perturbation	point
holes	two	limit	model
msun	lattice	quarks	order
function	dimensions	results	diagram
young	scaling	potential	systems
supermassive	numerical	staggered	two
accretion	simulation	chemical	theory
rate	transition	masses	system
solar	ising	simulations	study
initial	phase	theta	breaking
galactic	twodimensional	continuum	spin
central	temperature	volume	first

unigram topics

Blei and Lafferty (2009)

Incorporating Multi-Word Expressions into topics

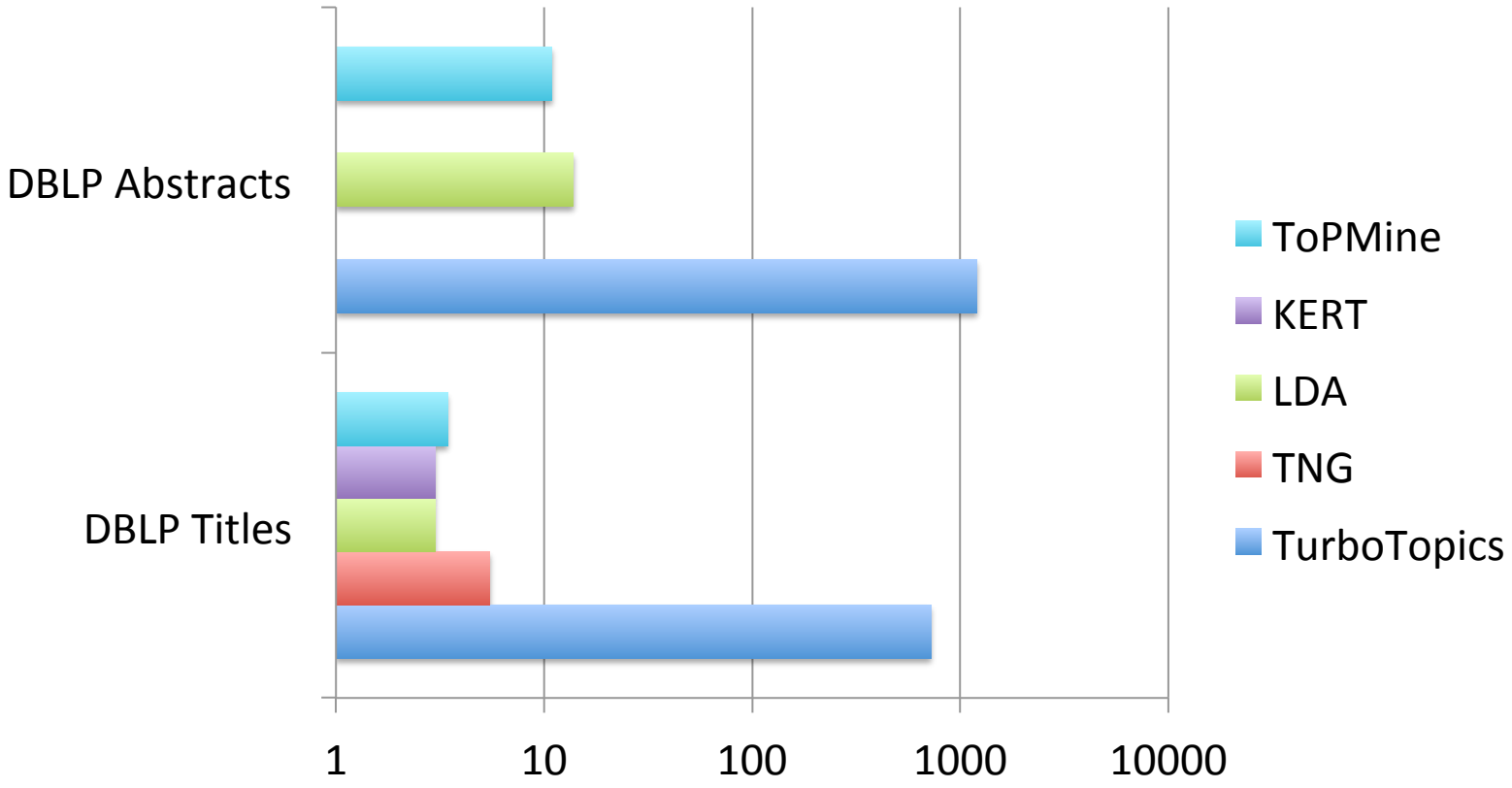
- Pre-process the corpus to glob MWEs together, e.g., “white house” -> white_house
  - Not feasible for domain-specific MWEs
- Learn the MWEs on the fly by looking at topic-assignment sequences
  - TurboTopics (Blei and Lafferty, 2009)
- Learn the MWEs *during* topic modeling
  - Most scalable approach is Top-Min (El-Kishky et al., 2014)

TopMine example phrases

	Topic 1	*Topic 2*	*Topic 3*	*Topic 4*
1-grams	plant	church	palestinian	bush
	nuclear	catholic	israeli	house
	environmental	religious	israel	senate
	energy	bishop	arab	year
	year	pope	plo	bill
	waste	roman	army	president
	department	jewish	reported	congress
	power	rev	west	tax
	state	john	bank	budget
	chemical	christian	state	committee
n-grams	energy department	roman catholic	gaza strip	president bush
	environmental protection agency	pope john paul	west bank	white house
	nuclear weapons	john paul	palestine liberation organization	bush administration
	acid rain	catholic church	united states	house and senate
	nuclear power plant	anti semitism	arab reports	members of congress
	hazardous waste	baptist church	prime minister	defense secretary
	savannah river	united states	yitzhak shamir	capital gains tax
	rocky flats	lutheran church	israel radio	pay raise
	nuclear power	episcopal church	occupied territories	house members
	natural gas	church members	occupied west bank	committee chairman

El-Kishky et al. (2014)

Phrase detection isn't always fast though



Documents can contain much more than just text

The Application of the PSO Based Community Discovery Algorithm in Scientific Paper Management SNS Platform*

Ruixin Ma[1], Guishi Deng[2], Xiao Wang[2], and Ailinna[2]

[1] School of Software, Dalian University of Technology
teacher_mrx@126.com
[2] School of Management, Dalian University of Technology
denggs@dlut.edu.cn

Abstract. The development of SNS provides a new platform and application prospect for the realization of Personalized Recommendation System. It is becoming a fire new research hot spot in social science and e-commence about how to apply community discovery algorithm (CDA) to find community structures in large network and to effectively conduct personalized recommendation. In terms of our recent research, we applied PSO based CDA as principle to divide social communities and based on this to conduct personalized recommendation in a scientific paper management system. The system's running results proved that by applying PSO based CDA, the accuracy of PR and the popularity of the platform had been both improved greatly.

Keywords: SNS, community structure, Personalized Recommendation, PSO based CDA.

1 Introduction

Social network lays emphasis on erpersonal interaction and reflects user's essential need for internet service. In recent years, SNS has become an increasing concern because it allows users to create and to manage things independently by themselves. This concern not only comes from the need to analyze network users' behavior model but also is dominated by the request to maintain the network cohesion. As a result of sparse communication and identity conceals, the effects of PR are not very efficient in normal network.

However, the emergence of SNS brings network users with a safer and more reliable platform[1] which enables users to interact with others more convenient and relieved. This paper's results sufficiently proved that SNS optimized the consequent of social communication and constructed the foundation to realize PRS.

The rest of this paper is organized as follows. In section 2, we introduce the current development of SNS. In section 3, common community discovery algorithms are presented. Section 4 introduces the application of PSO based CDA in scientific paper management system. Subsequently, running comparison is presented and finally, section 6 gives conclusion.

* Supported by"the Fundamental Research Funds for the Central Universitie.

W. Zhang (Ed.): Software Engineering and Knowledge Engineering, AISC 162, pp. 661–667
springerlink.com

Chapter Two

First Steps in My New Country

My father gave me 50 piasters – that was the name of the sub-division of the pound. My mother made me take my awful, rough Bulgarian coat, and wearing my freshly washed but creased white dress, I made my way towards the main gate.

The policeman, in shirt sleeves, was sitting on a chair at the entrance to his little booth, reading a newspaper. He looked up as I came nearer and smiled:

'Bulgaria?'

I nodded. He spoke to me for a while, and although I didn't understand a word of what he said, I realised that he was talking about me. His winks and side-smiles, made that quite clear, but I didn't much like to think what the subject matter of his discourse was. He pointed to himself eventually, and said:

'Polonia.'

I understood that he was telling me that he was Polish.

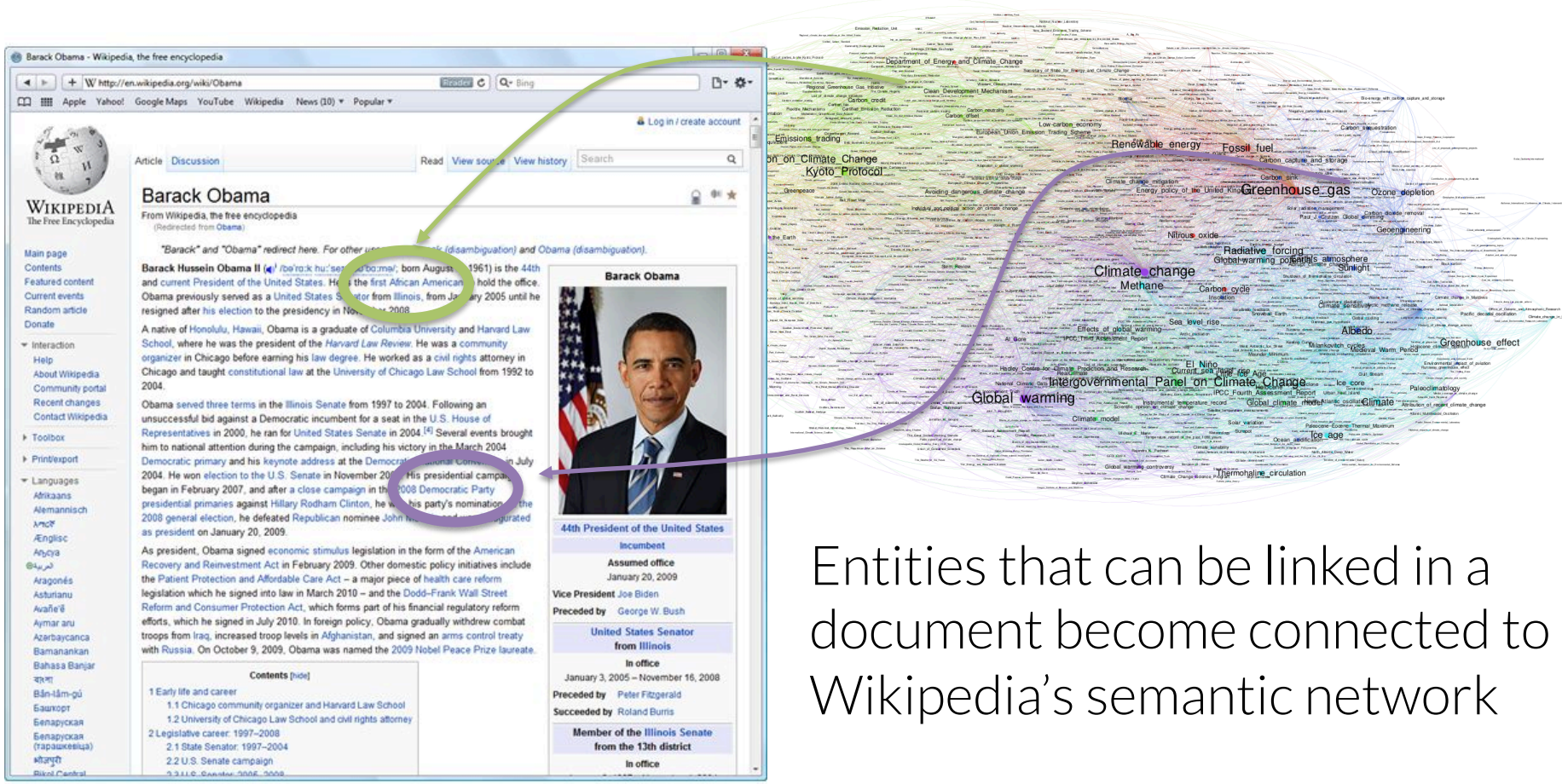
'Yes, Warsaw.' I said.

'No, no Krakow!'

It was all Poland to me, but obviously, to him, it made a

Lots of work on structured document similarity

Adding knowledge to the document representation

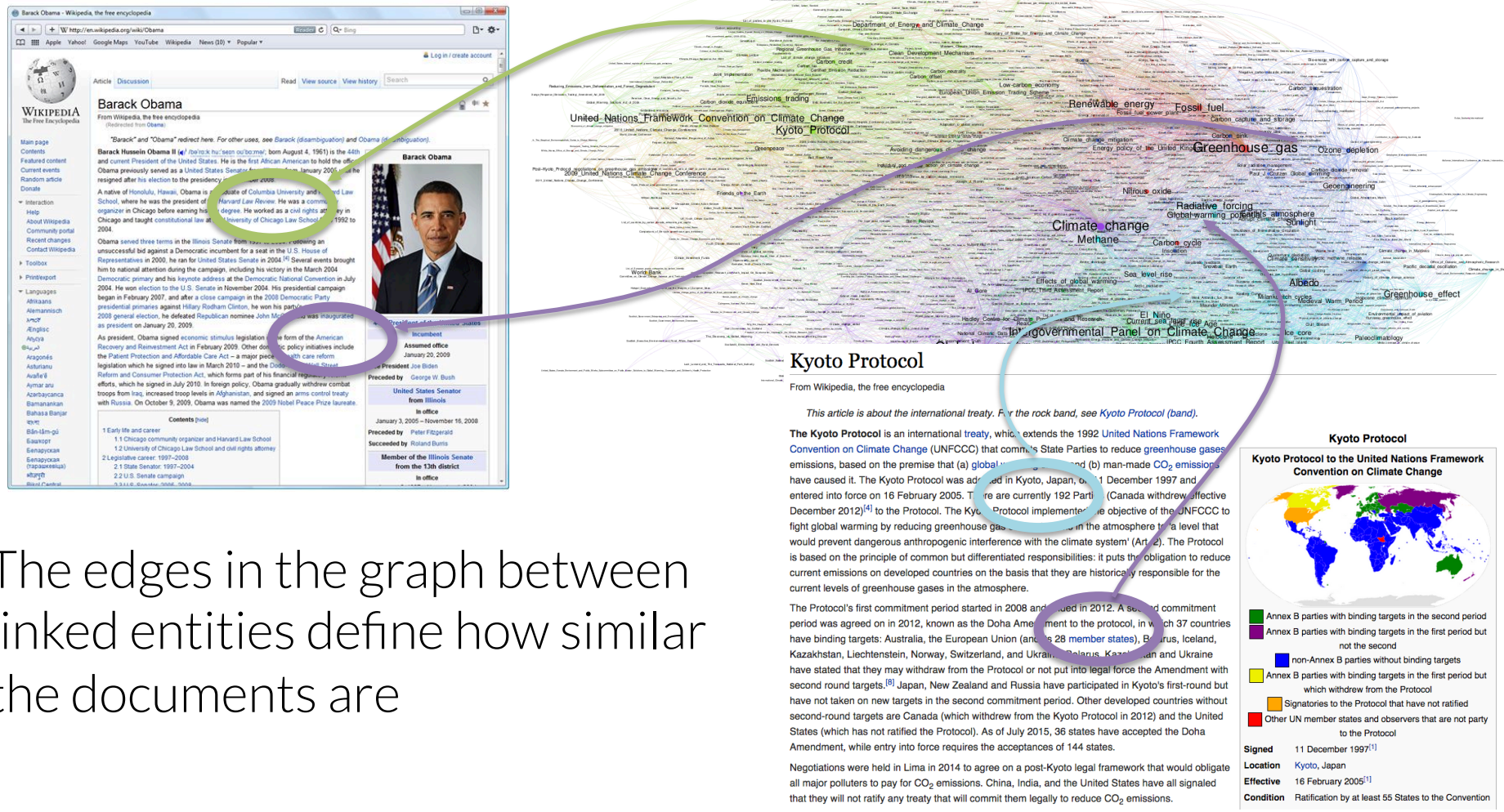


Entities that can be linked in a document become connected to Wikipedia's semantic network

Image credit: http://www.emapsproject.com/blog/archives/1572

Schuhmacher and Ponzetto (2014)

Adding knowledge to the document representation



The edges in the graph between linked entities define how similar the documents are

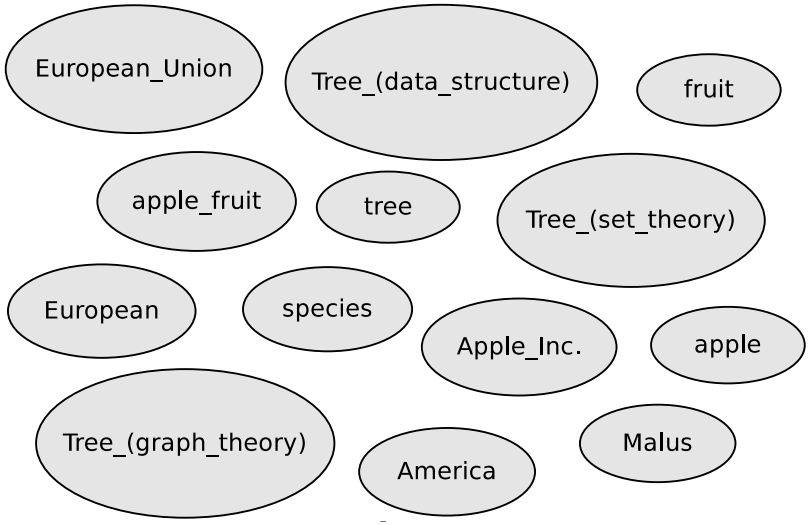
Image credit: http://www.emapsproject.com/blog/archives/1572

Schuhmacher and Ponzetto (2014)

Representing the document within a knowledge base

Start by representing the words in the document as potential concept nodes in the graph

"Species of the European apple tree (Malus) are in America too." → {"European", "apple", "tree", "Malus", "species", "America"}

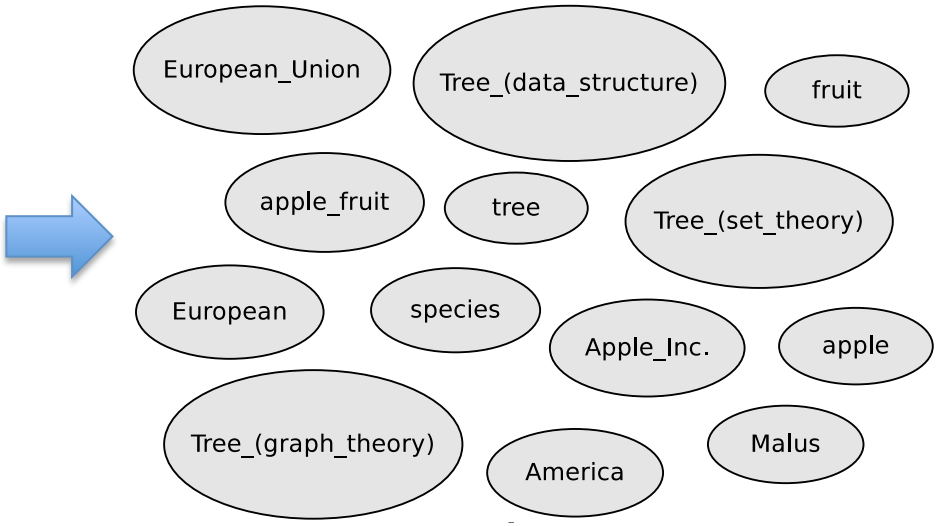


Franco-Salvador et al. (2014)

Representing the document within a knowledge base

Start by representing the words in the document as potential concept nodes in the graph

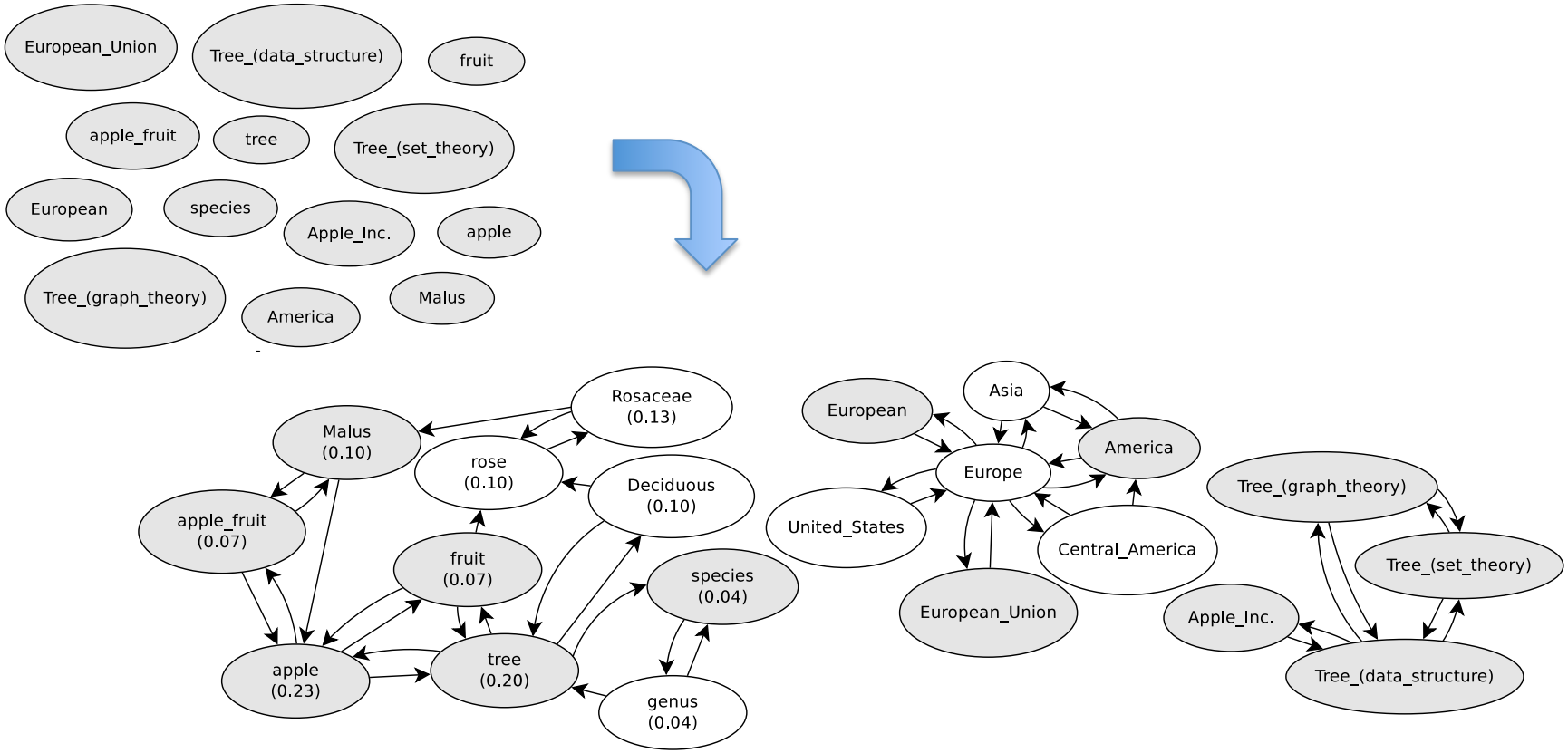
"Species of the European apple tree (Malus) are in America too." → {"European", "apple", "tree", "Malus", "species", "America"}



Franco-Salvador et al. (2014)

Representing the document within a knowledge base

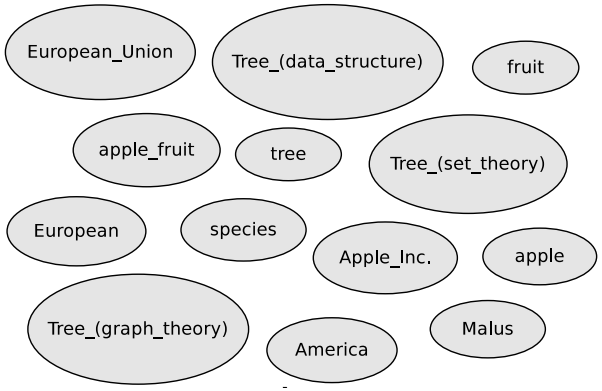
Link the potential concepts in the knowledge base using their semantic relations to find the correct interpretation



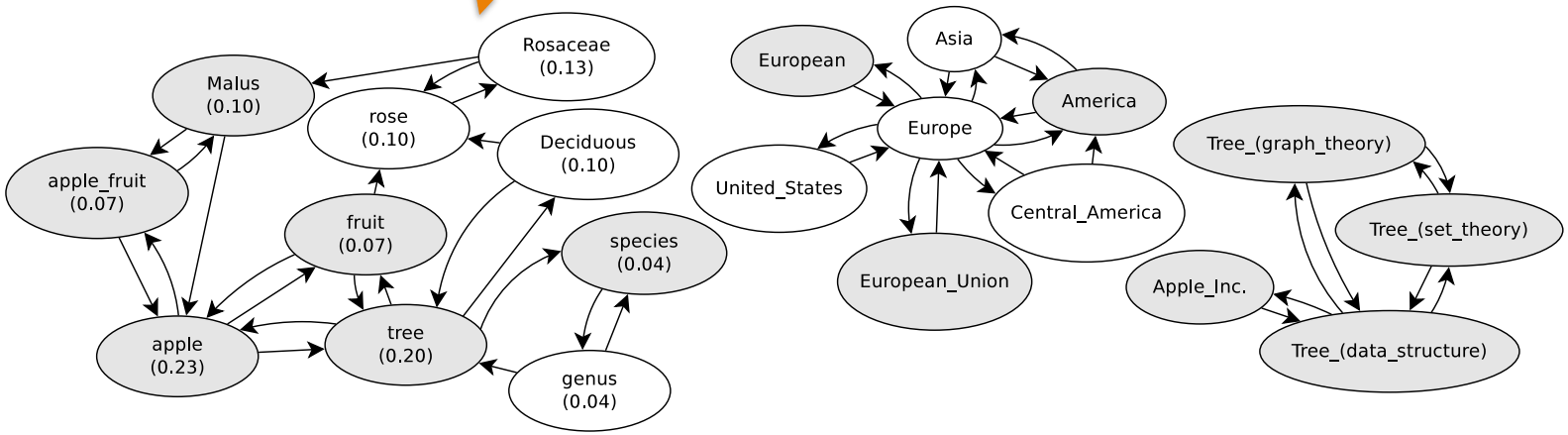
Franco-Salvador et al. (2014)

Representing the document within a knowledge base

Link the potential concepts in the knowledge base using their semantic relations to find the correct interpretation



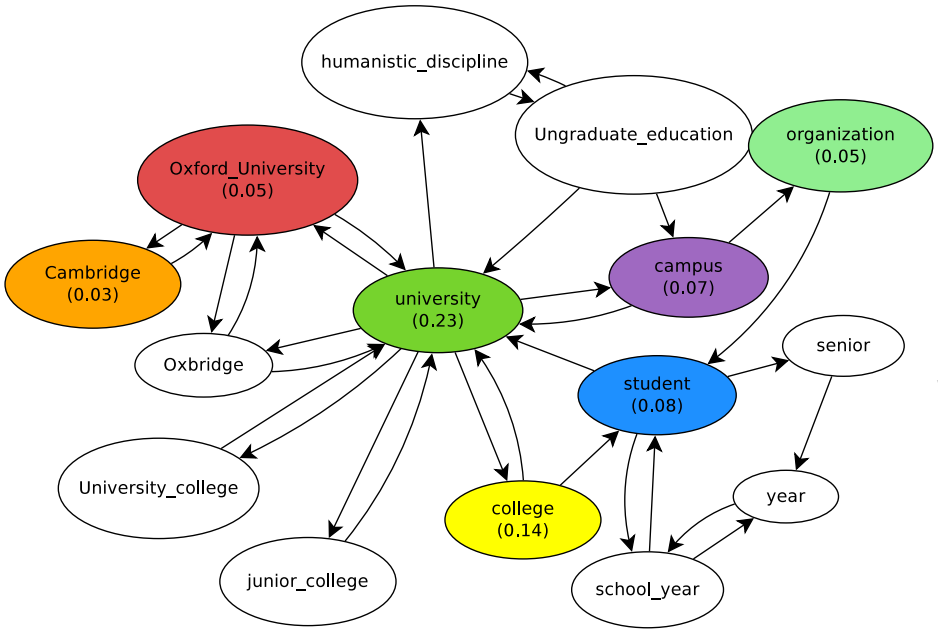
Select the largest component as the correct interpretation



Franco-Salvador et al. (2014)

Representing the document within a knowledge base

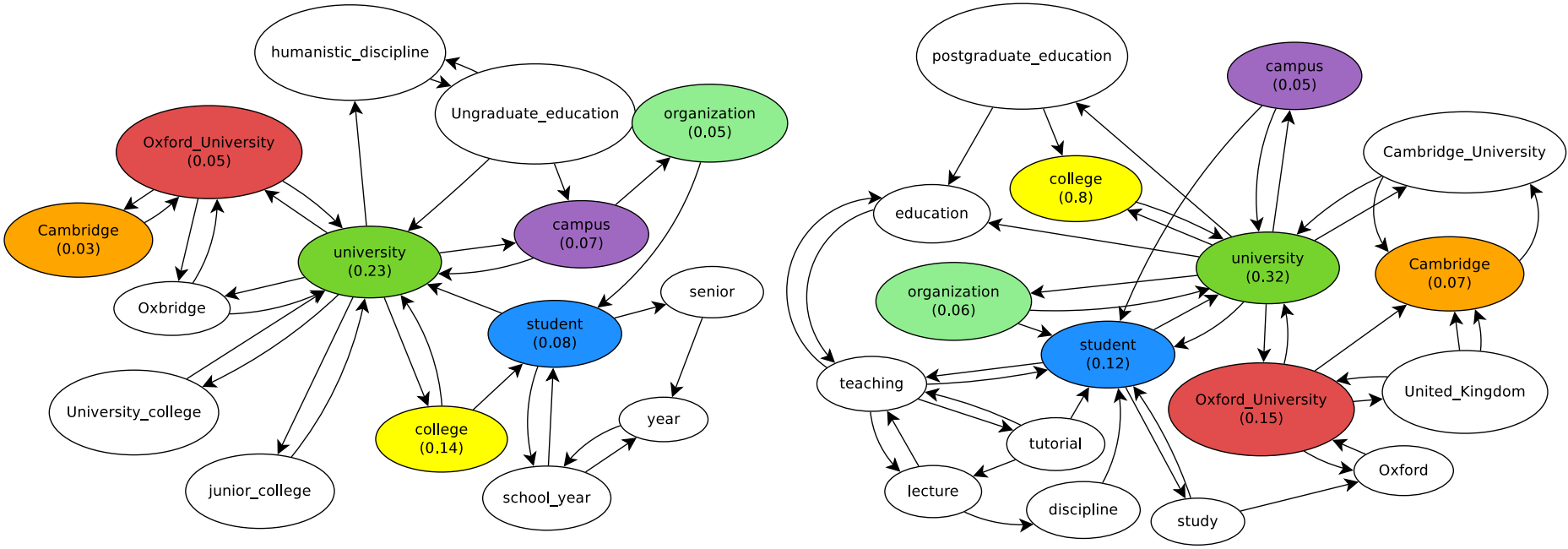
Weight the importance of concepts to a document by running PageRank over its graph



Franco-Salvador et al. (2014)

Representing the document within a knowledge base

Compare graphs by making them vectors over concepts and using cosine similarity

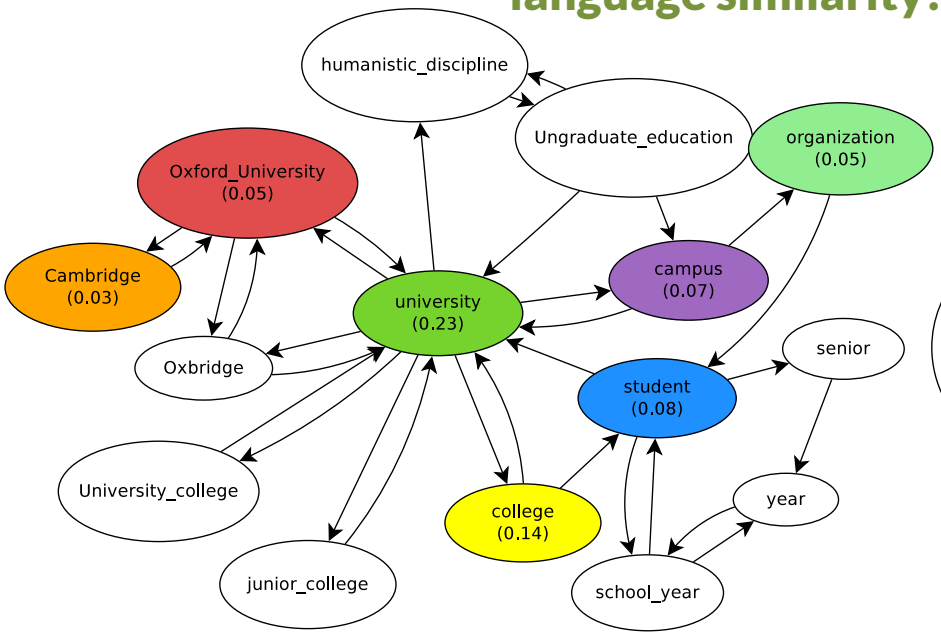


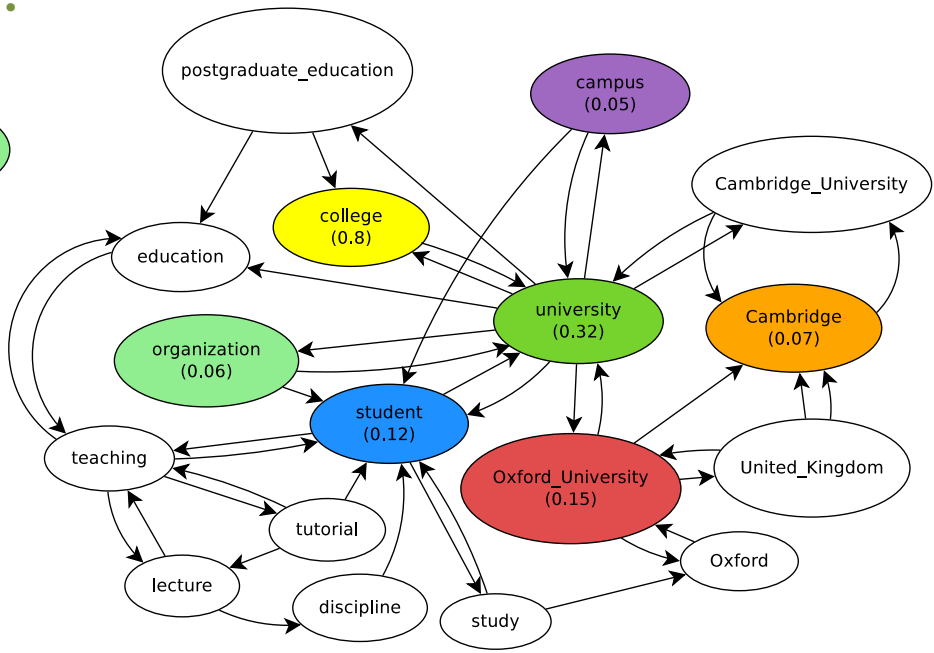
Franco-Salvador et al. (2014)

Representing the document within a knowledge base

Compare graphs by making them vectors over concepts and using cosine similarity

Since the underlying conceptual knowledge base is multilingual, the representation is capable of cross-language similarity!





Franco-Salvador et al. (2014)

Other recent works have tried an LSA-like approach with new dimensionality reductions

- Non-negative Matrix Factorization
  (Xu et al., 2003)
- Concept Factorization
  (Xu and Gong, 2004)
- Locally-Consistent Concept Factorization
  (Cai et al. 2011)
  - Non-linear dimensionality reduction

Main issues are computational complexity and representational opaqueness

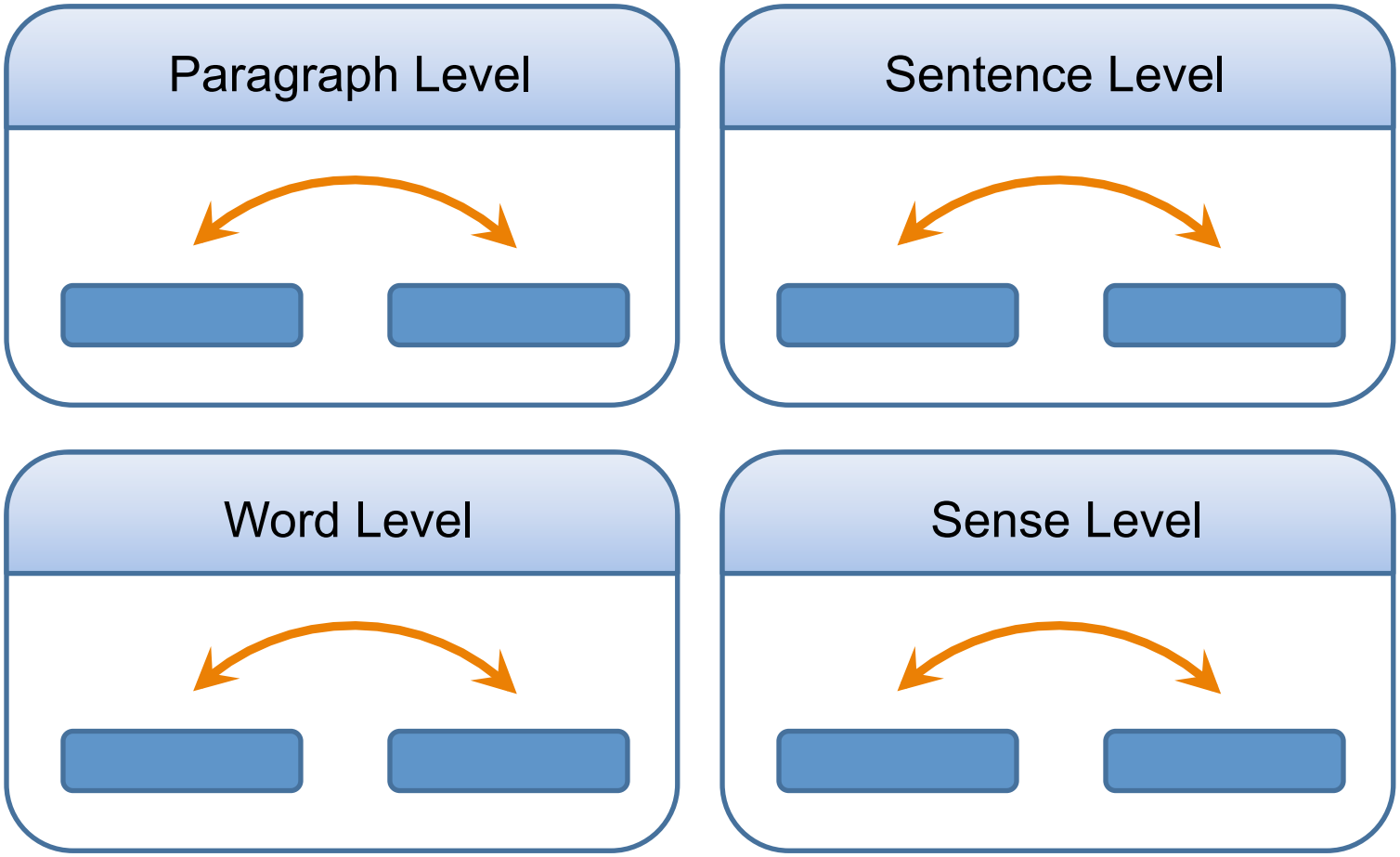
Document Similarity Summary

- Topic Models are still state of the art for bag of words documents
  - If you don't know the number of topics…
    - Use a hierarchical dirichlet process (HDP)
  - If your data contains many multi-word expressions…
    - Use a model that finds these for you, like TurboTopics or TopMiner
- If all of your documents have regular structure
  - Consider using structured document models
- If your document contain many entities
  - Consider linking to a Knowledge Base

Cross-Level Semantic Similarity

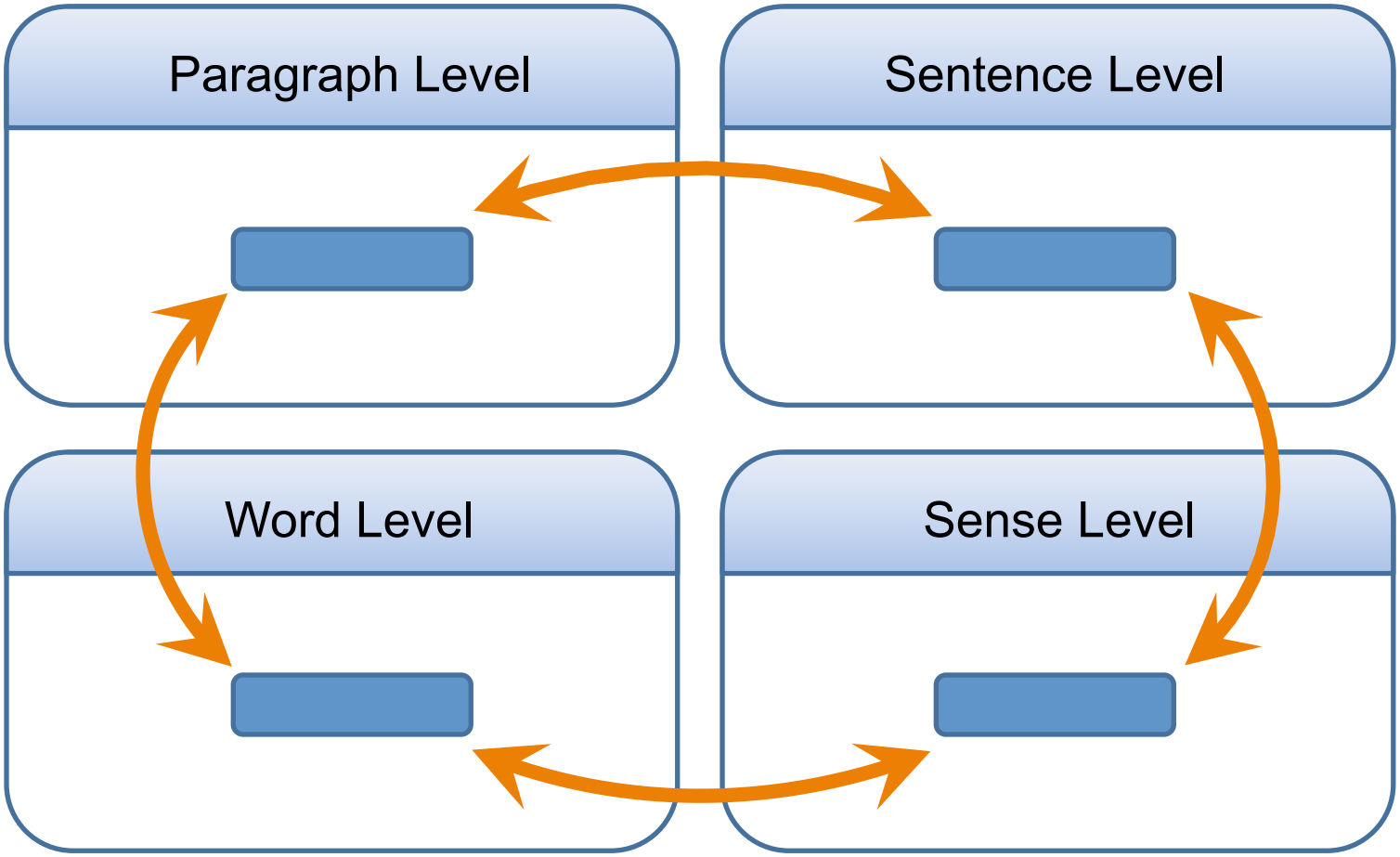
Semantic Similarity

Mostly focused on similar types of lexical items



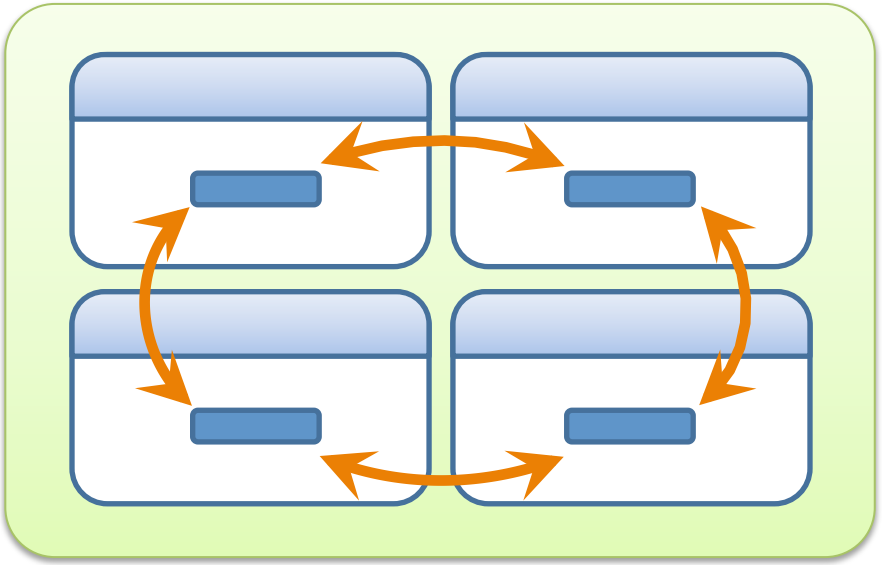
Semantic Similarity

What if we have different types of inputs?

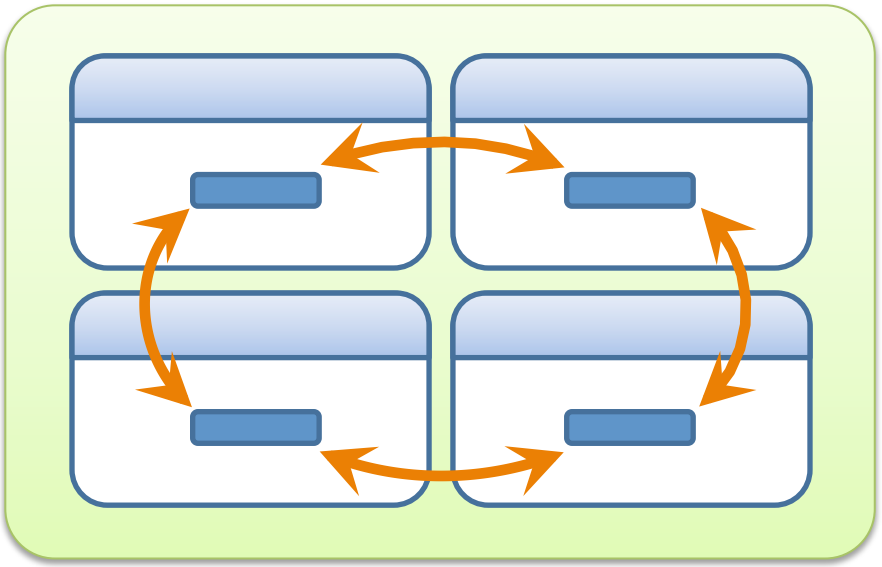


CLSS: Cross-Level Semantic Similarity

A new type of similarity task



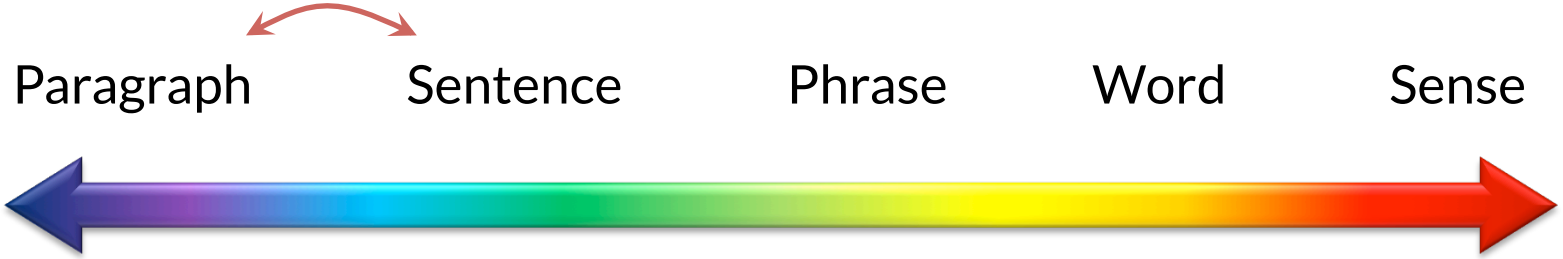
CLSS: Cross-Level Semantic Similarity



- Multiple types of comparison
- Incorporate multiple genres of text
- Push towards computing the similarity of anything

CLSS: Comparison Types

Paragraph to Sentence



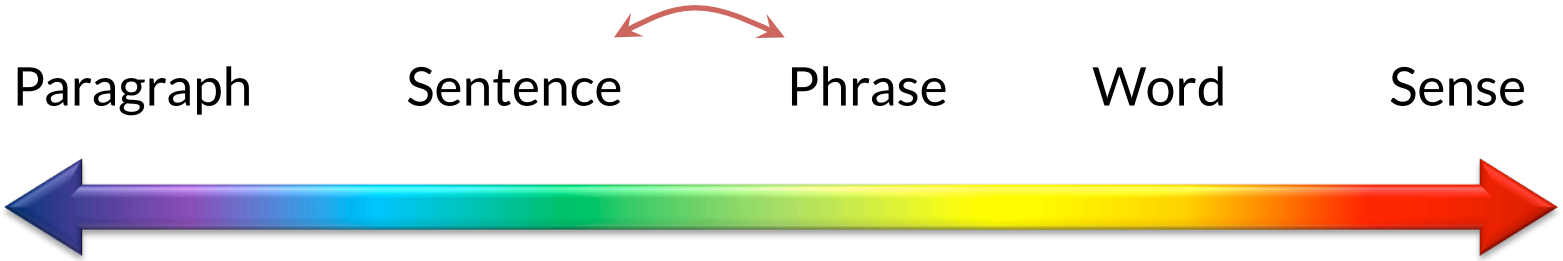
CLSS: Comparison Types

Paragraph to Sentence

Sentence to Phrase

The 30-year-old woman has had no contact with the outside world.

30-year-old female recluse



CLSS: Comparison Types

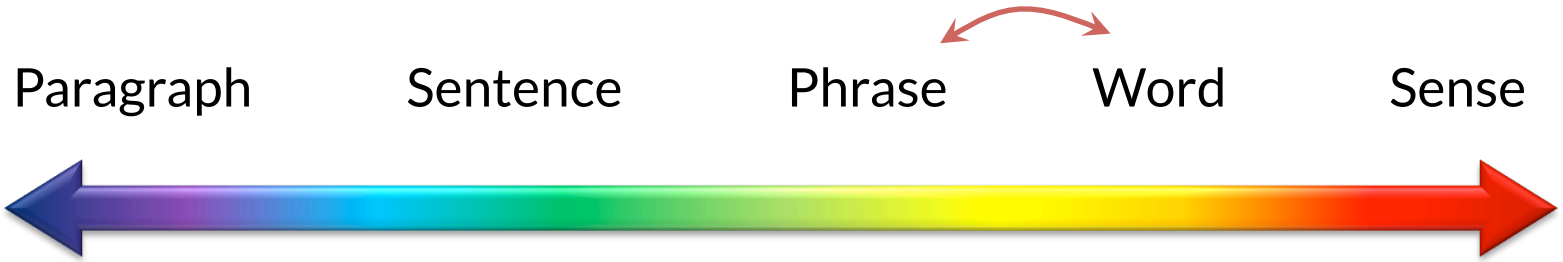
Paragraph to Sentence

Sentence to Phrase

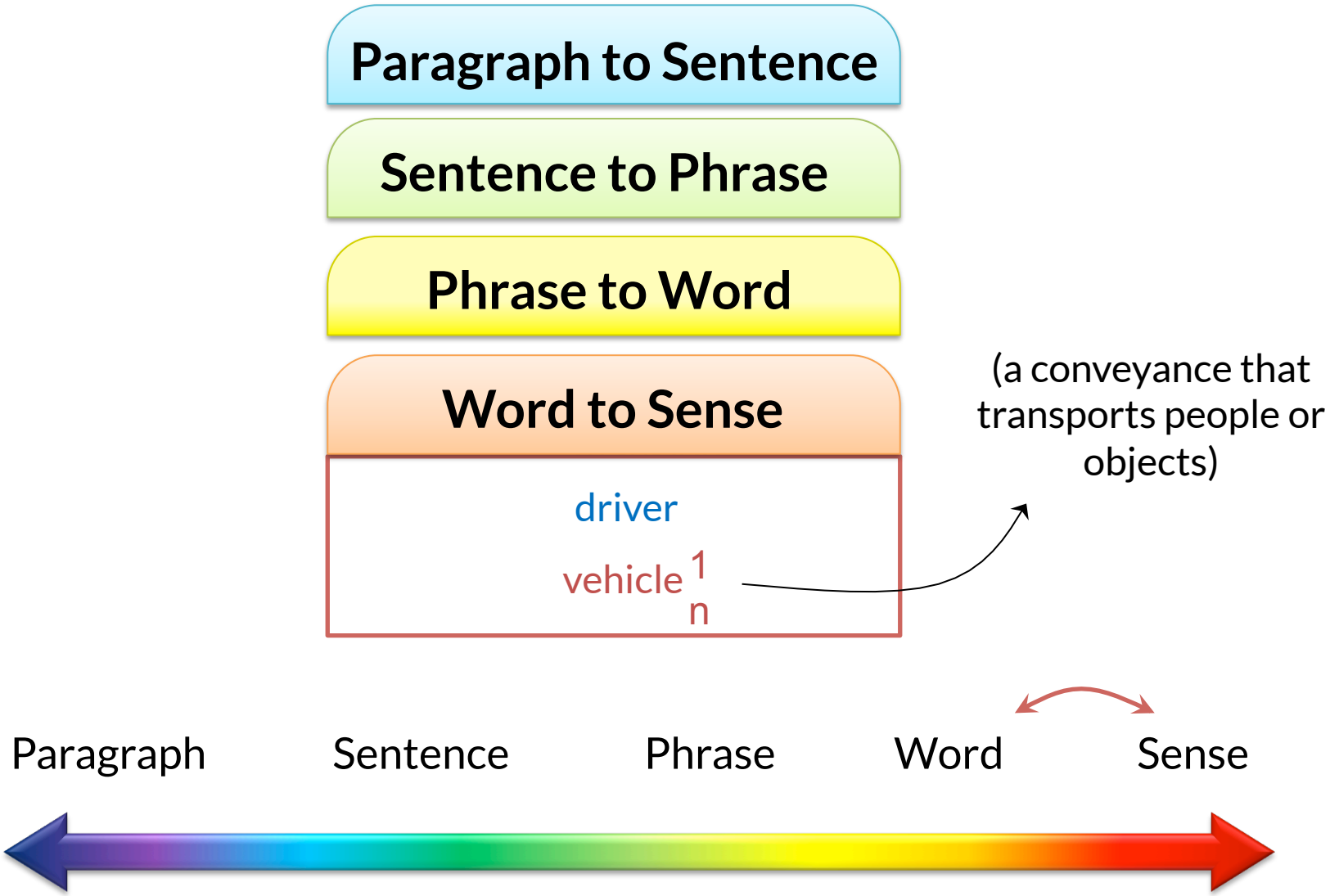
Phrase to Word

a large, expensive house

mansion



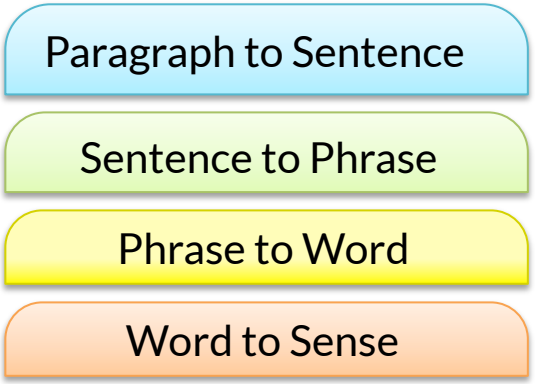
CLSS: Comparison Types



Task Data

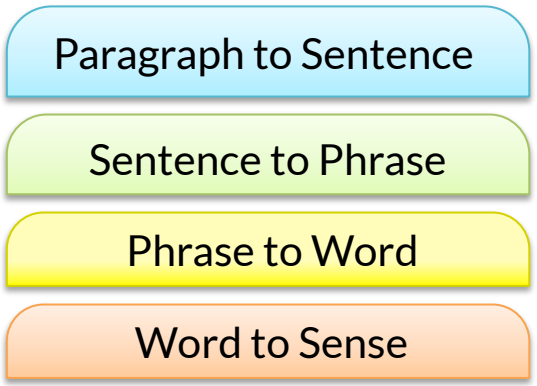
4000 pairs in total

500 pairs per type



Training set

500 pairs per type



Test set

Task Data

A wide range of domains and text styles

Paragraph to Sentence	Sentence to Phrase	Phrase to Word
Newswire	Newswire	Newswire
Travel	Travel	Slang
Question Answering	Question Answering	Idiomatic
Scientific	Scientific	Lexicographic
Metaphoric	Slang	Descriptive
Review	Idiomatic	Search

Word to Sense pairs

"Regular"	"central" *vs.* essential#a#1 "tyre" *vs.* automobile#n#1
Word not in WordNet	"zombify" *vs.* resurrect#v#3 "drank" *vs.* opiate#n#1
Sense not in WordNet	"red" *vs.* communist#a#1 "shiraz" *vs.* grape#n#1

Rating Scale

4 -- *Nearly* identical

3 -- Similar, but not identical

2 -- Related but not similar

1 -- On the same topic, but not closely related

0 -- Completely unrelated

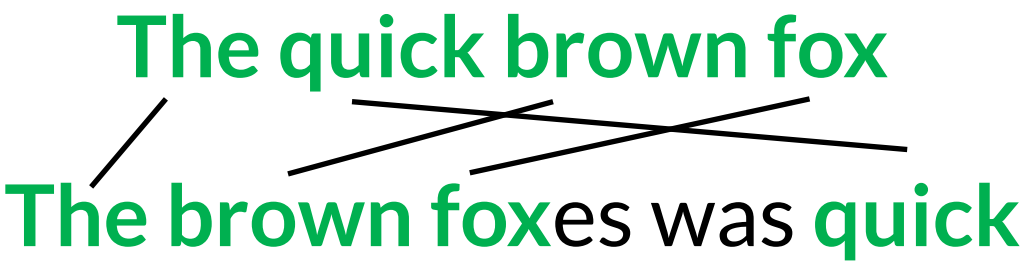
Comparison Baselines

- Longest Common Substring (LCS)

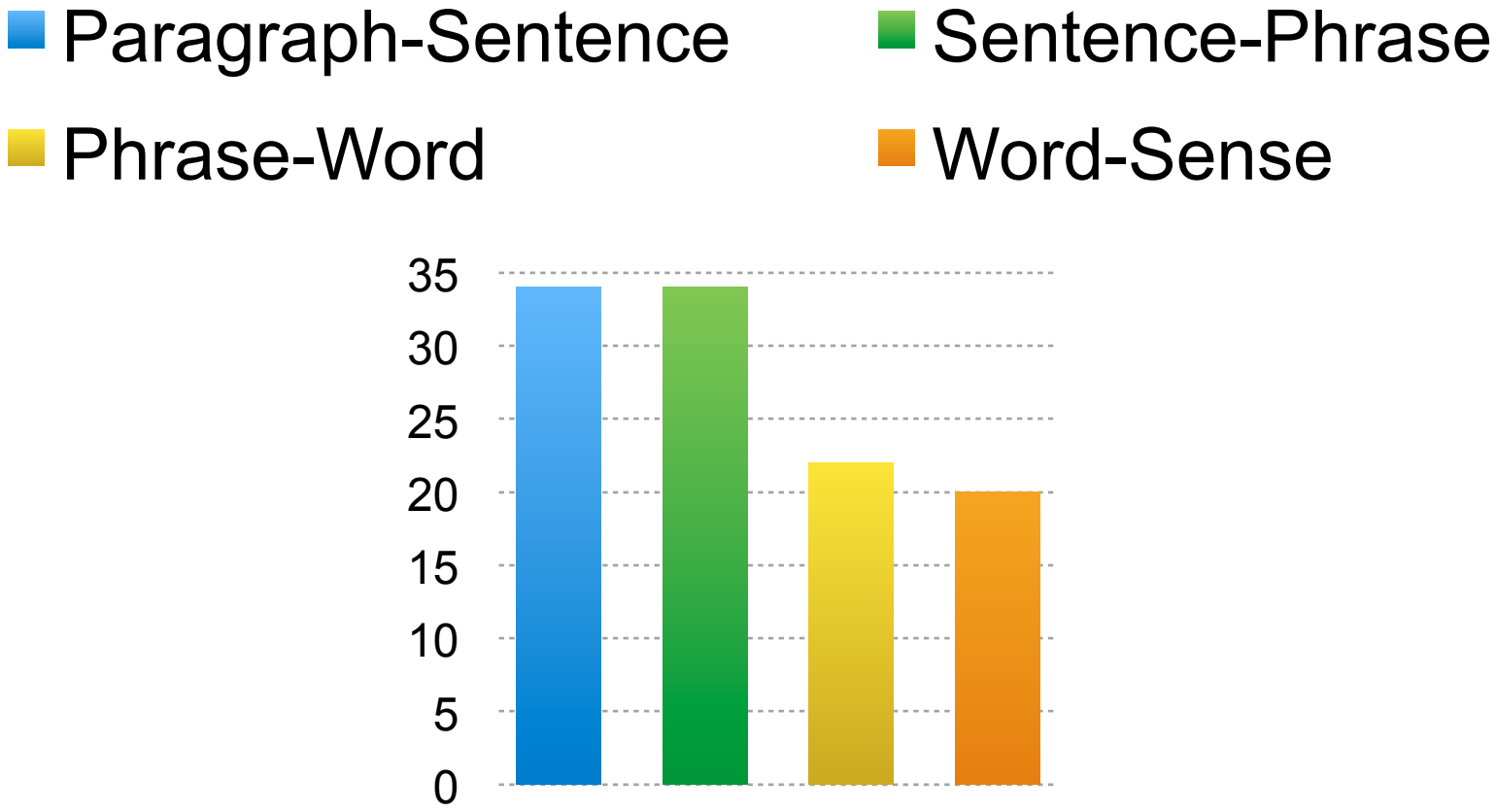
The quick **brown fox**

The **brown fox** was quick

- Greedy String Tiling (GST)

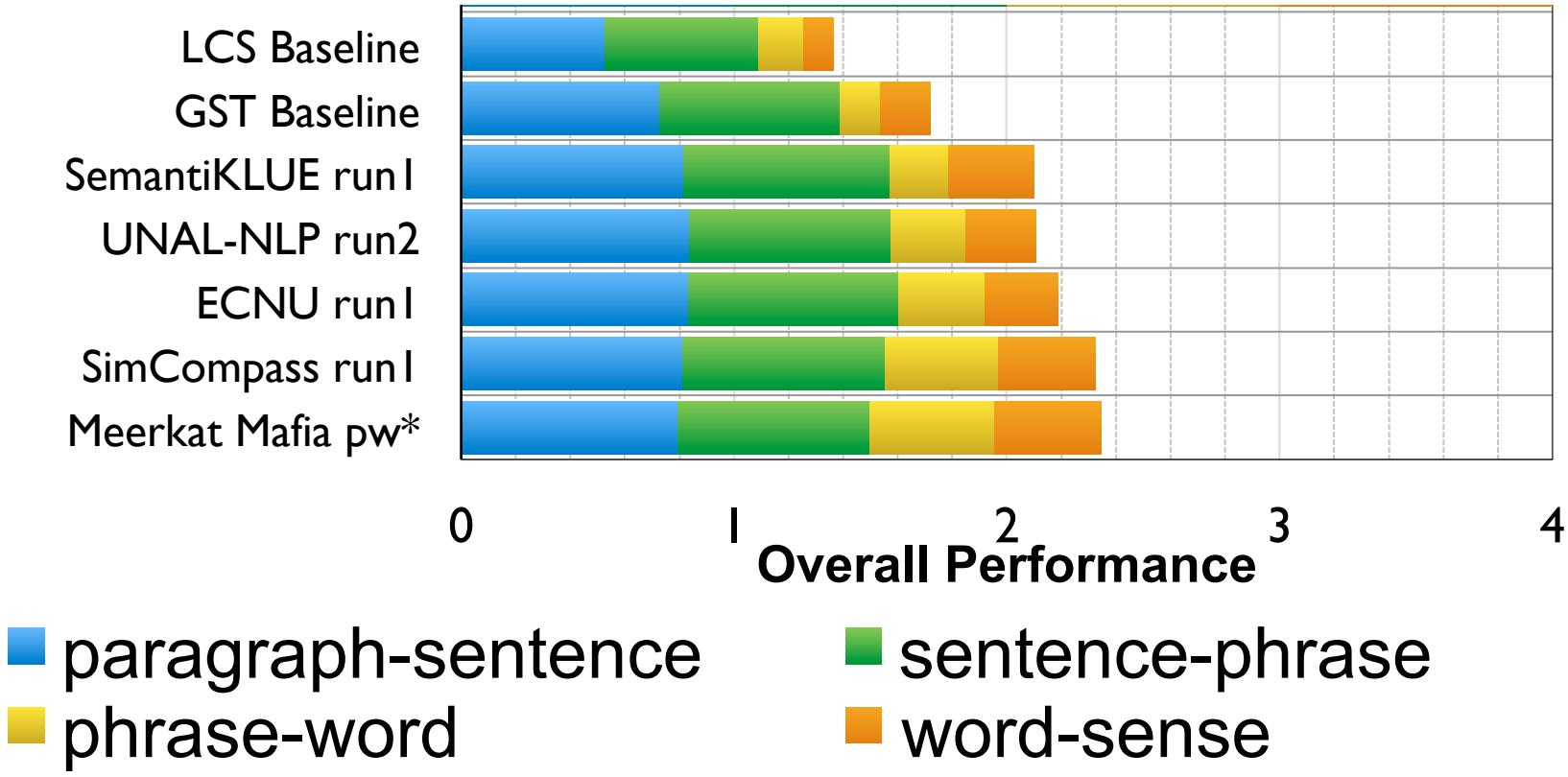


Number of participants

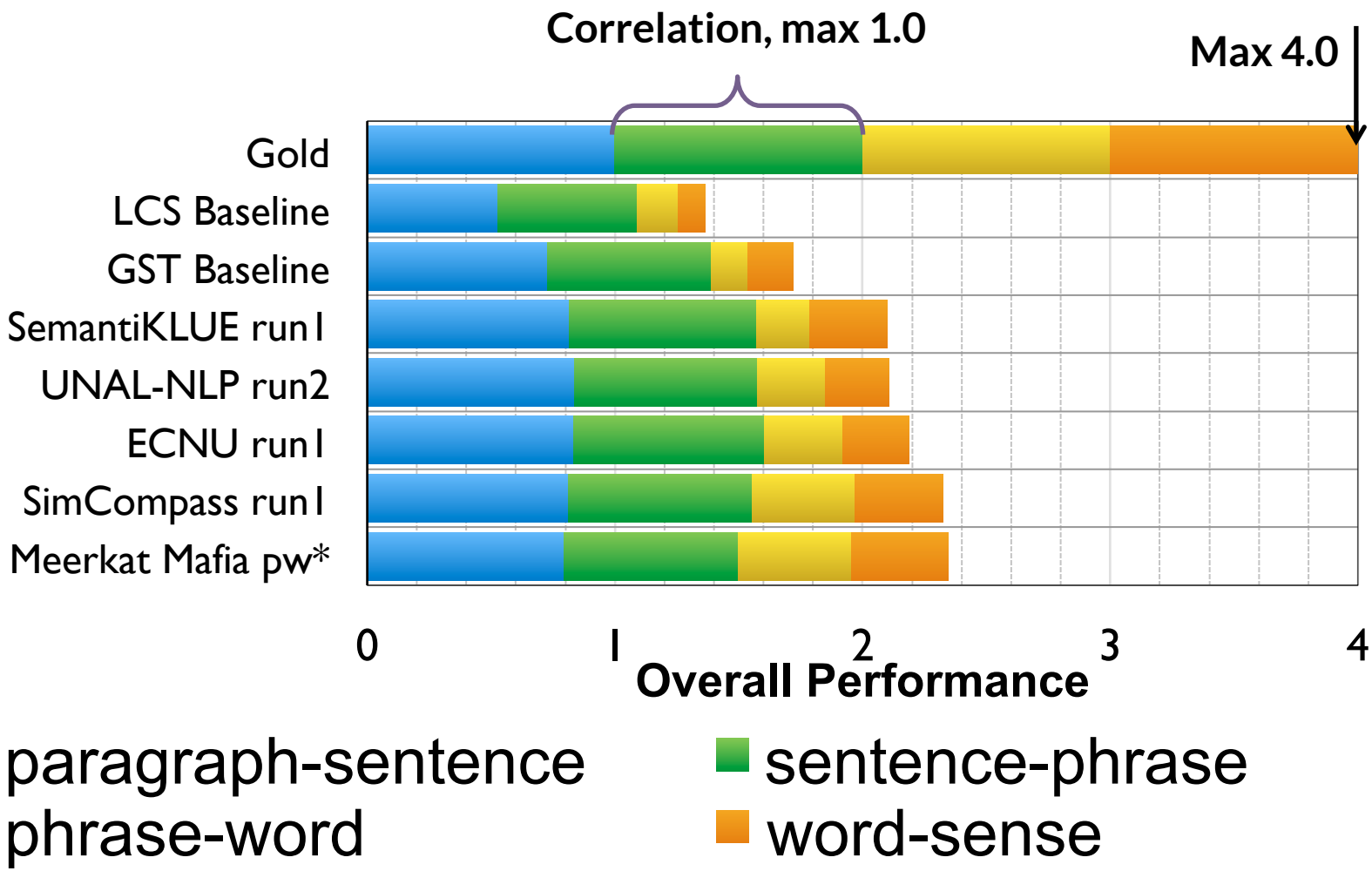


38 Systems total from 19 teams

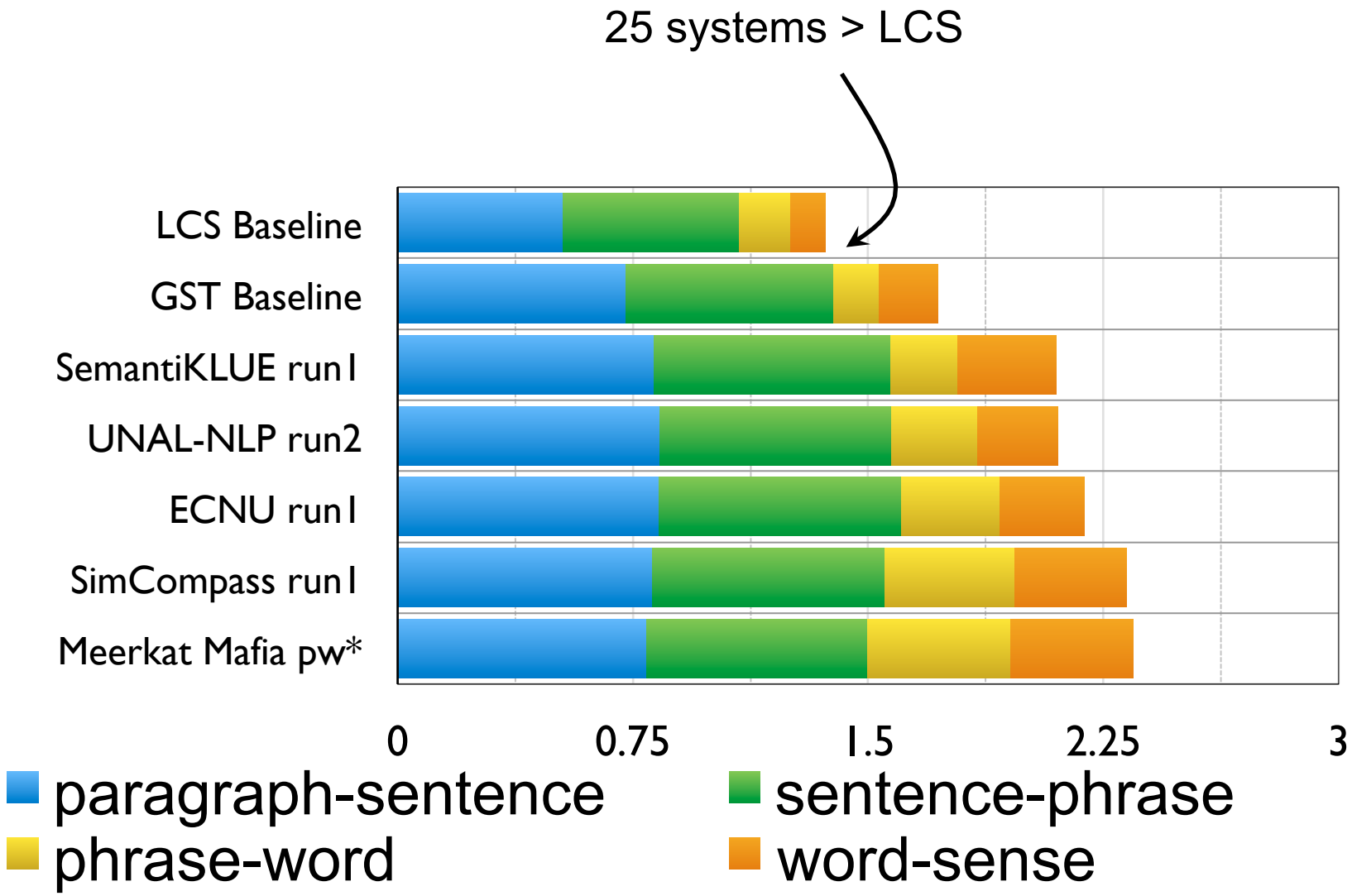
Top 5 Systems and Baselines



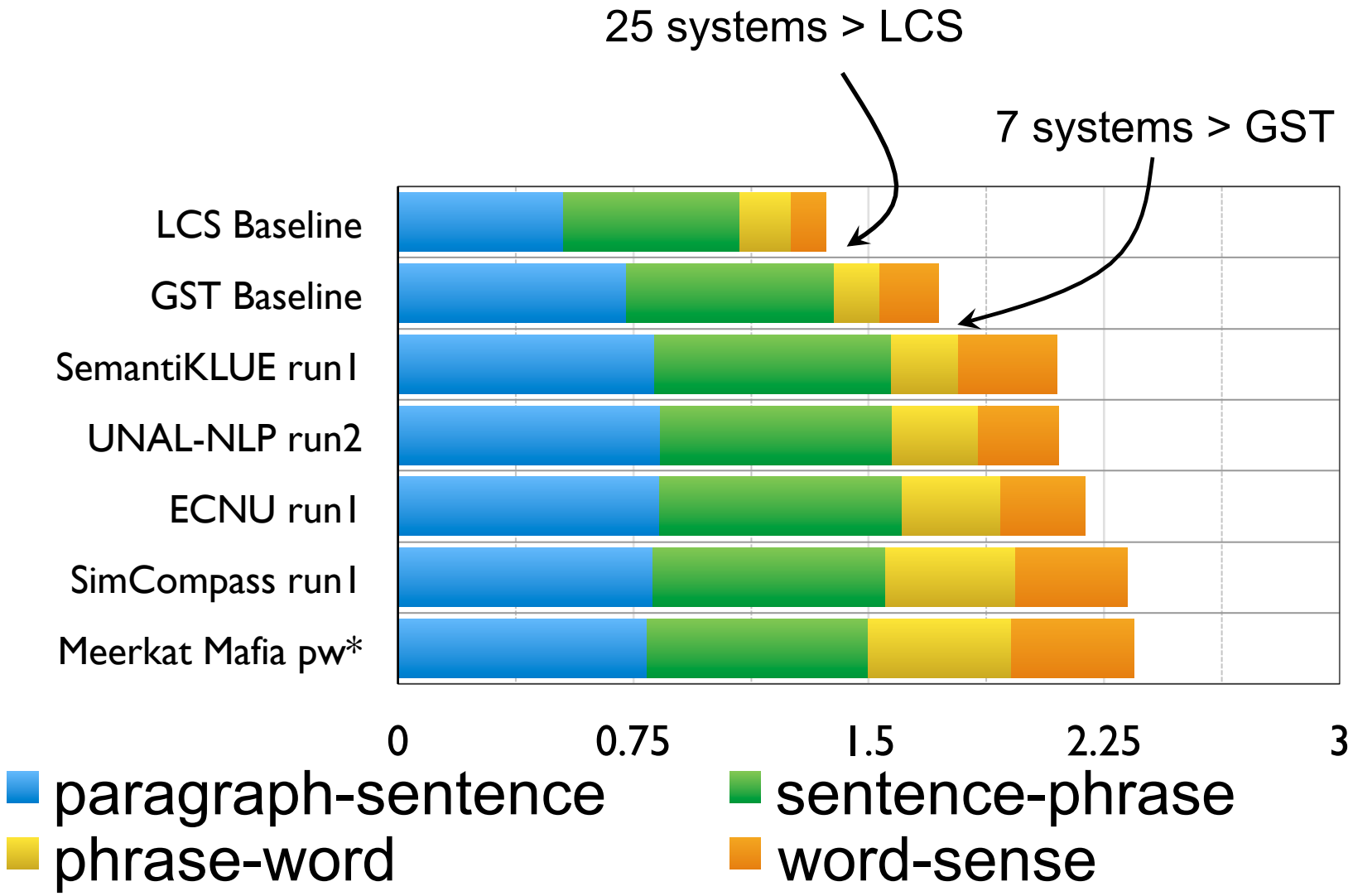
Top 5 Systems and Baselines



Where do the baselines stand?



Where do the baselines stand?



SimCompass - Banea et al (2014)

Highest overall performance among all competing systems.

Multi-feature regression model:

- Knowledge-based
  - Different WordNet-based measures

- Corpus-based
  - Deep Learning Word Embeddings, Skip-gram (Mikolov et al, 2013)

Other novel features:

- Transform texts to a sets of topic centroids; then check for closest topics

ENCU - Zhu and Lan (2014)

Among the top three systems

Multi-feature regression model:

- String-based
- Knowledge-based
  - Different WordNet-based measures

- Corpus-based
  - LSA
- Syntactic-based

Other novel features:

Using metrics for Machine Translation evaluation for semantic similarity, e.g., TER, METEOR, BLEU, etc.

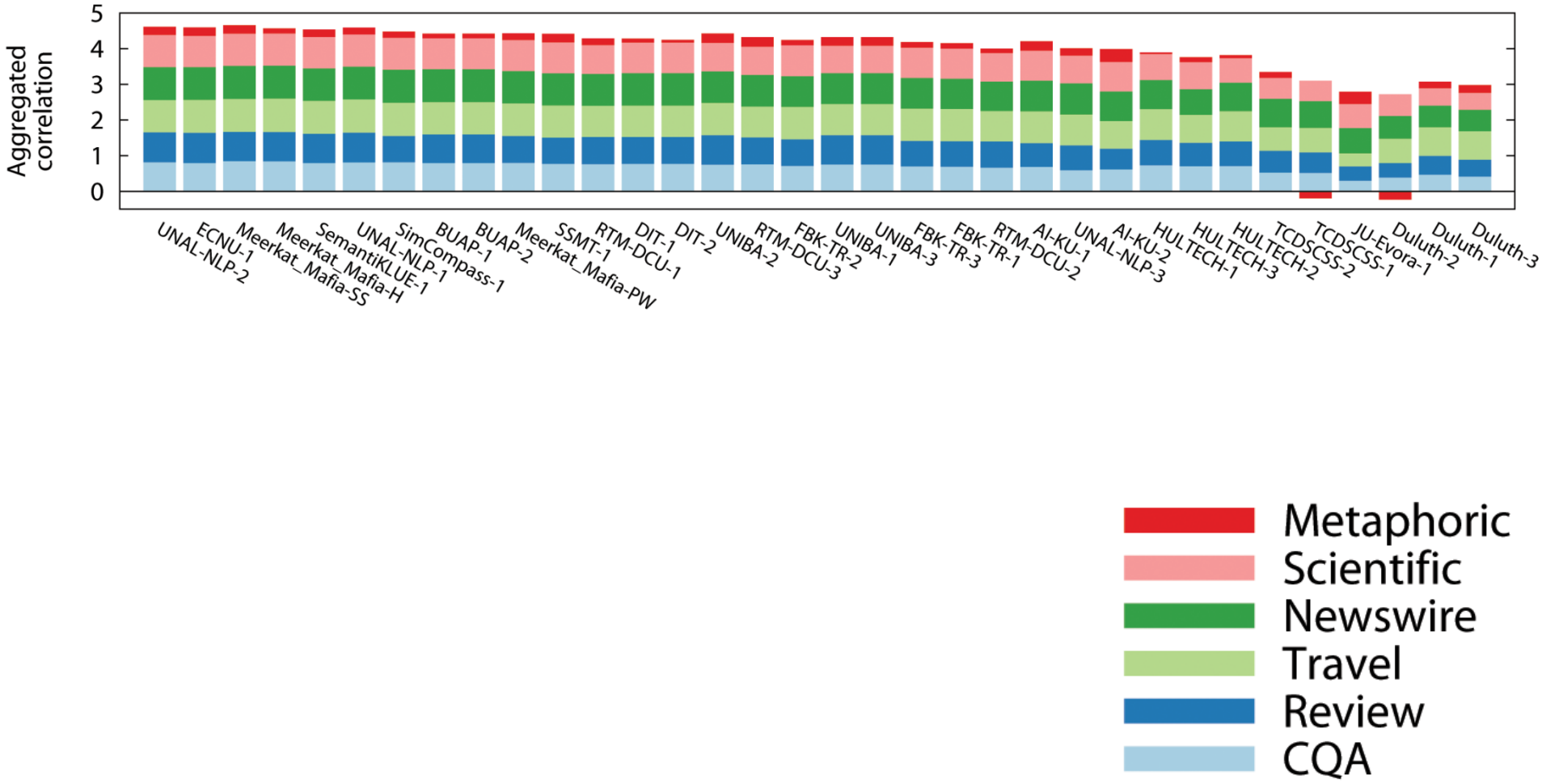
UNAL-NLP - **Jimenez et al (2014)**

Third best system overall

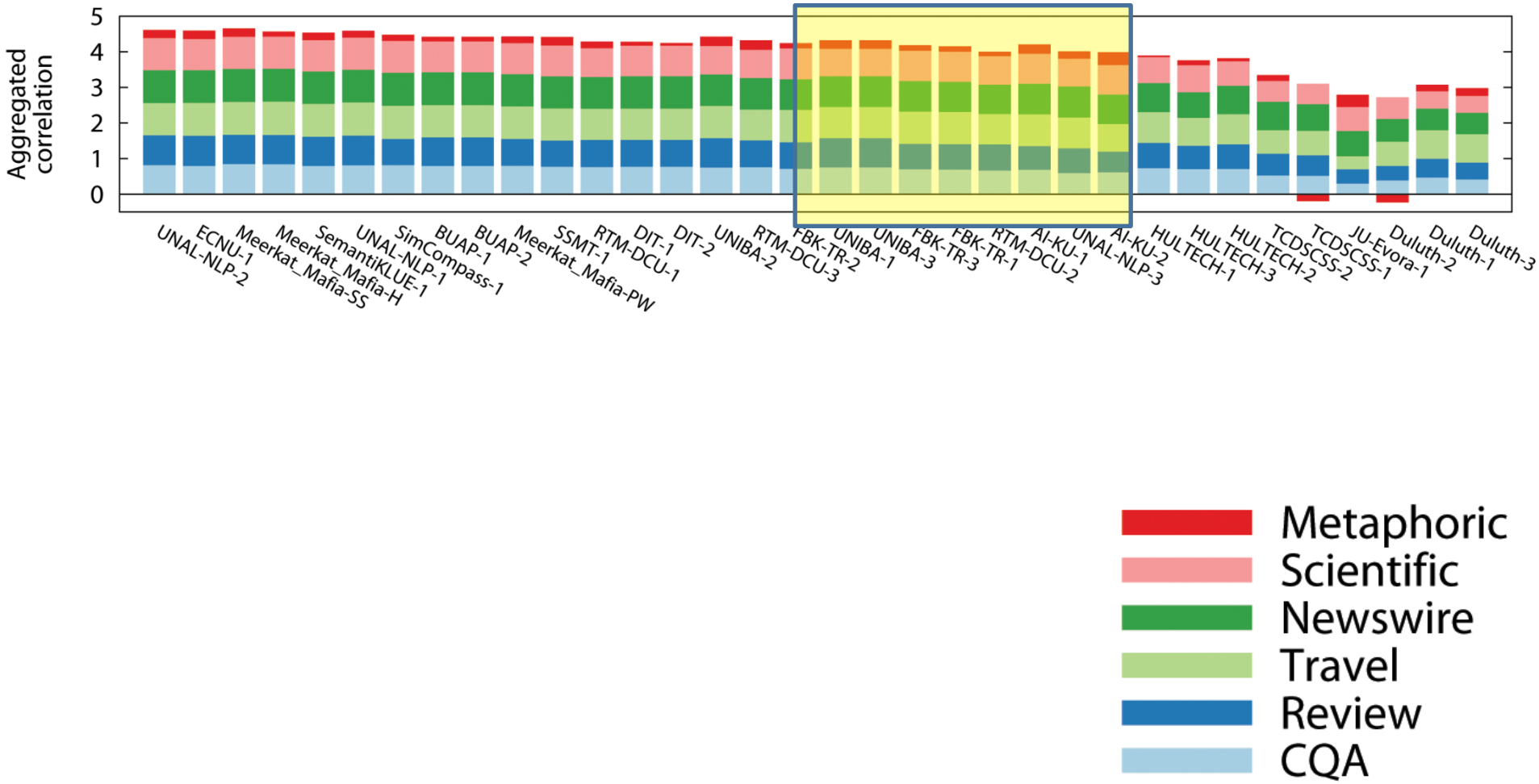
Utilizes only a set of simple **string-similarity features** based on soft cardinality (Jimenez et al, 2010).

UNAL-NLP *run1*, ranked 5th, is unsupervised: mirroring the potential for unsupervised semantic similarity measured seen in the recent work of Sultan et al (2014, 2015).

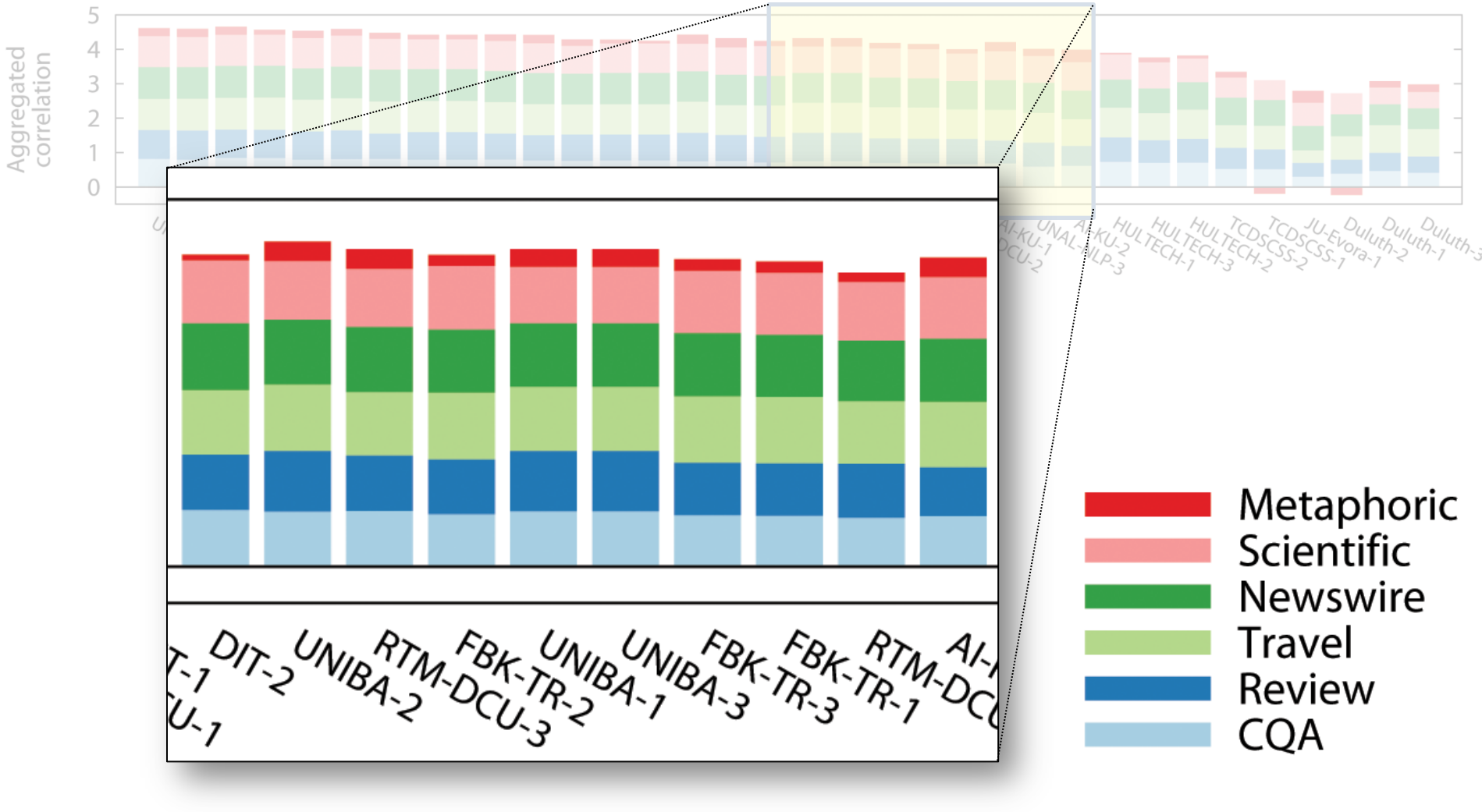
Correlation per genre
paragraph-to-sentence



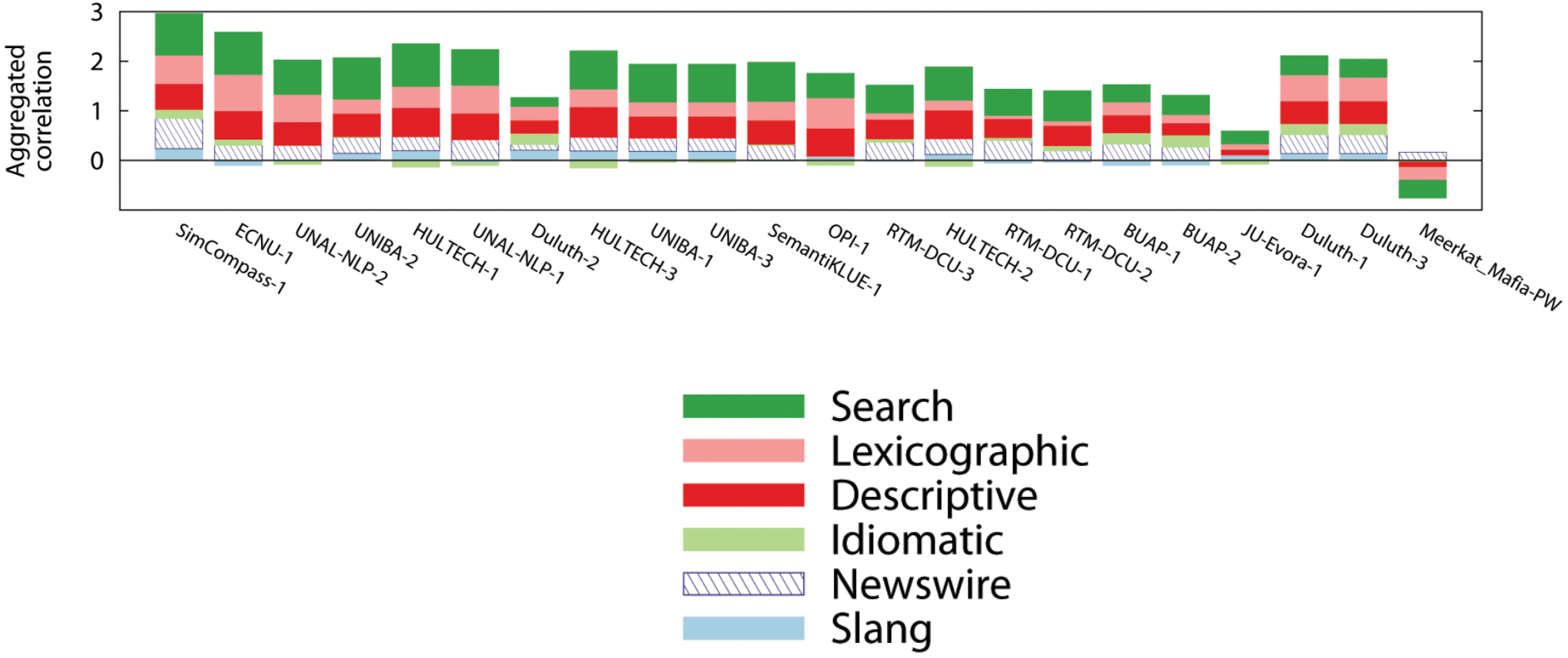
Correlation per genre
paragraph-to-sentence



Correlation per genre
paragraph-to-sentence



Correlation per genre
phrase-to-word



What makes the task difficult?

Handling OOV words and novel usages

How often do draik eggs come in Merifoods in Meridell?

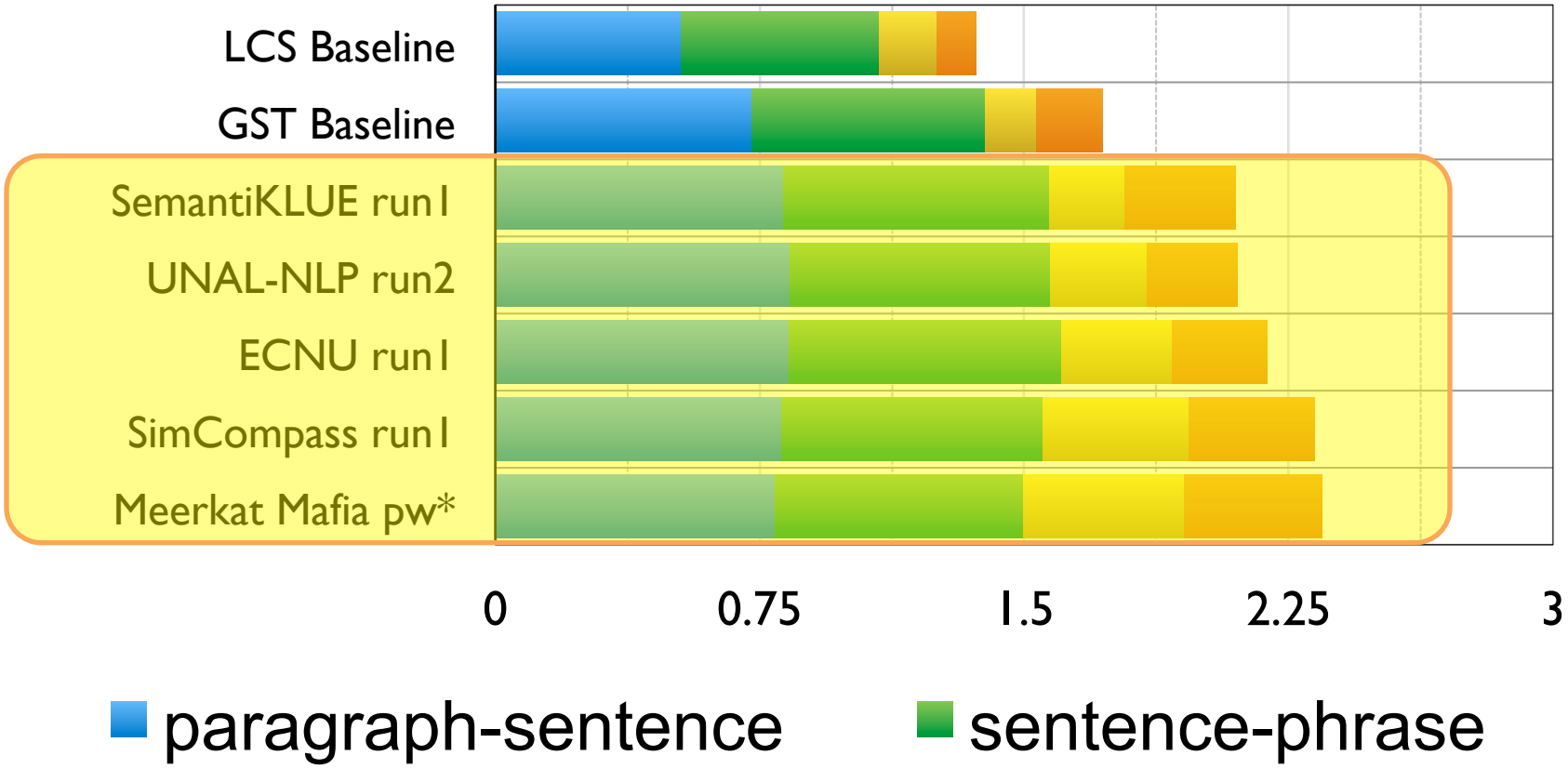
frequency of draik eggs in Merifoods

Hard feelings

grudge

WordNet alone is too limited

Include multiple dictionaries or
use distributional methods



Dealing with social media text

can i watch 4od bbc iplayer etc with 10GB useage allowence?

online television streaming for bbc

Can d Internet companies see which websyts ive bin visiting?

internet provider's knowledge of my actions

Fables

A Groom used to spend whole days in currycombing and rubbing down his Horse, but at the same time stole his oats and sold them for his own profit. “Alas!” said the Horse, “if you really wish me to be in good condition, you should groom me less, and feed me more.”

Horses need food to look their best.

Fables in real world

The Fields Medals are regarded as mathematics' Nobel Prize, and are awarded every four years. All the previous 52 winners of the Fields have been men since its inception in 1936.

Mathematics is a male-dominated research area.

Open Source Tools for Semantic Similarity

Sense
Word
Phrase
Sentence
Para/Doc

Tools

WordNet::Similarity

- Word and sense similarity (Ted Pederson)
  - in Perl
    - also available in Java, by Hideki Shima
      WS4J: http://code.google.com/p/ws4j/
  - Many common WordNet Similarity measures
    - Leacock & Chodorow (1998)
    - Jiang & Conrath (1997)
    - Resnik (1995)
    - Lin (1998)
    - Hirst & St-Onge (1998)
    - Wu & Palmer (1994)
    - The extended gloss overlap measure by Banerjee and Pedersen (2002)
    - Two measures based on context vectors by Patwardhan (2003).

Sense
Word
Phrase
Sentence
Para/Doc

Tools

Align, Disambiguate and Walk: ADW (ACL 2013)

- Multi-level similarity
  - From word senses to texts
  - All inputs have comparable representations

- Implicit word sense disambiguation

- Publicly available in Java

https://github.com/pilehvar/adw

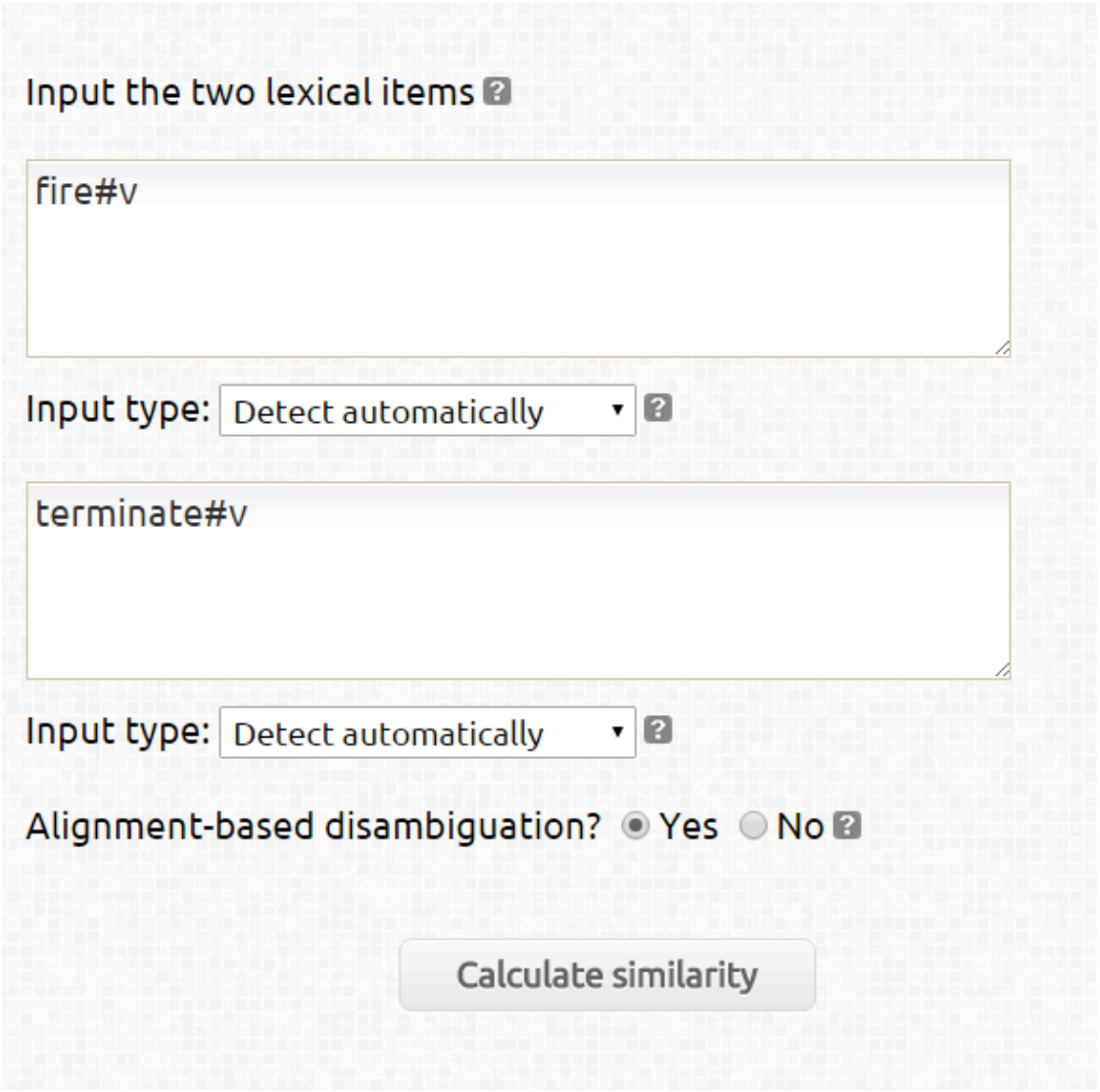
Sense
Word
Phrase
Sentence
Para/Doc

Tools

Align, Disambiguate and Walk: ADW

Online demo at

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



Sense
Word
Phrase
Sentence
Para/Doc

Tools

NLTK

http://www.nltk.org/

- A large NLP package with support for many kinds of operations on text
- Integrated with WordNet with easy support for most sense- and word-similarity measures
- Written in Python

Sense
Word
Phrase
Sentence
Para/Doc

Tools

Spacy

http://spacy.io/

- A large NLP package with support for many kinds of operations on text
  - Fast POS taggers, parsers, with state of the art-level performance

- Built in support for representing words with dependency-based word2vec vectors (Levy and Goldberg, 2014)

- Written in Python

- Free for OSS / Research, Commercial license available

Sense
Word
Phrase
Sentence
Para/Doc

Tools

DKProSimilarity

https://github.com/dkpro/dkpro-similarity

- Open source framework for text similarity, Java
- Best system SemEval STS-12 Task
- Several similarity measures, including:

algorithms.lexical	GreedyStringTiling, Levenshtein, NGramBased, ...
algorithms.lsr	Based on WordNet or Wikipedia
algorithms.style	FunctionWordFrequency, MTLD, TypeTokenRatio
algorithms.vsm	Vector-space models, e.g. ESA
algorithms.wikipedia	Special Wikipedia measures, e.g., WikipediaLinkMeasure

Sense
Word
Phrase
Sentence
Para/Doc

Tools

TakeLab

http://takelab.fer.hr/sts/

- Open source framework for text similarity, Python
- Among the top five in STS-12
- Several similarity measures, including:

Lexical	WordNet-based measures from NLTK
Knowledge-based	GreedyStringTiling, Levenshtein, NGramBased, etc
Corpus-based	Latent Semantic Analysis
Syntactic	Syntactic role similarity, syntactic dependency similarity
Other	Normalized differences, number overlap, etc.

Tools

S-Space Package

https://github.com/fozziethebeat/S-Space

- Open source framework for word distributions
- Written in Java
- Support for common weighting (e.g., PMI) and matrix factorizations (e.g,. SVD)
- Implements many common algorithms in a single interface
  - LSA, word2vec, COALS, GloVe, random indexing
- Integrated pre-processing support using Stanford CoreNLP

Sense
Word
Phrase
Sentence
Para/Doc

Tools
DISSECT

http://clic.cimec.unitn.it/composes/toolkit/

- Open source framework for word distributions
- Written in Python
- Support for common weighting (e.g., PMI) and matrix factorizations (e.g,. SVD)
- Designed around compositionality
  - Easy to build representation for larger phrases

Sense
Word
Phrase
Sentence
Para/Doc

Tools

word2vec

https://code.google.com/p/word2vec/

- Tomas Mikolov (in C)
- Efficient implementation of the continuous bag-of-words and skip-gram architectures for **word** representation
- Dependency-based version available from Omer Levy
  - https://bitbucket.org/yoavgo/word2vecf
- Also available in
  - Java: DL4J, Deep Learning 4 Java http://deeplearning4j.org/word2vec.html
  - Spark MLib: https://spark.apache.org/docs/latest/mllib-feature-extraction.html#word2vec
  - Python: as a part of gensim http://radimrehurek.com/2013/09/deep-learning-with-word2vec-and-gensim/

Sense
Word
Phrase
Sentence
Para/Doc

Tools

http://nlp.stanford.edu/projects/glove/

GloVe: Global Vectors for Word Representation

- Written by Jeffrey Pennington, Richard Socher, Christopher D. Manning (in C) as an alternative to word2vec

- Efficient implementation, with pre-trained vectors available

- Also available in
  - Java: DL4J, Deep Learning 4 Java
    http://deeplearning4j.org/word2vec.html

Sense
Word
Phrase
Sentence
Para/Doc

Tools

Gensim

https://radimrehurek.com/gensim/

- Originally written for high-performance LSA
- Now includes support for many kinds of topic modeling and word2vec
  - Usually where new algorithms get first implemented
- Fast and written in Python
- Commercial support available

Sense
Word
Phrase
Sentence
Para/Doc

Tools

doc2vec

https://radimrehurek.com/gensim/models/doc2vec.html

- Implemented in Python as a part of gensim

- Efficient implementation of the continuous bag-of-words and skip-gram architectures for paragraph-level representations

Sense
Word
Phrase
Sentence
Para/Doc

Tools

http://mallet.cs.umass.edu/

MALLET - MAchine Learning for LanguagE Toolkit

- A software package for building all kinds of probabilistic models from text
- Scalable and fast support for LDA and the hierarchical Pachinko Allocation Model
- Written in Java

Sense
Word
Phrase
Sentence
Para/Doc

Tools

Other topic modeling software

- Huge list of topic modeling software available at http://www.cs.columbia.edu/~blei/topicmodeling_software.html
  - with an active mailing list too
- Adams et al. (2010) hierarchical topic model
  - http://hips.seas.harvard.edu/files/tssb.tgz
- Highlights include:
  - LDA in C (fast!)
  - HDP in C
  - TurboTopics in Python

Sense
Word
Phrase
Sentence
Para/Doc

Resources

Out of vocabulary or rare words

- Medial Subject Headings (MeSH)
  - https://www.nlm.nih.gov/mesh/
- Wiktionary
  - https://www.wiktionary.org
- Wordnik: “world's biggest online English dictionary”
  - https://www.wordnik.com/
- Collaborative International Dictionary of English
  - http://gcide.gnu.org.ua/
- Moby Thesaurus II
  - http://goo.gl/fzRRCF
- The Free On-line Dictionary of Computing
  - http://foldoc.org/

Sense
Word
Phrase
Sentence
Para/Doc

Resources

Crown

https://github.com/davidjurgens/crown

- Extension of WordNet with new synsets and lexicalizations
  - 2X the size of WordNet
    - Slang, archaic forms, idioms, technical words, …
- Released as stand-off dictionaries, so compatible with all WordNet libraries
  - NLTK, WordNet::Similarity

Sense
Word
Phrase
Sentence
Para/Doc

Resources

BabelNet

http://babelnet.org/

- Combination of many resources into a single representation
  - WordNet, Wikipedia, Wiktionary
- Can be combined with Babelfy to disambiguate text to sense level
- Support for cross-lingual mapping of concepts across 271 languages
- Written in Java, but has REST API as well

Pre-trained Word Vectors

- Word2vec
  - https://code.google.com/p/word2vec/
  - https://github.com/3Top/word2vec-api
- Baroni and Lenci, Distributional memory
  - http://clic.cimec.unitn.it/dm/
- GloVe
  - http://nlp.stanford.edu/projects/glove/
- Faruqui and Dyer (ACL 2014)
  - http://wordvectors.org/
- Huang et al (2012), Multiple Word Prototypes
  - http://www-nlp.stanford.edu/~ehhuang/wordrep.zip
- Levy and Goldberg (2014), dependency-based word embeddings
  - https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/

Open Problems in Semantic Similarity

Open Problem: Irregular Language

can i watch 4od bbc iplayer etc with 10GB useage allowence?

online television streaming for bbc

Can d Internet companies see which websyts ive bin visiting?

internet provider's knowledge of my actions

Open Problem: Multi-word Expressions (MWEs)

- Most approaches either ignore MWEs or recognize those from fixed lists of MWEs
  - Problematic unless lemmatizing
  - Even more problematic with syntactic rearrangement

We need to **sort out** the problem

We need to **sort** the problem **out**

Open Problem: Multi-word Expressions (MWEs)

- New SemEval-2016 task on super-sense tagging seems like a promising direction for addressing this

> I PRP googled VBD restaurants NNS in IN the DT area NN and CC Fuji NNP Sushi NNP came VBD up RB and CC reviews NNS were VBD great JJ so RB I PRP made VBD a DT carry VB out RP order NN

the goal is to predict the representation

> I googled communication restaurants GROUP in the area LOCATION and Fuji _ Sushi GROUP came _ up communication and reviews COMMUNICATION were stative great so I made _ a carry _ out possession _ order communication

example from the Task's website

Open Problem: Cross-Language Similarity

- Beneficial for Machine Translation evaluation or even applications like plagiarism detection
- **Recent benchmarks** by Camacho-collados et al. (2015) and Leviant and Reichart (2015)

  Cross-lingual datasets constructed based on **RG-65** (FR, DE, EN, FA, ES, and PT) and **WS353** (EN, DE, IT, and RU)

  http://lcl.uniroma1.it/similarity-datasets/

  http://technion.ac.il/~irakr/MultilingualVSMdata.html

Open Problem: Syntax

- **Syntax matters**
  - "Man bites dog"
  - "Dog bites man"
  - "Pitbull bites man"

Vector addition would fail in these cases

- **Compositionality** can help here but more analysis is needed
  - Recent SICK benchmark designed to explicit test for compositional ability (Marelli et al., 2014)
- Possible solution with **Abstract Meaning** Representations (AMRs)
  - Check out SemEval-2016's task!

Open Problem: Punctuations!

A woman without her man is nothing.

A woman: without her, man is nothing.

Open Problem: Variable-Sized Input

The 30-year-old woman has had "no contact with the outside world."	30-year-old female recluse

Prius

A fuel-efficient hybrid car

An automobile powered by both an internal combustion engine and an electric motor, reducing its dependence on fossil fuels

Requires smarter compositionality

Open Problem: Ambiguity

- Multiple interpretations can wreak havoc when text is limited

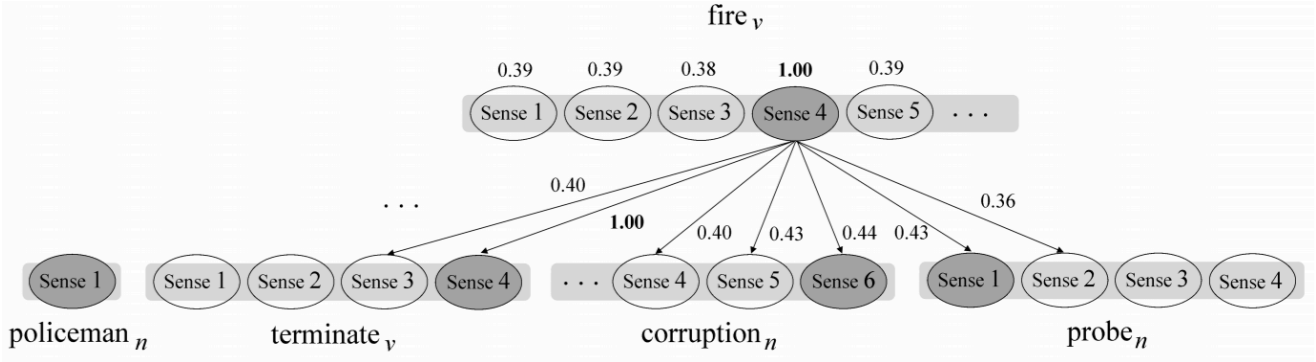
The boss **fired** his worker.

An employee was **terminated** from work by his boss.

A worker was **shot** by his boss.

Open Problem: Ambiguity

- Alignment-based disambiguation of ADW

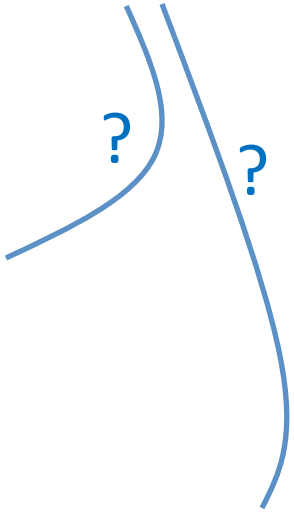


- WSD is a solution, but is still a long way off

Babelfy

Open Problem: Subjectivity vs. objectivity

As of 2012, there are 2.1 million hybrids on U.S. roads.



Hybrid cars are getting quite popular in U.S.

US hybrid vehicle market share grew by 41% in 2012.

Open Problem: Uncovered words

- Words might not have been covered in the corpus or by the lexicon;

- For instance, some WordNet OOV words:

  - prequel#n
  - fanbase#n
  - screenshot#n
  - bookmark#v
  - programmatic#a
  - broadband#n
  - And many more regular terms

  - photoshop#v
  - space_cadet#n
  - homewrecker#n
  - And many more slang terms

Open Problem: Evaluation

- Many evaluation tasks make it easy to pick-and-choose which results to report
  - 20+ choices for word similarity!
  - What exactly is state of the art?

- Similarity itself is **not an end-task**, yet most approaches are only tested on STS benchmarks, not in any application.
  - No easily-pluggable application-based tests

Semantic Similarity Frontiers: From Concepts to Documents

David Jurgens
jurgens@stanford.edu
Stanford University

Mohammad Taher Pilehvar
pilehvar@di.uniroma1.it
Sapienza University of Rome

Slides, bibliography, extended reading list, and all other materials available at http://tiny.cc/similarity-tutorial

ERC grant 259234

Bonus: must-see similarity papers at EMNLP!

- J. Li and D. Jurafsky: **Do Multi-Sense Embeddings Improve Natural Language Understanding?**
- H. He et al: **Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks**
- D. Kiela et al: **Specializing Word Embeddings for Similarity or Relatedness**
- J. Wieting and D. Roth: **Latent Variable Regression for Text Similarity and Textual Entailment**
- Sergienya and Schutze: **Learning Better Embeddings for Rare Words Using Distributional Representations**
- A. Gupta et al: **Distributional vectors encode referential attributes**