Semantic Similarity Frontiers: From Concepts to Documents

David Jurgens
Stanford University
Mohammad Taher Pilehvar
Sapienza University of Rome

EMNLP 2015
CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING
LISBON

erc
ERC grant 259234

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

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)

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)

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)

Gal

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

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

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

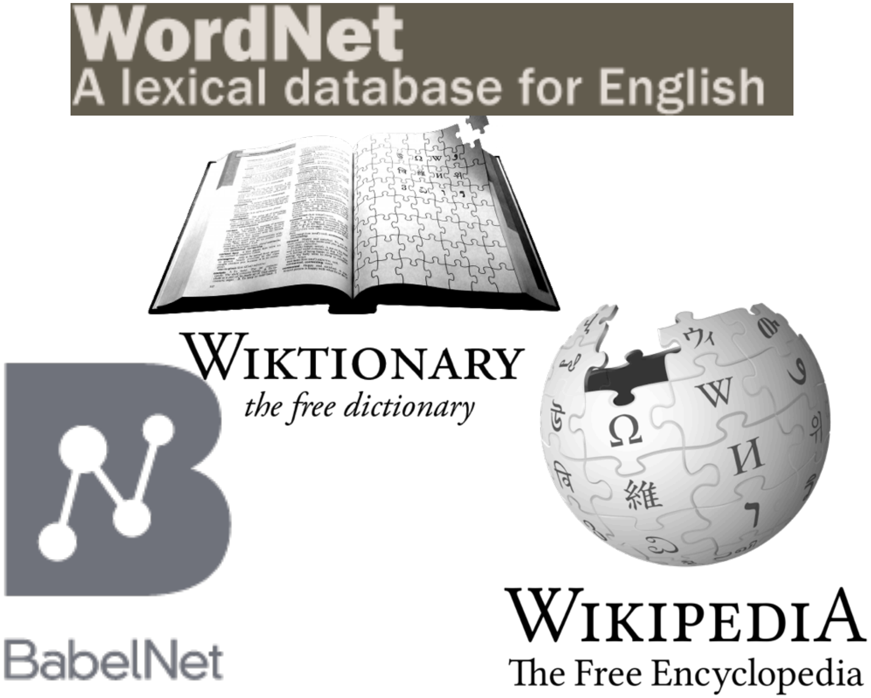
- Consistently interpretable similarity scores with explanations of why similar
- Works well for different types of text (news, web, social media, …)
- Applicable to multiple linguistic types (words, phrases sentences)

Typically, two main resources for measuring similarity

Massive corpora of text documents

Typically, two main resources for measuring similarity

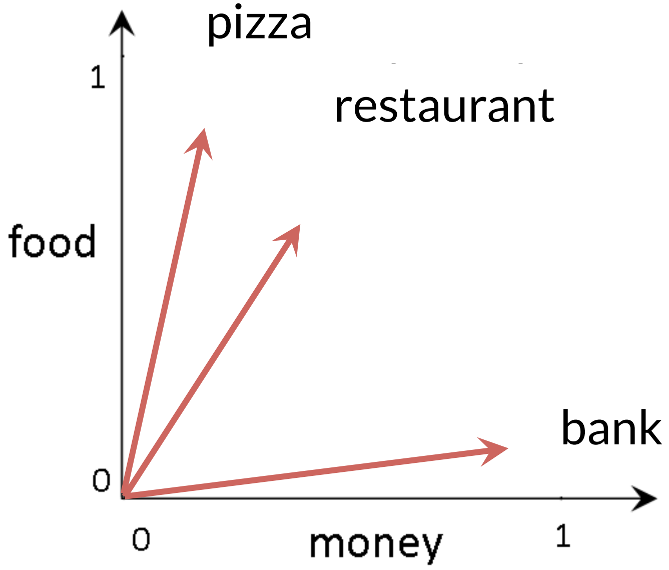
Massive corpora of text documents



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

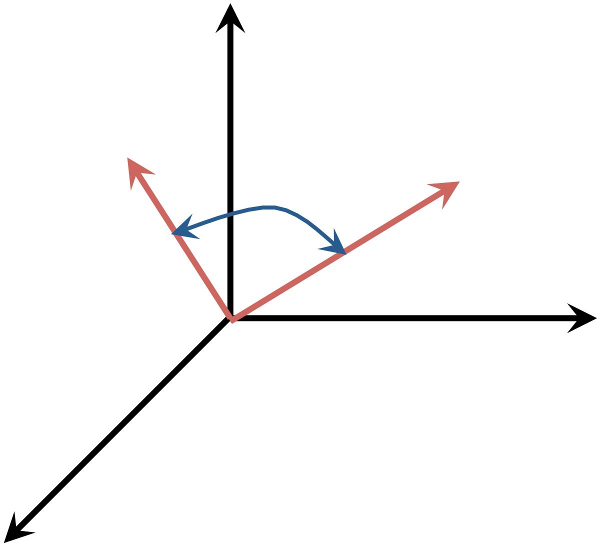
$$D_{JS}\left(\mathcal{S}_1, \mathcal{S}_2\right) = \frac{1}{2} D_{KL}\left(\mathcal{S}_1 \middle\| \frac{\mathcal{S}_1 + \mathcal{S}_2}{2}\right) + \frac{1}{2} D_{KL}\left(\mathcal{S}_2 \middle\| \frac{\mathcal{S}_1 + \mathcal{S}_2}{2}\right)$$

Vector Space Models

Vector comparison techniques

Cosine distance

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



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

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

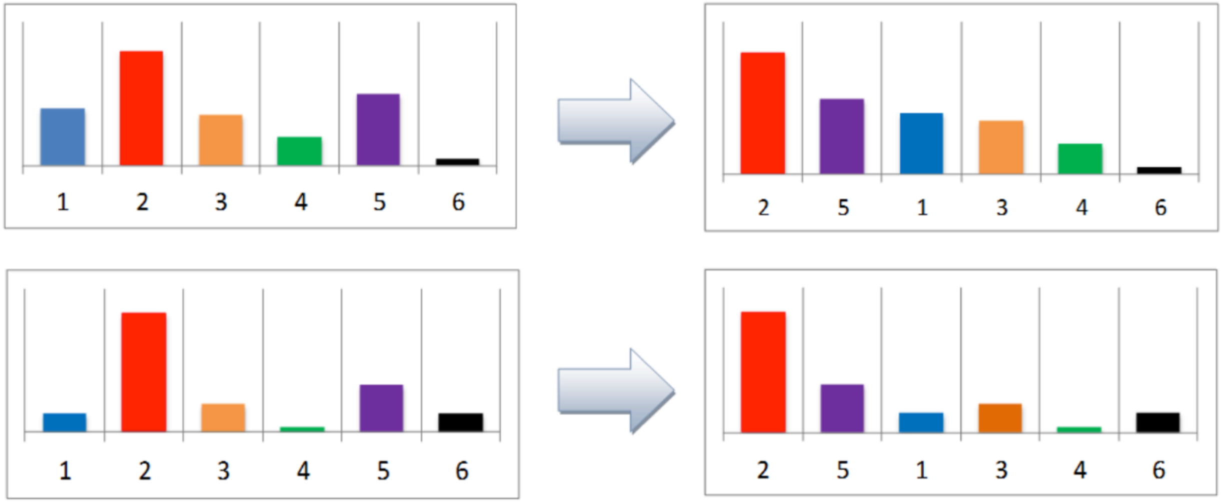
A parameter that determines the relative importance of the top elements.

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!

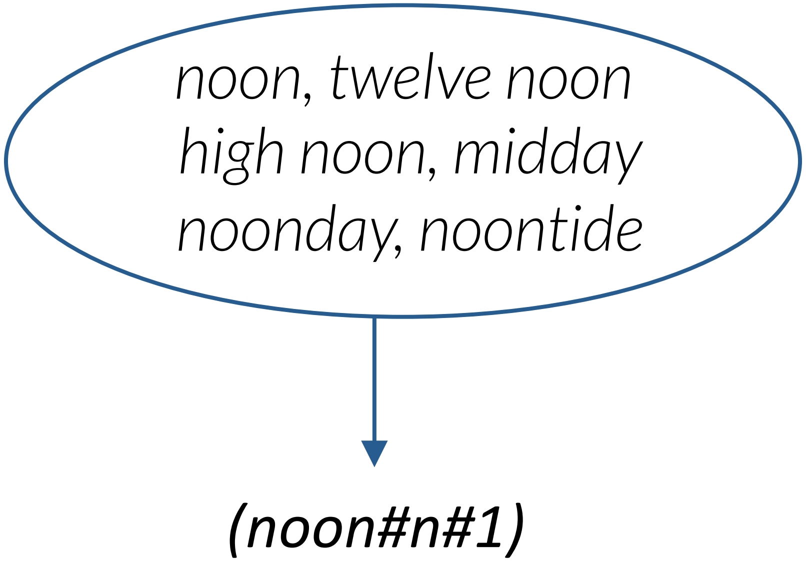
Semantic Similarity

between word senses

Concepts vs. senses

A WordNet synset (concept):

the middle of the day



noon, twelve noon
high noon, midday
noonday, noontide

(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

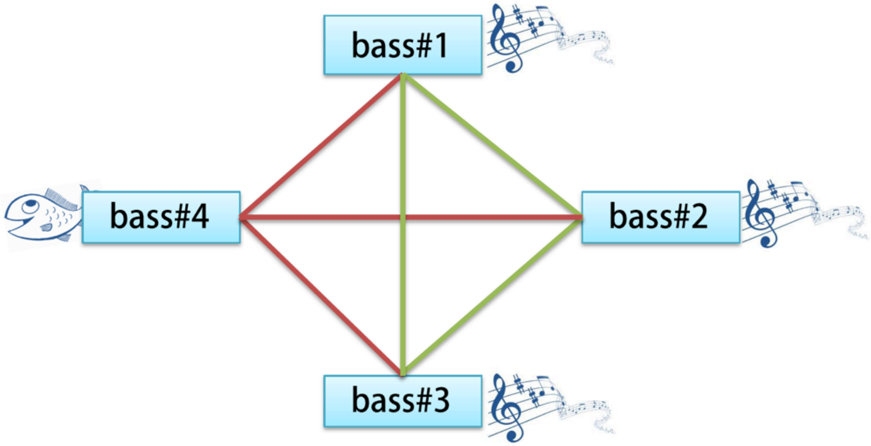
Applications - specific

WSD

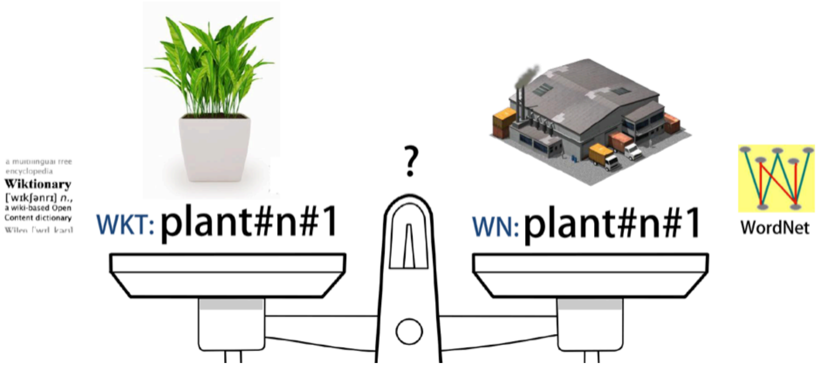
install the updated application

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

Coarsening



Alignment



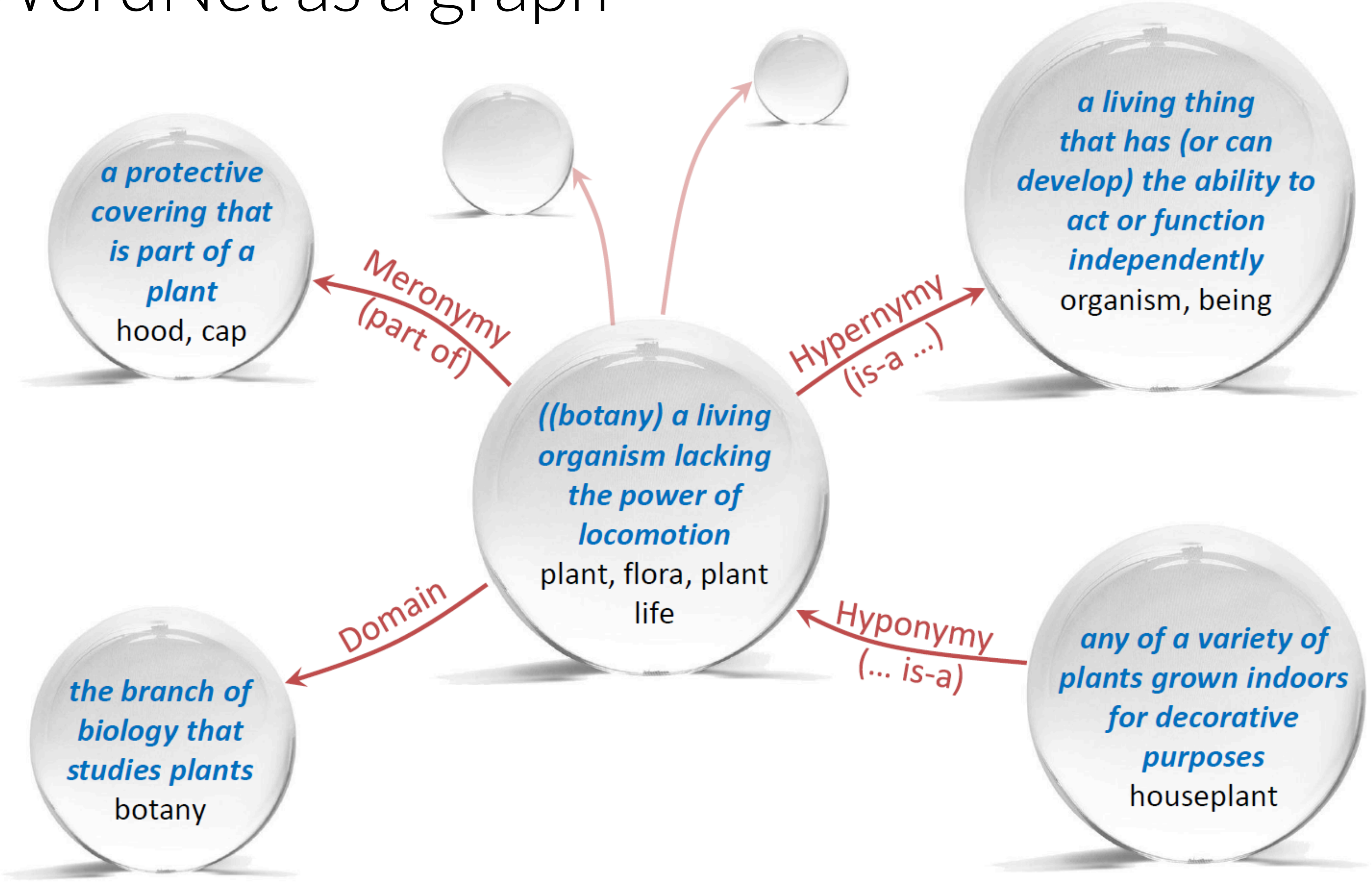
Sense Similarity Techniques

- 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



a protective covering that is part of a plant
hood, cap

Meronymy (part of)

a living thing that has (or can develop) the ability to act or function independently
organism, being

Hypernymy (is-a ...)

((botany) a living organism lacking the power of locomotion
plant, flora, plant life

Domain

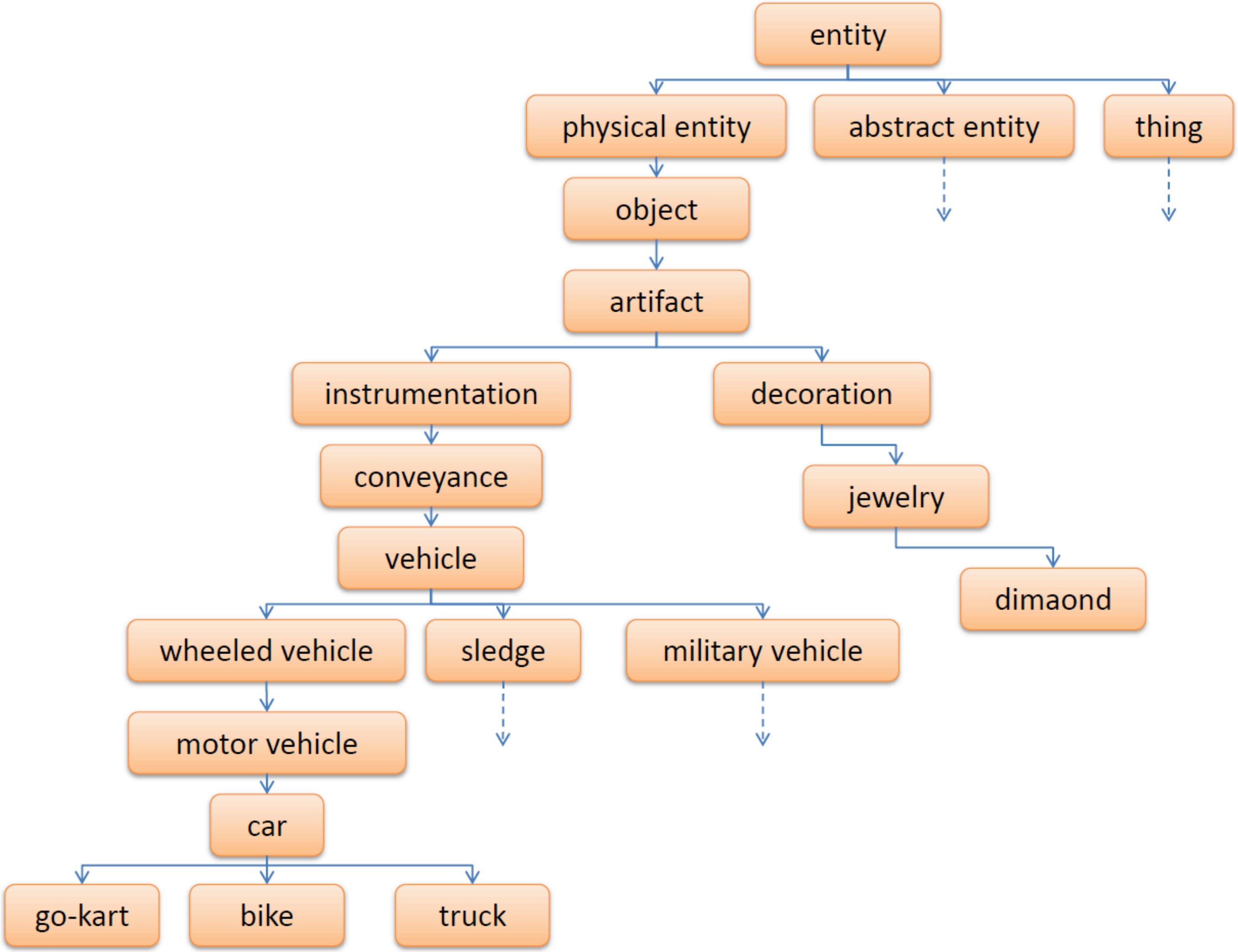
the branch of biology that studies plants
botany

Hyponymy (... is-a)

any of a variety of plants grown indoors for decorative purposes
houseplant

Sense Similarity Techniques

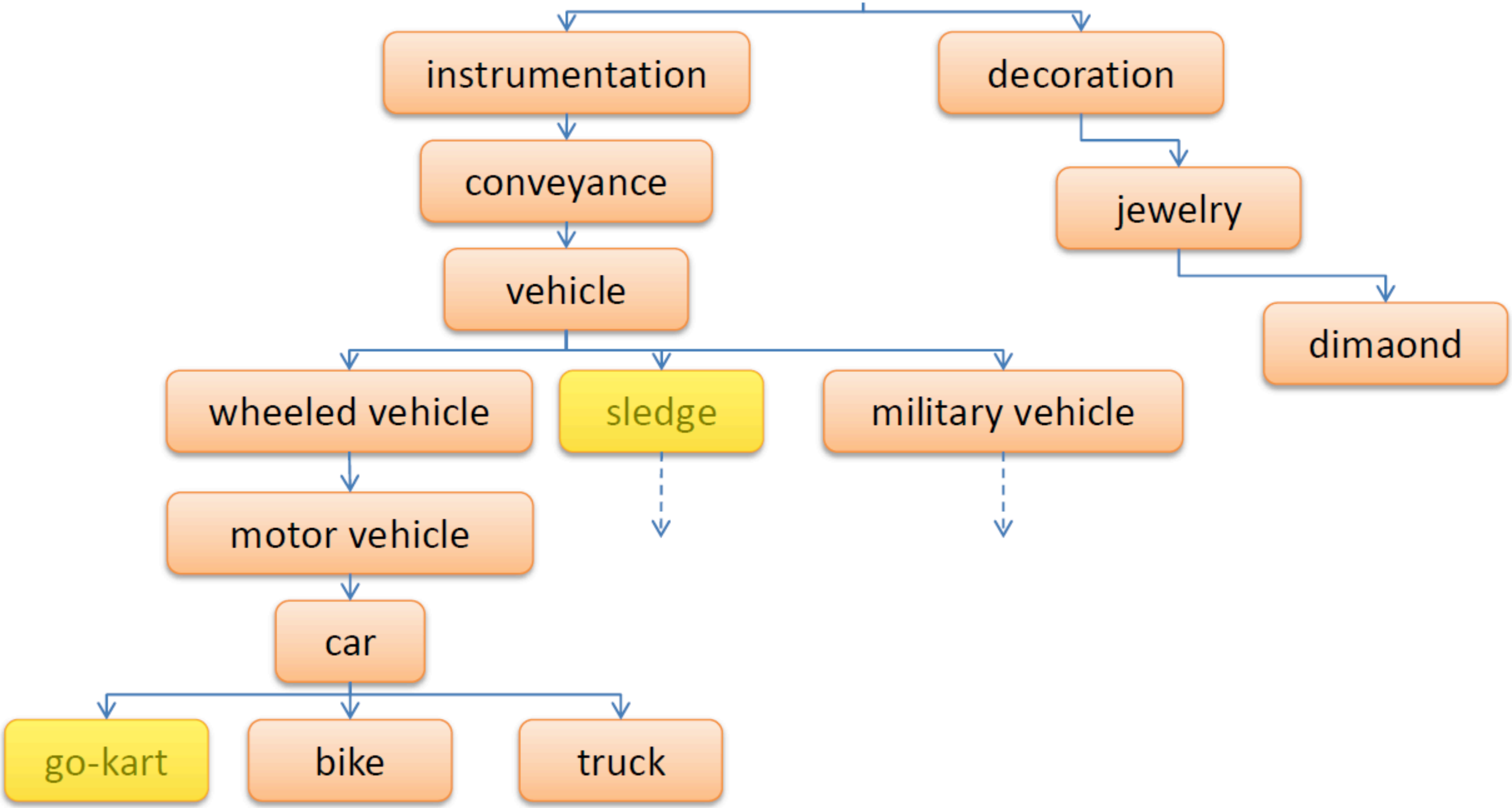
Tied to sense inventories: WordNet graph distance



Sense Similarity Techniques

Tied to sense inventories: WordNet graph distance

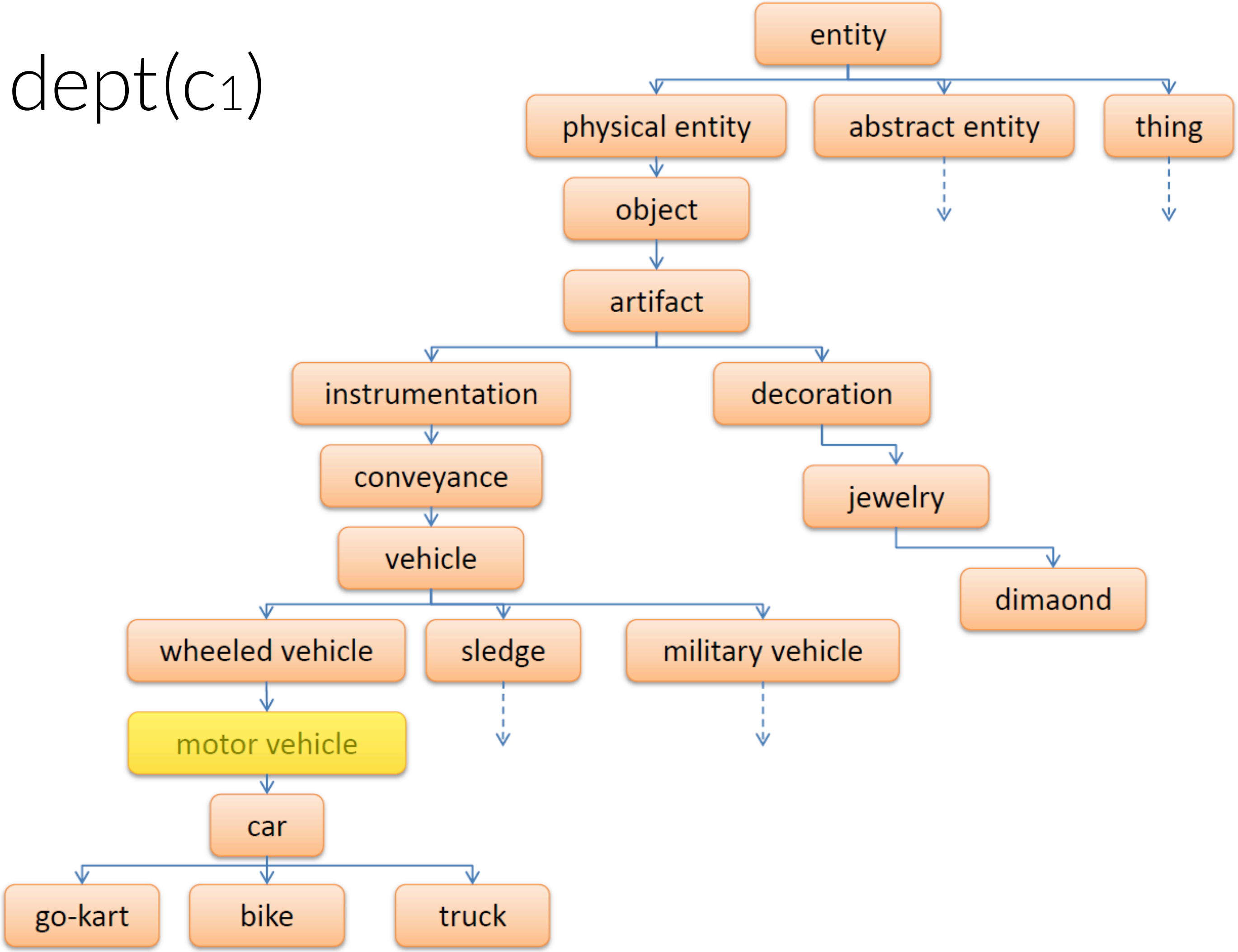
$\mathrm{len}(c_1,c_2)$



Sense Similarity Techniques

Tied to sense inventories: WordNet graph distance

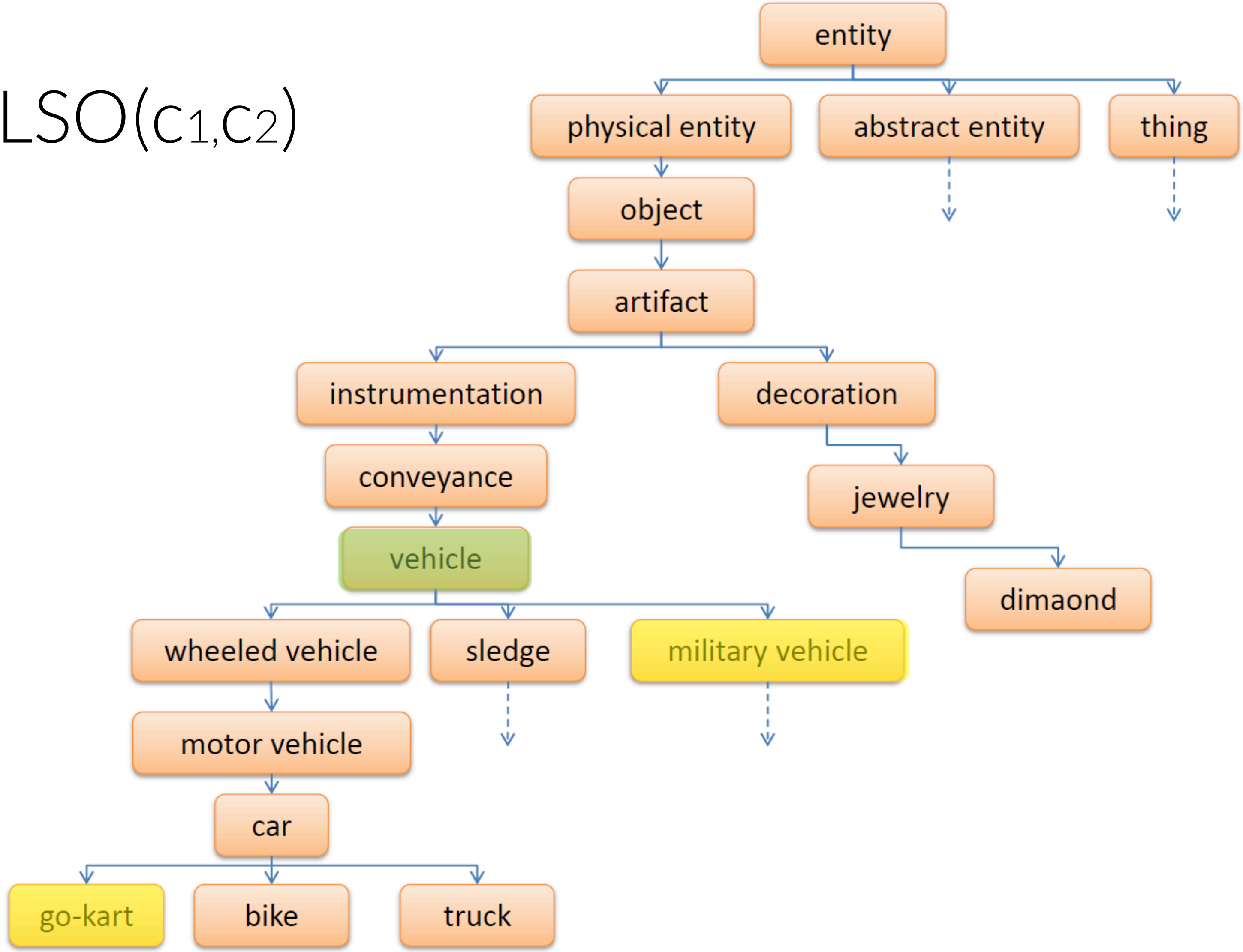
dept(c_1)



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)

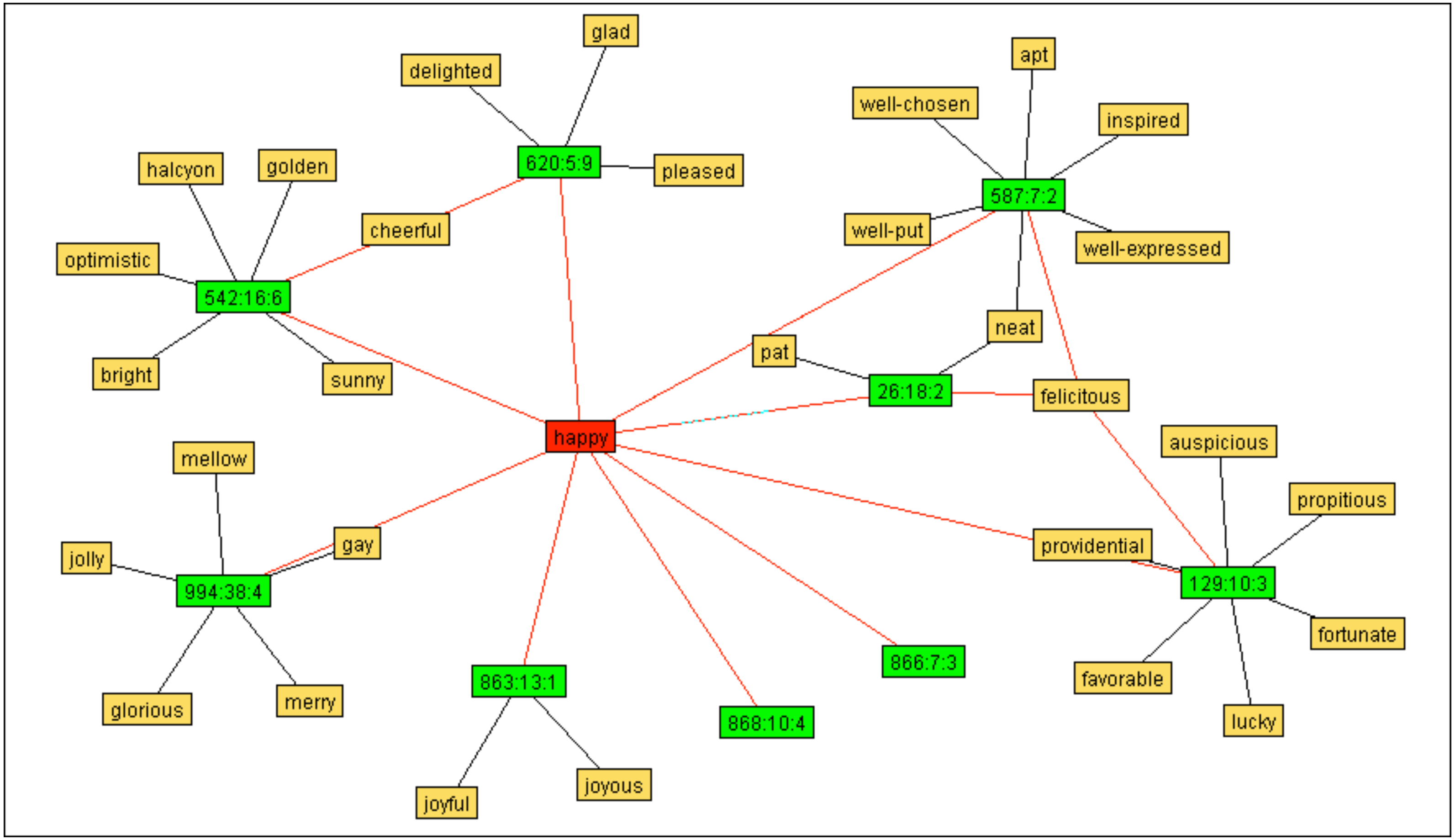
- Combined with statistics from corpora

  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

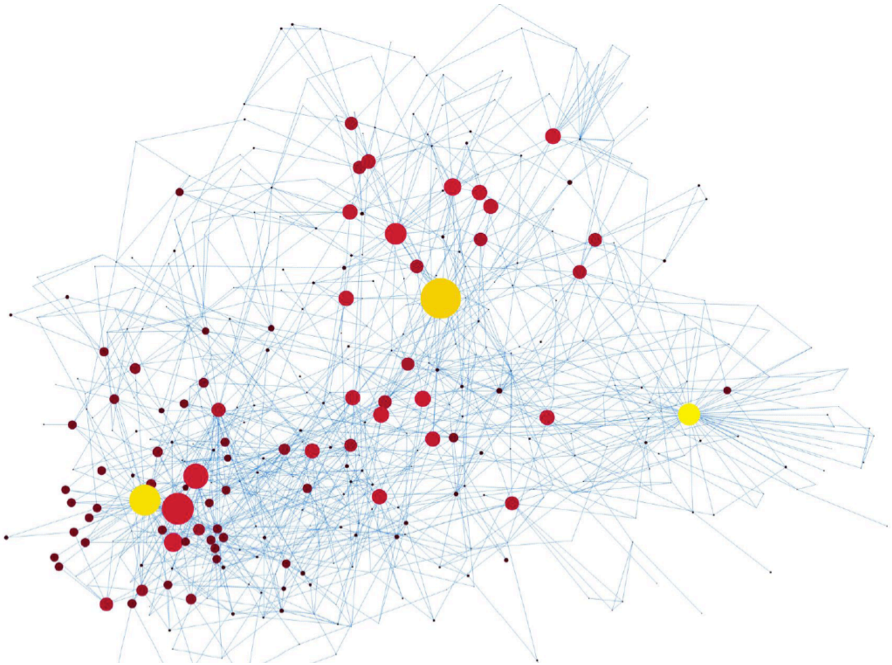
Tied to sense inventories: Explicit semantic representation

Random walks on semantic networks

The Personalized PageRank algorithm

Semantic similarity: Pilehvar et al (2013)

WSD: Agirre et al (CL 2014)



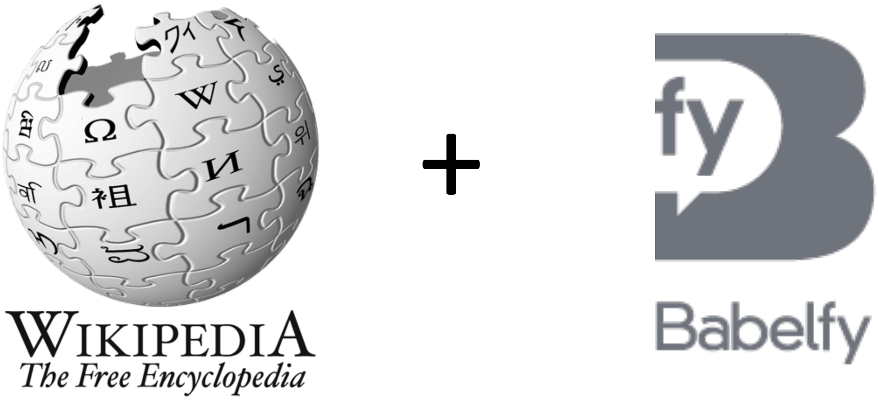
Sense Similarity Techniques

Tied to sense inventories: Explicit semantic representation

Distributional

SensEmbed - word2vec sense embeddings

Iacobacci et al (2015)

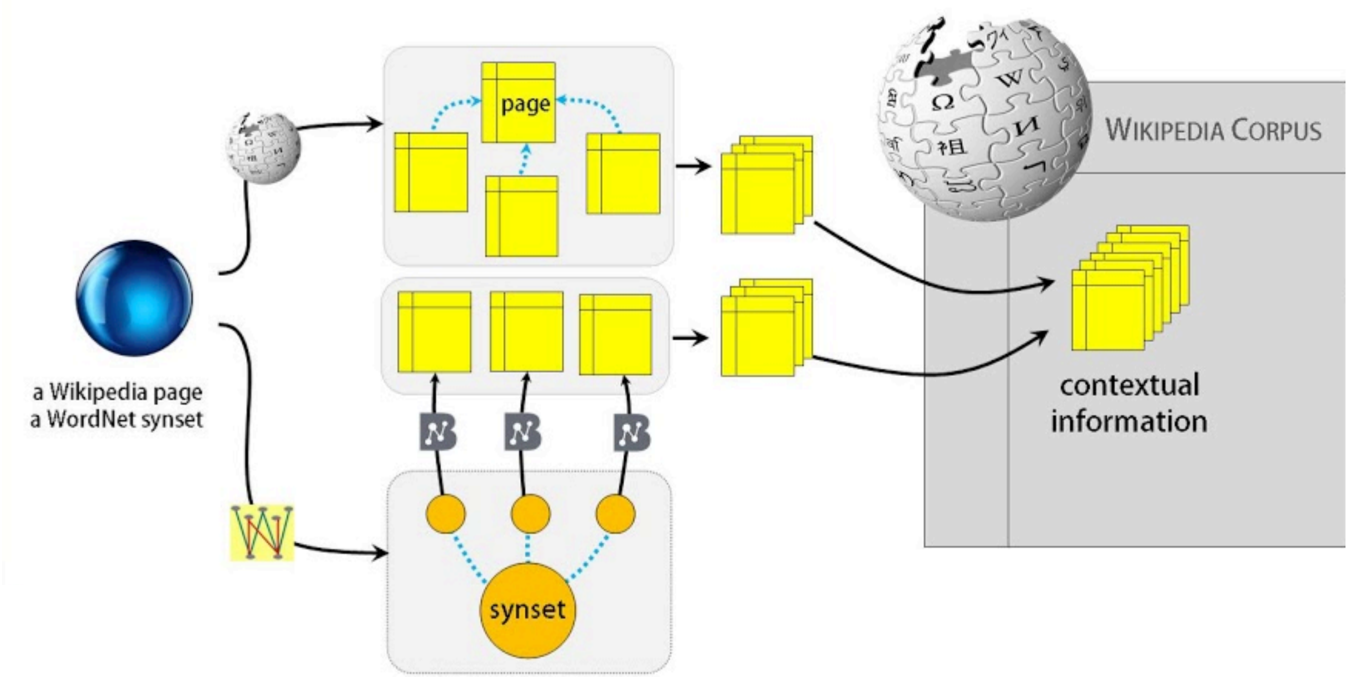


Sense Similarity Techniques

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

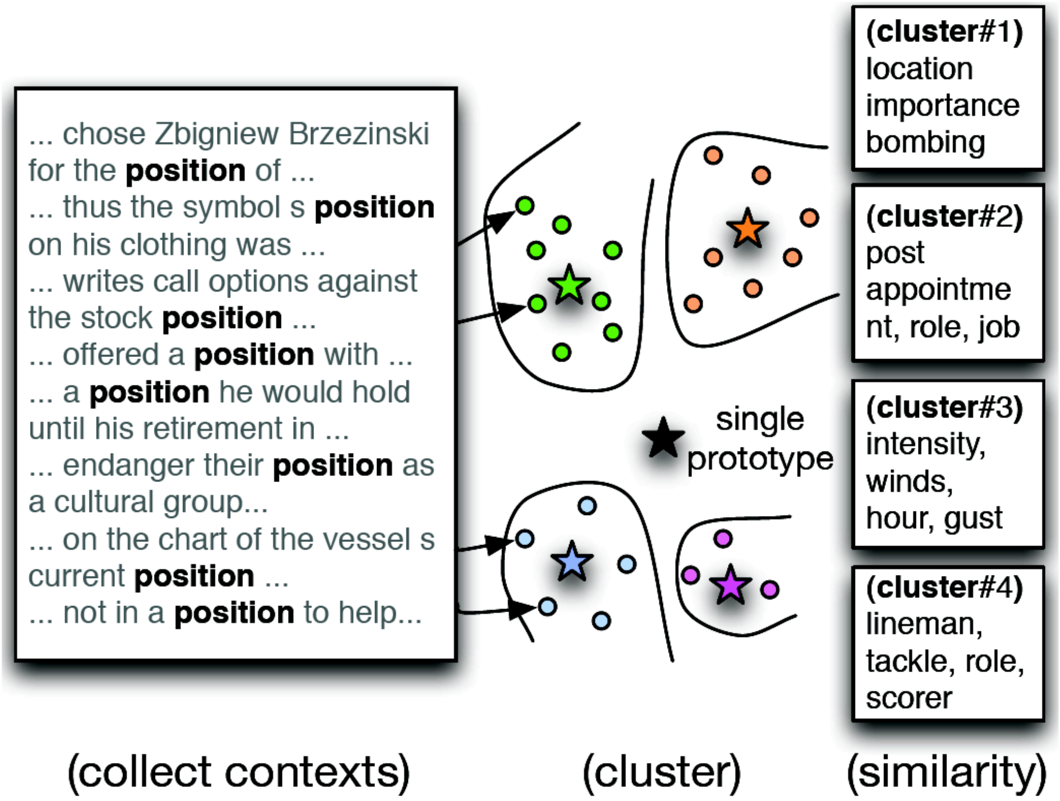
multi-prototype or **topic-based** representations

Usually based on clustering

Sense Similarity Techniques

Not Tied to sense inventories

Reisinger and Mooney (2010)



... chose Zbigniew Brzezinski for the **position** of ...
... thus the symbol s **position** on his clothing was ...
... writes call options against the stock **position** ...
... offered a **position** with ...
... a **position** he would hold until his retirement in ...
... endanger their **position** as a cultural group...
... on the chart of the vessel s current **position** ...
... not in a **position** to help...

single prototype

(cluster#1) location importance bombing

(cluster#2) post appointme nt, role, job

(cluster#3) intensity, winds, hour, gust

(cluster#4) lineman, tackle, role, scorer

(collect contexts) (cluster) (similarity)

Sense Similarity Techniques

Not Tied to sense inventories

Reisinger and Mooney (2010)

Measuring similarity - isolated words:

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

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

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

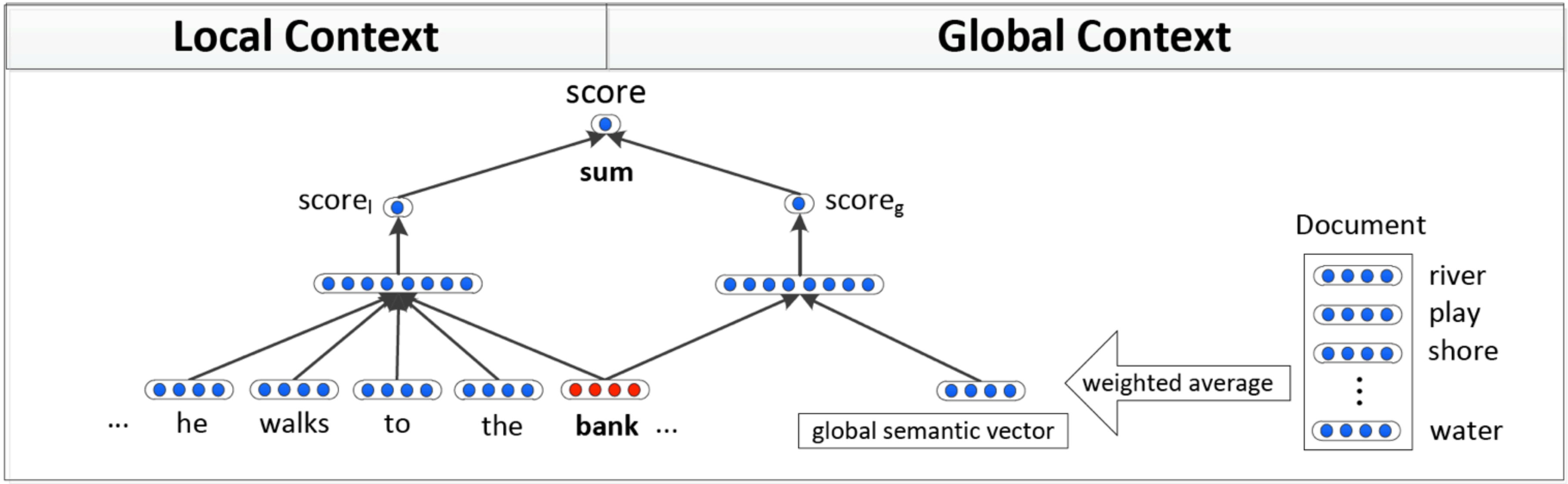
$$\text{MaxSimC}(w, w') \stackrel{\text{def}}{=} d(\hat{\pi}(w), \hat{\pi}(w'))$$

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

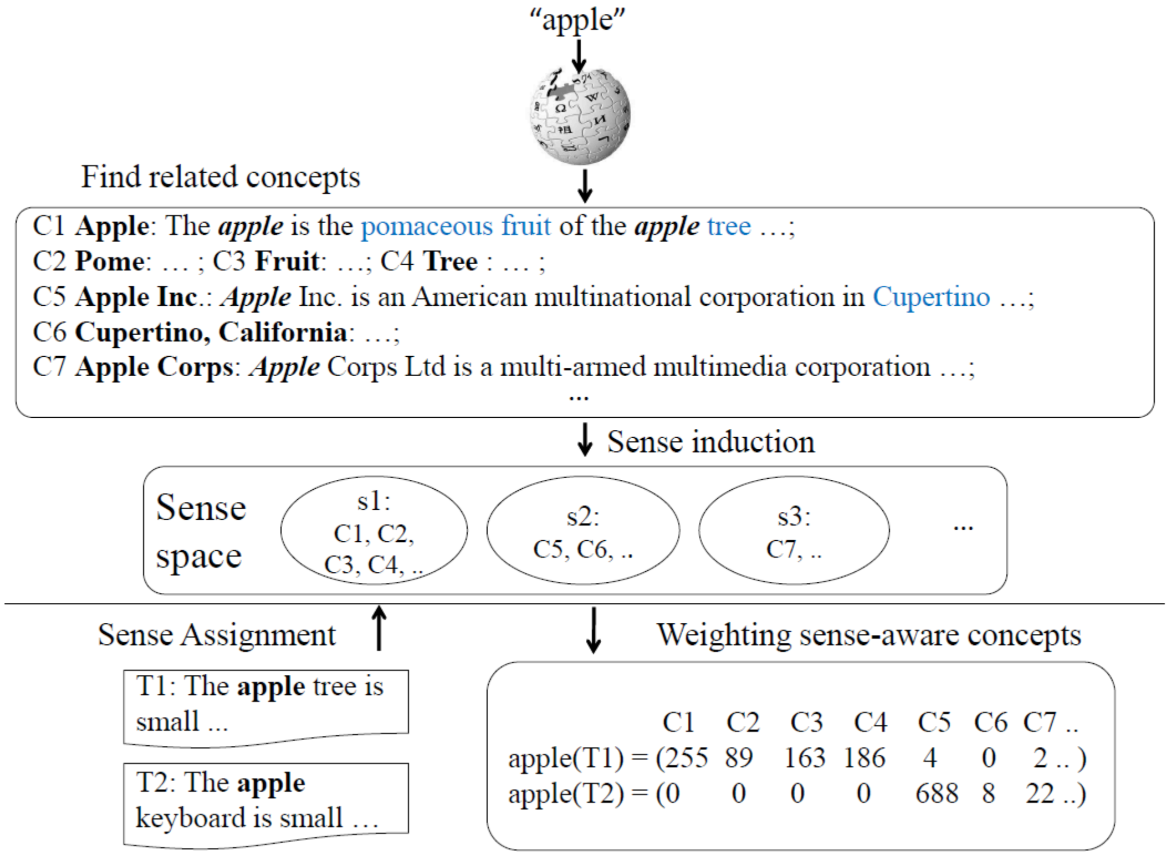
Online non-parametric clustering

Sense Similarity Techniques

Not Tied to sense inventories

SaSA - Sense-aware Semantic Analysis

Wu and Giles (AAAI 2015)



"apple"

Find related concepts

C1 **Apple**: The ***apple*** is the pomaceous fruit of the ***apple*** tree …;
C2 **Pome**: … ; C3 **Fruit**: …; C4 **Tree** : … ;
C5 **Apple Inc**.: ***Apple*** Inc. is an American multinational corporation in Cupertino …;
C6 **Cupertino, California**: …;
C7 **Apple Corps**: ***Apple*** Corps Ltd is a multi-armed multimedia corporation …;
…

Sense induction

Sense space

s1: C1, C2, C3, C4, ..

s2: C5, C6, ..

s3: C7, ..

…

Sense Assignment

T1: The **apple** tree is small …

T2: The **apple** keyboard is small …

Weighting sense-aware concepts

	C1	C2	C3	C4	C5	C6	C7 ..
apple(T1) = (255	89	163	186	4	0	2 ..)
apple(T2) = (0	0	0	0	688	8	22 ..)

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

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

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

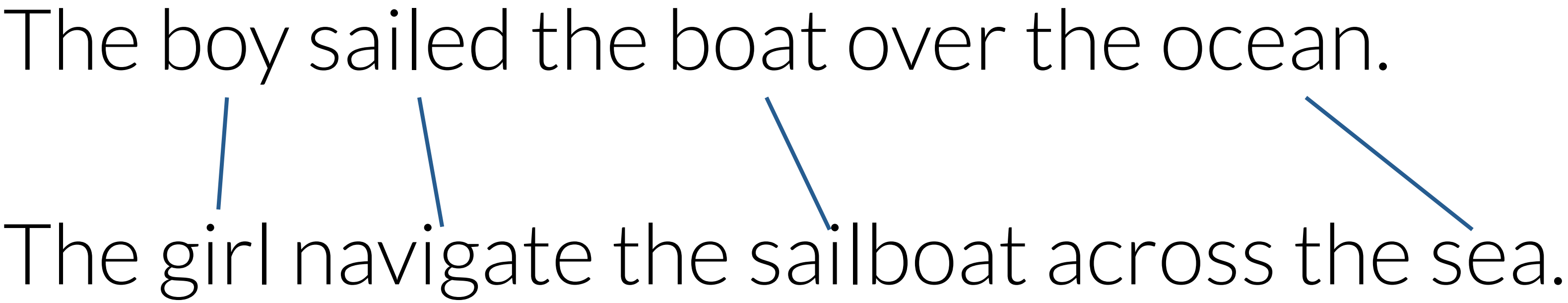
He went to the **ATM** to deposit the money.

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

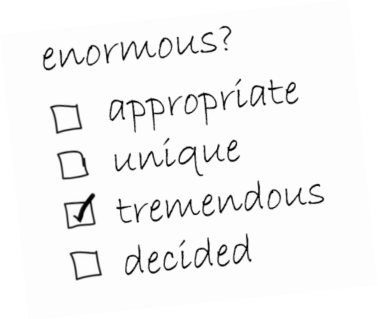
Numeric Word-Pair Similarity Tests

- 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

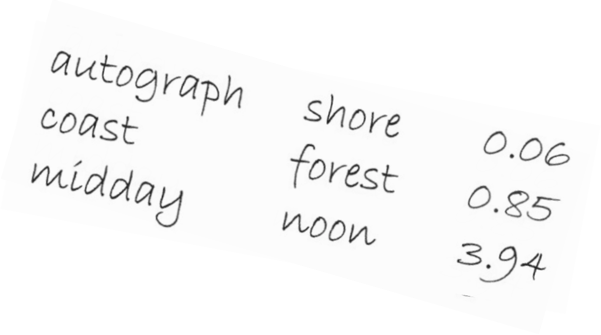
TOEFL Synonymy recognition



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)

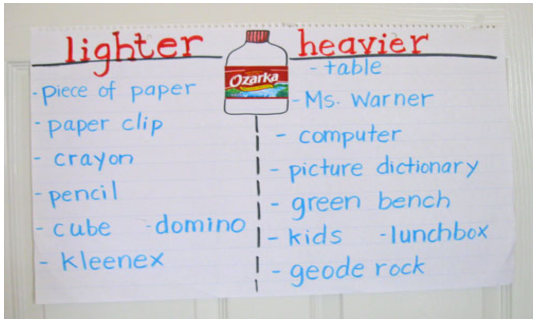
Learning semantic representations from text

1) Corpus

Learning semantic representations from text

1) Corpus

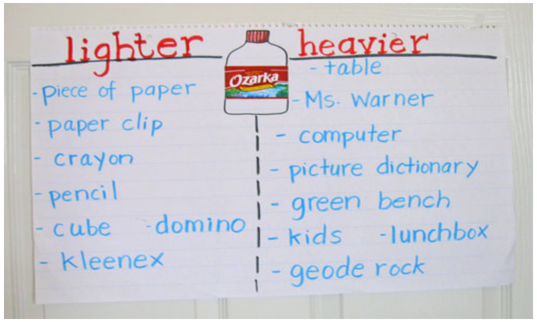
2) Preprocessing



Learning semantic representations from text

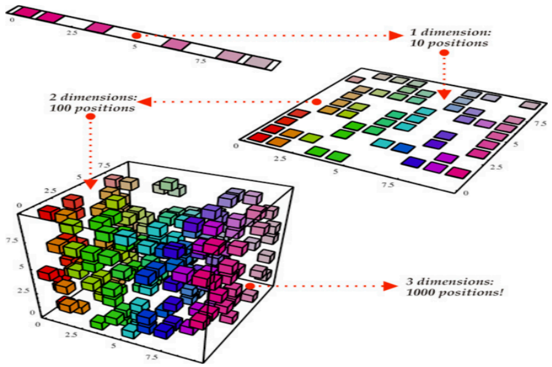
1) Corpus

2) Preprocessing



lighter	heavier
- Piece of paper	- table
- paper clip	- Ms. Warner
- crayon	- computer
- pencil	- picture dictionary
- cube -domino	- green bench
- kleenex	- kids -lunchbox
	- geode rock

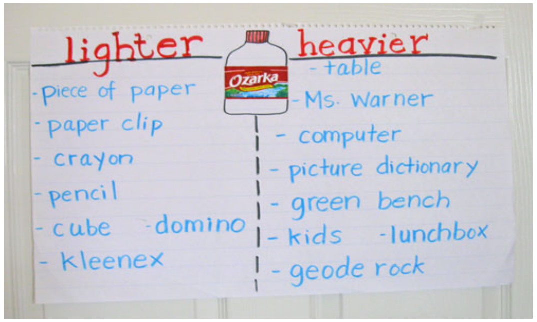
3) Dimensionality Reduction



Learning semantic representations from text

1) Corpus

2) Preprocessing



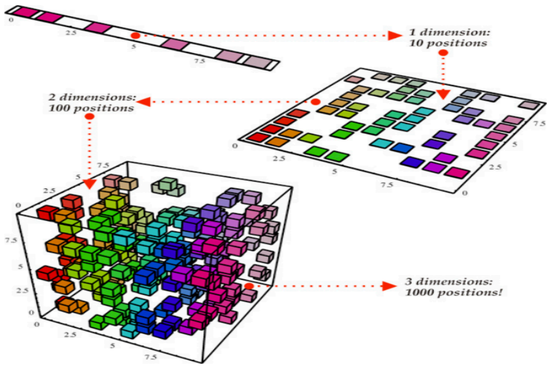
lighter | heavier

- piece of paper
- paper clip
- crayon
- pencil
- cube
- domino
- kleenex

Ozarka

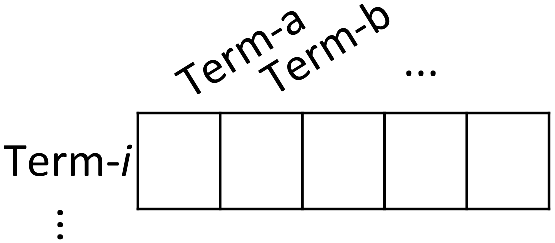
- table
- Ms. Warner
- computer
- picture dictionary
- green bench
- kids
- lunchbox
- geode rock

3) Dimensionality Reduction



4) Post Processing

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

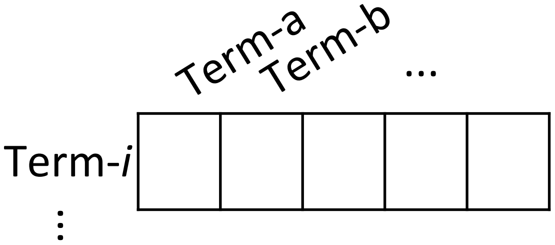


Cells record the number of times...

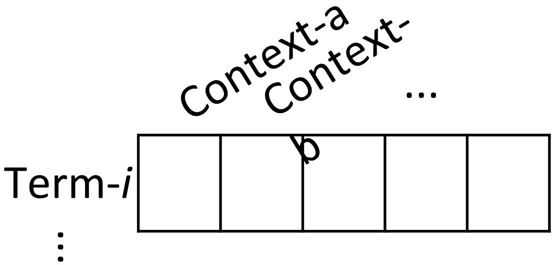
term *j* occurs in the context window of term *i*.

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*.

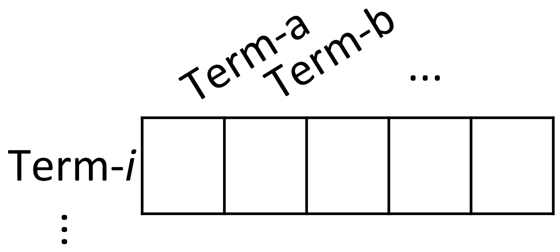


term *i* occurs in a context window

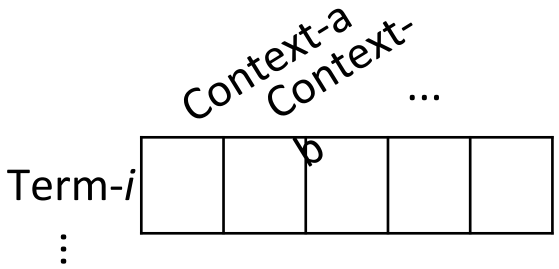
- w_{-2}, w_{-1}, w, w_1, w_2
- or analogously, with dependencies

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

Cells record the number of times...

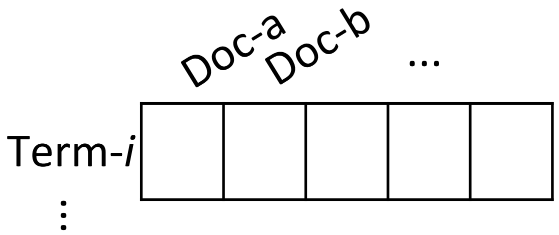


term *j* occurs in the context window of term *i*.



term *i* occurs in a context window

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

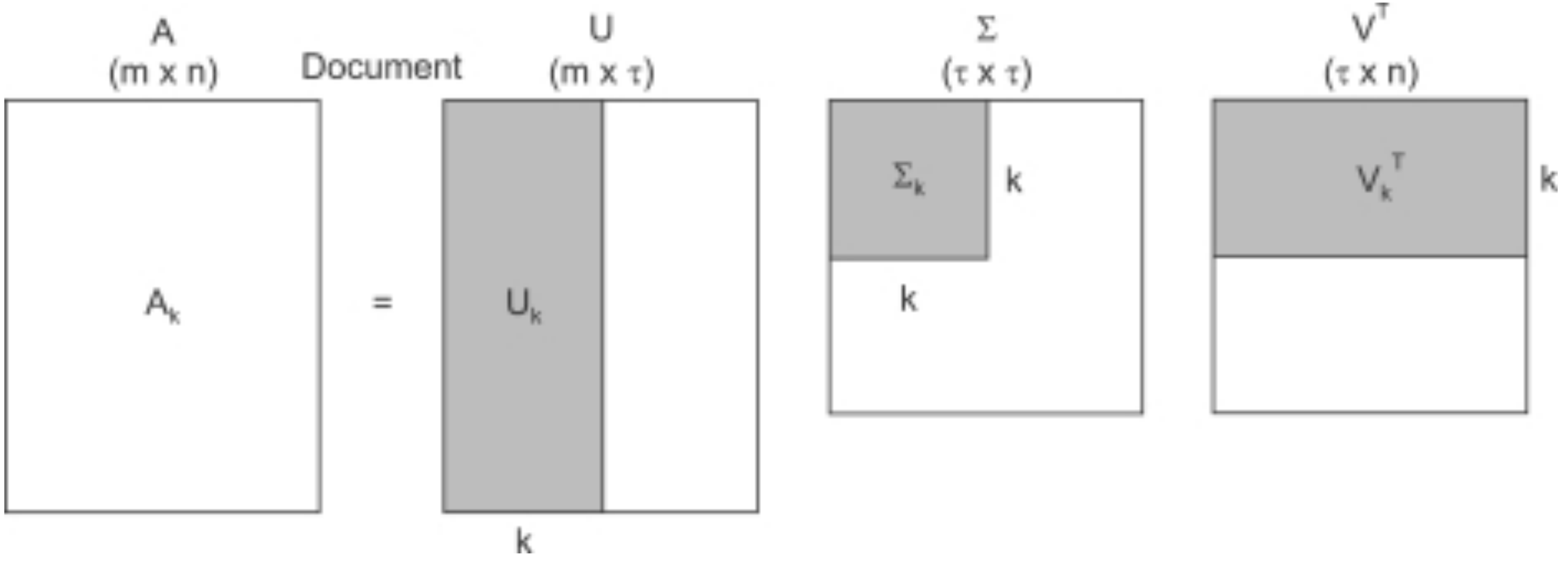


term *i* occurs in document *j*.

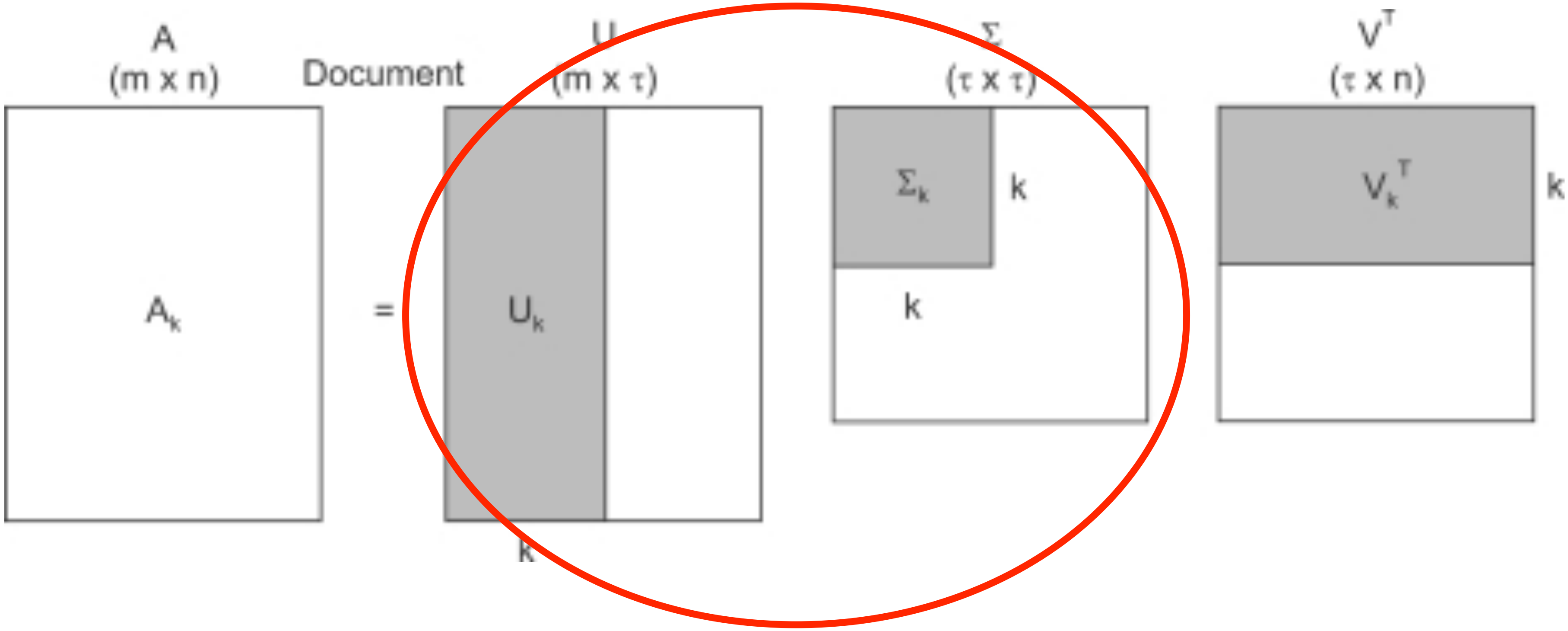
Raw word co-occurrence is rarely satisfactory as a representation

- 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)



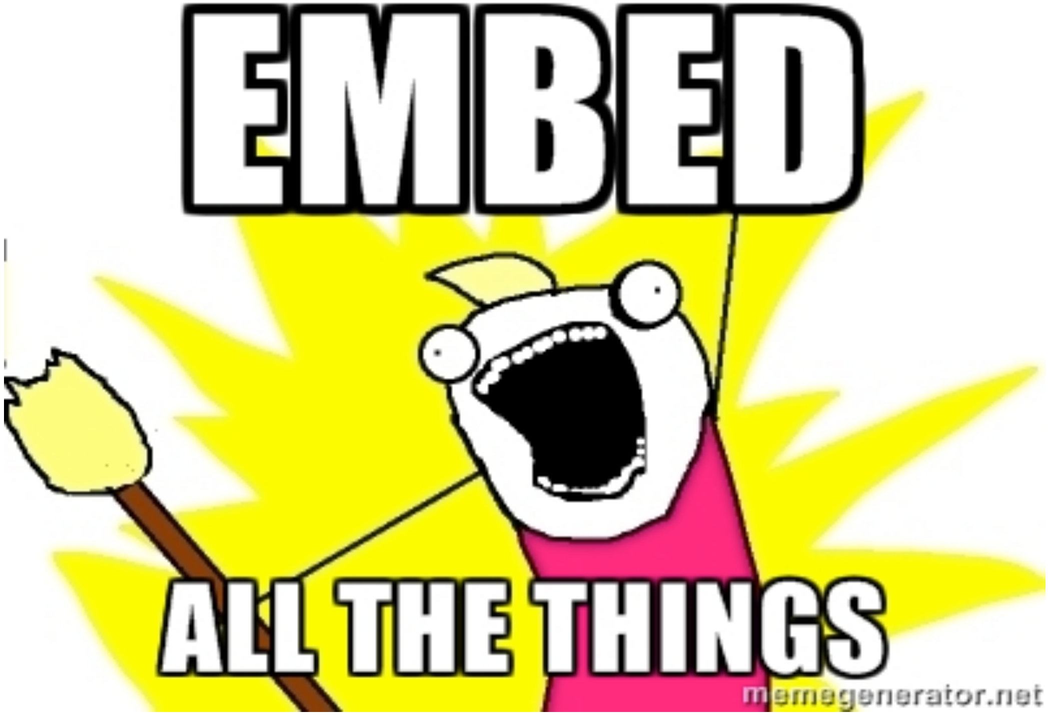
Standard Approach: Reduce the dimensionality using the Singular Value Decomposition (SVD)



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

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

also known as



word2vec

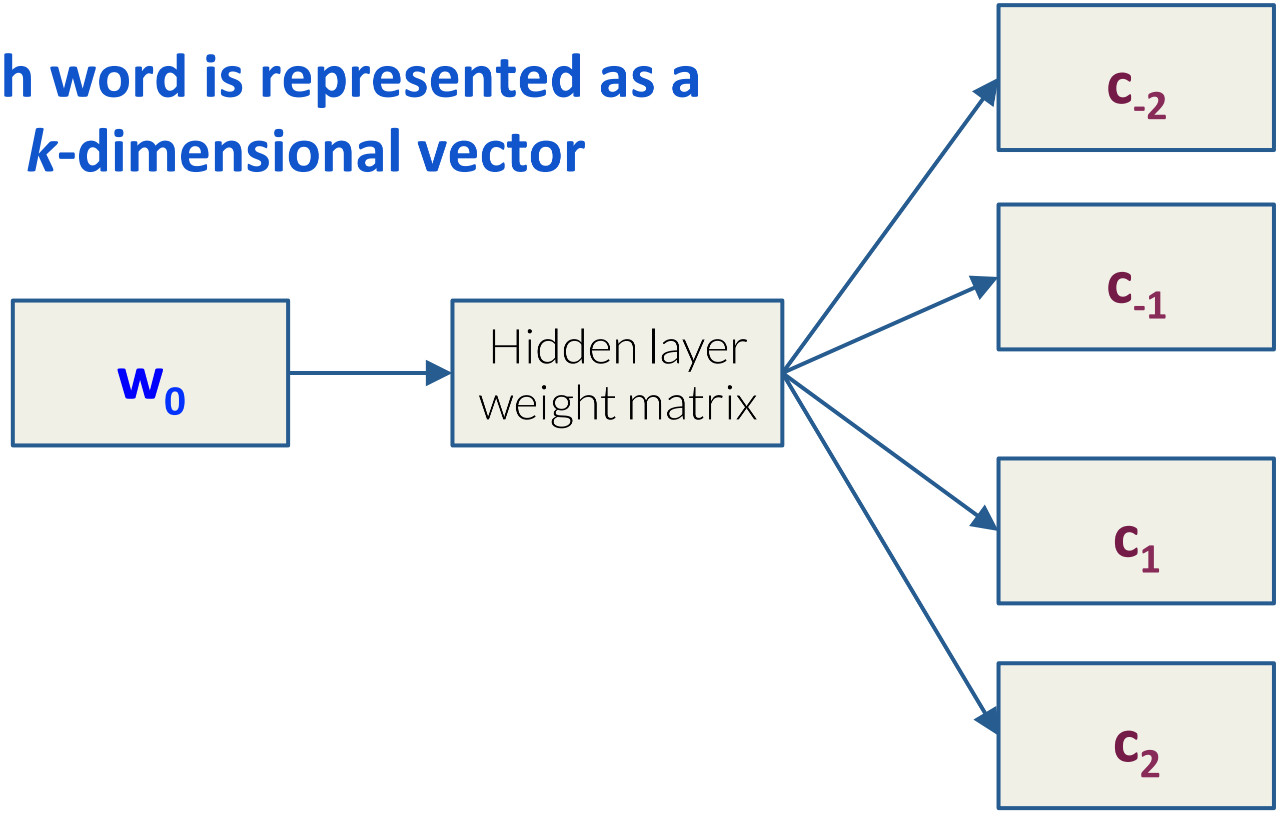
More a software system than an algorithm

- 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

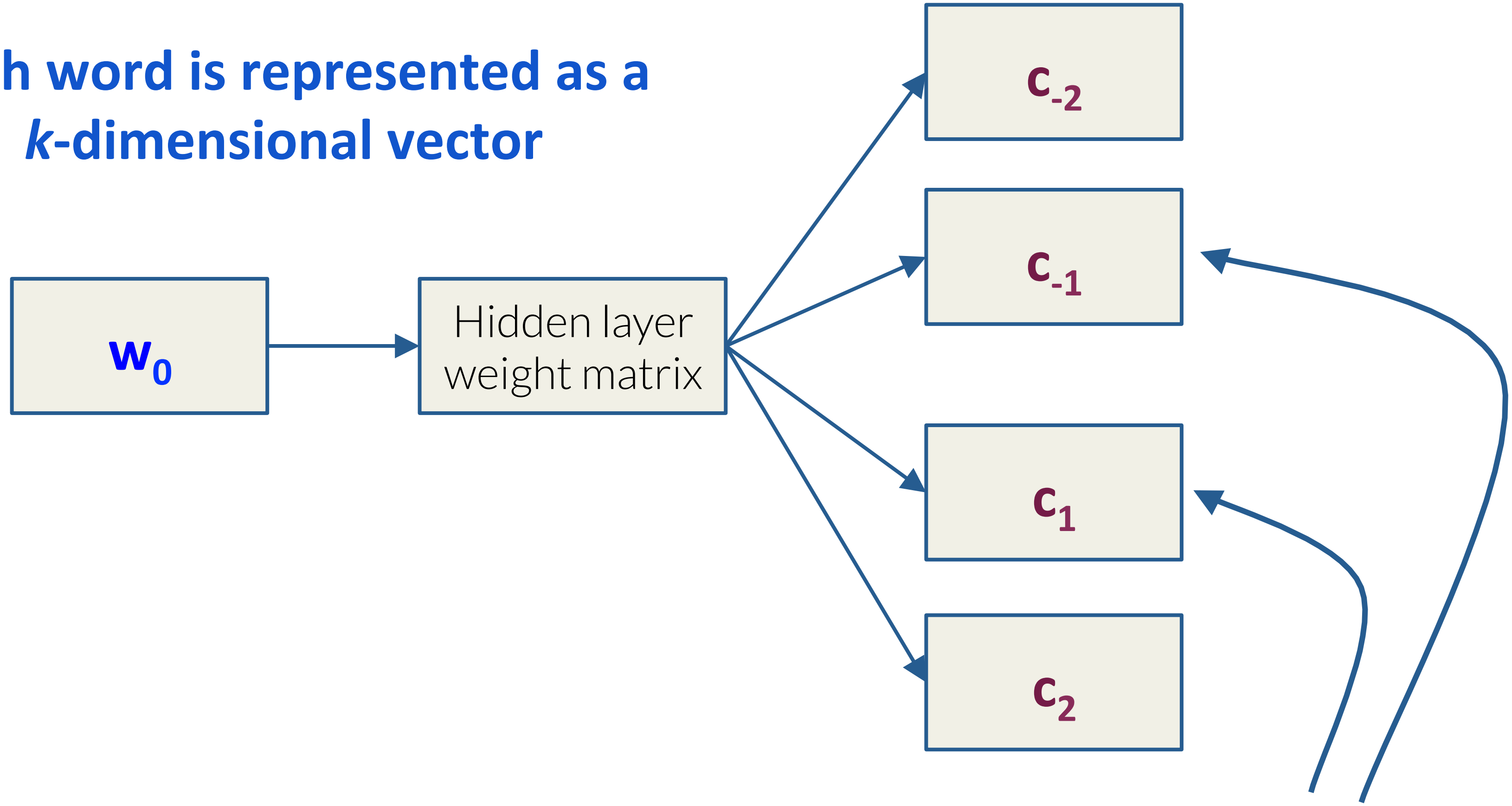
Each word is represented as a *k*-dimensional vector



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

word2vec: a neural look

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



w_0

Hidden layer weight matrix

c_{-2}

c_{-1}

c_1

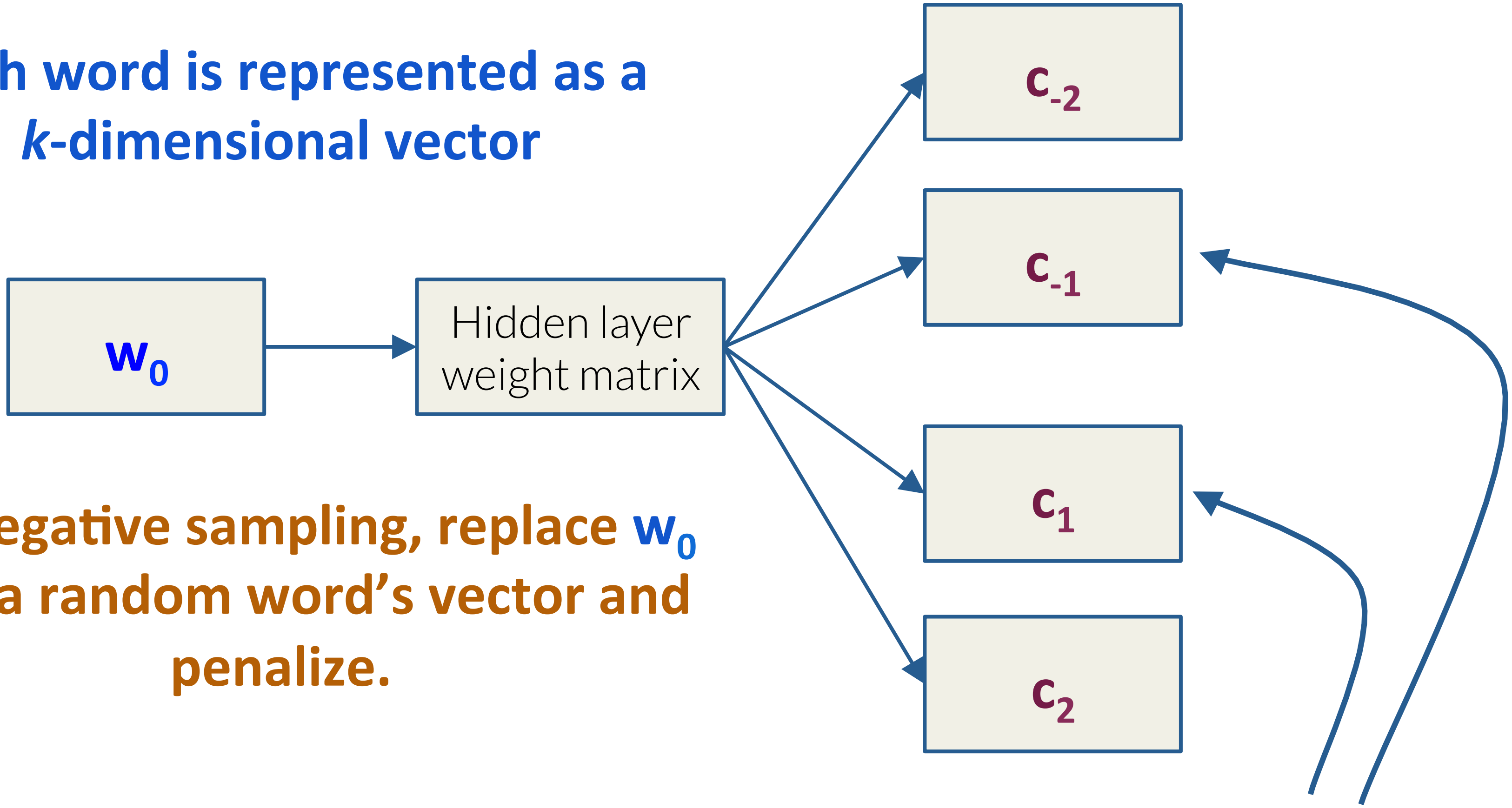
c_2

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)

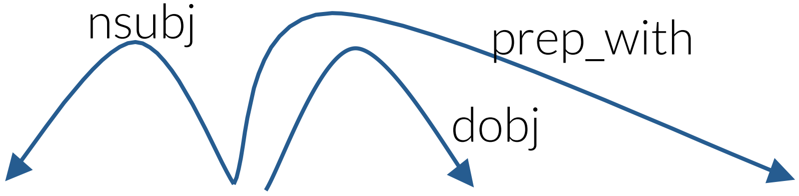
Could we get better performance with syntactic contexts?

Australian scientist discovers star with telescope

(Levy and Goldberg, 2014)

Could we get better performance with syntactic contexts?

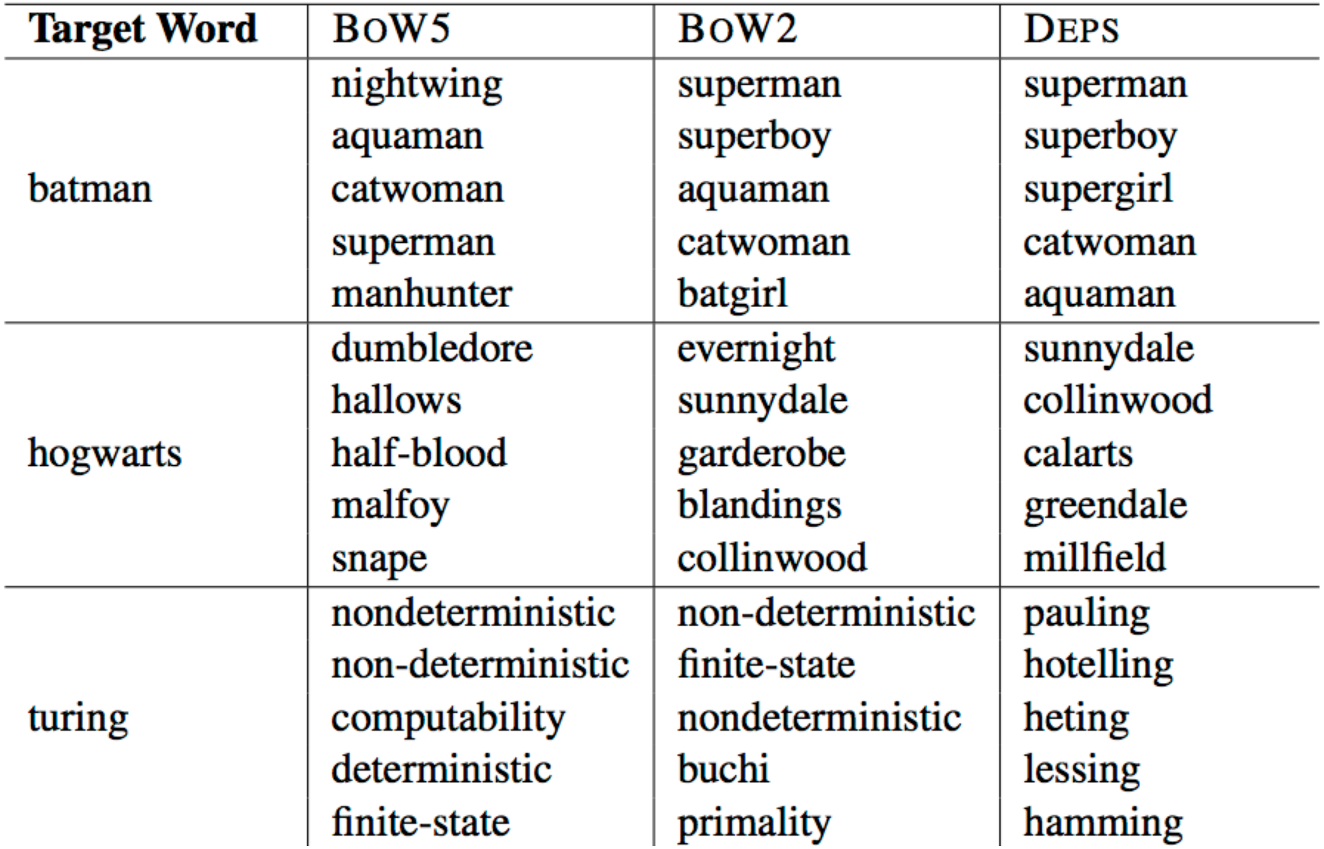
Australian scientist discovers star with telescope



Australian scientist discovers star with telescope

(Levy and Goldberg, 2014)

Dependency-based embeddings capture functional information



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

No quantitative results on standard benchmarks

(Levy and Goldberg, 2014)

Glove: capture the ratio of co-occurrence probabilities

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

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

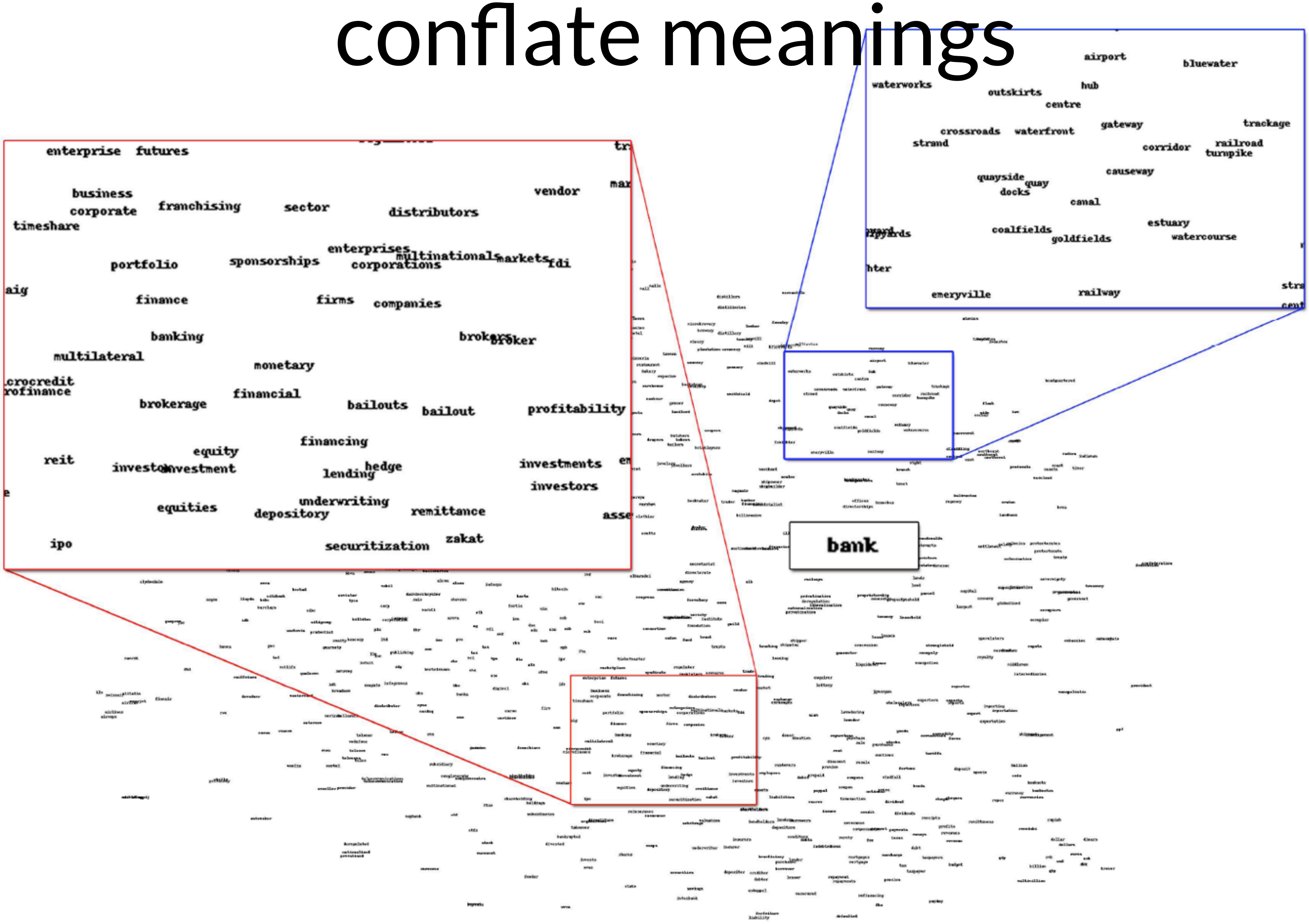
(Pennington et al., 2014)

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

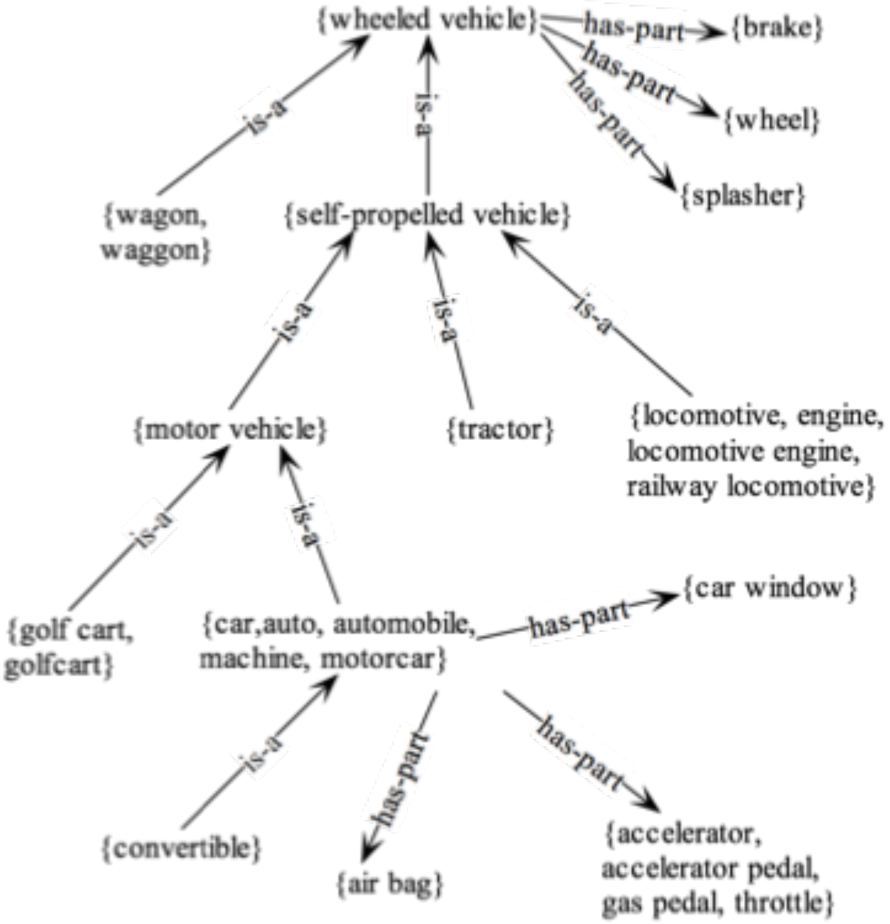
(Li and Jurafsky, 2015)

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?

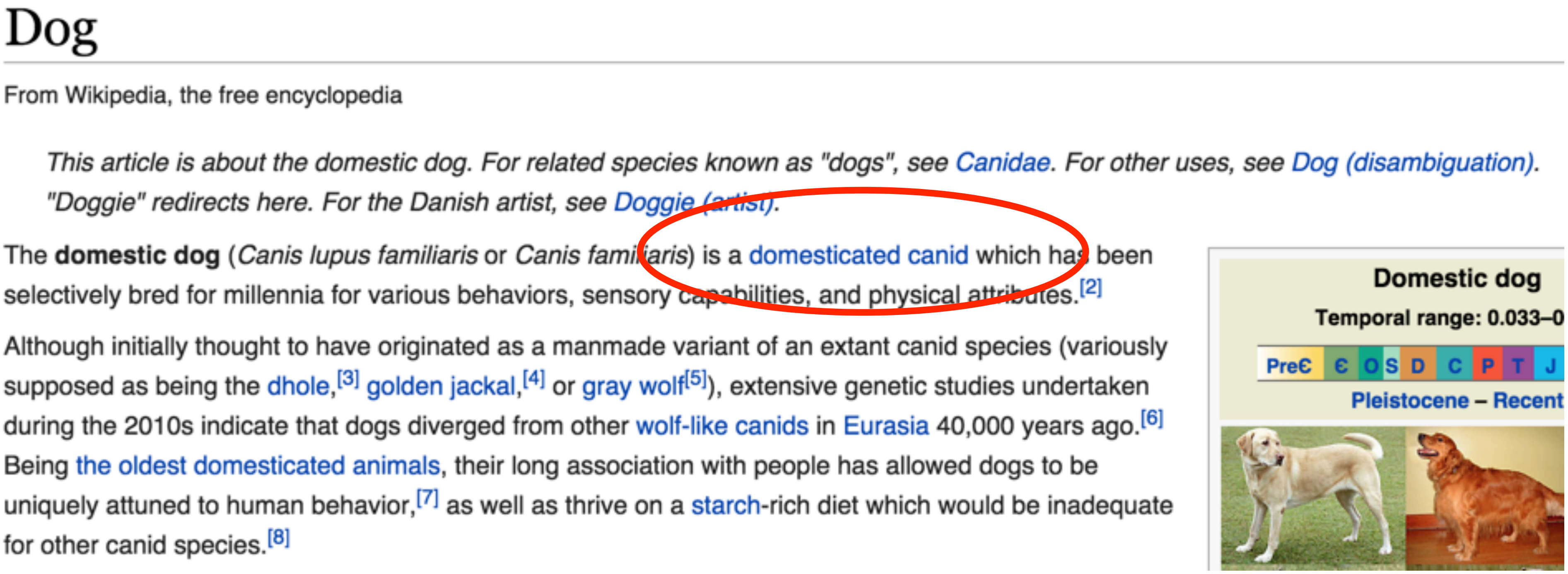
The structure of WordNet, Wikipedia, and other knowledge bases can be used to measure word similarity



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]

Domestic dog

Temporal range: 0.033–0

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

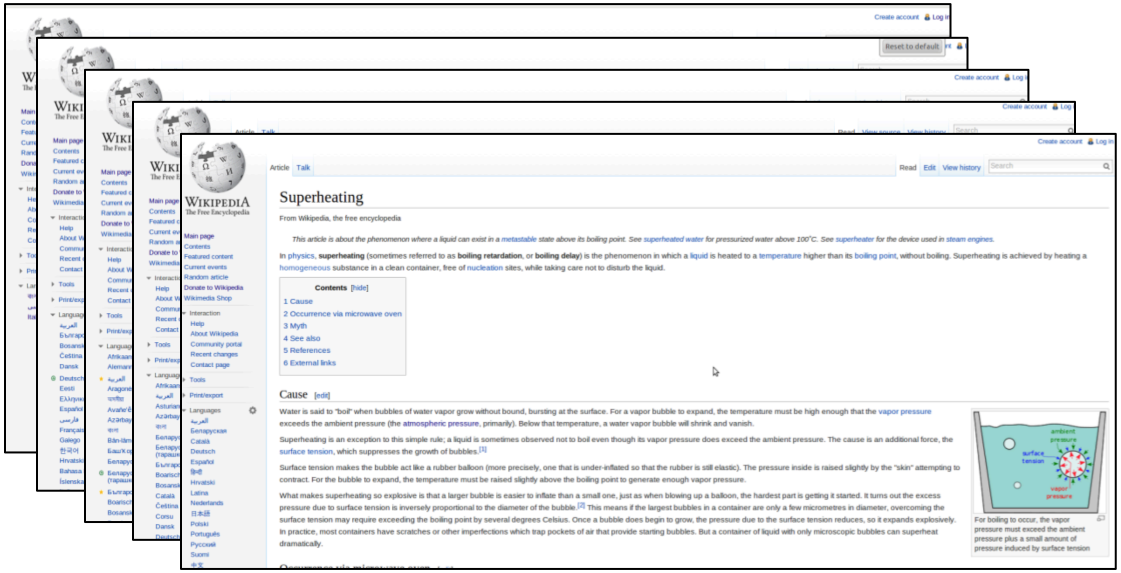
- Run PPR and compare vectors

	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



Superheating

From Wikipedia, the free encyclopedia

This article is about the phenomenon where a liquid can exist in a metastable state above its boiling point. See superheated water for pressurized water above 100°C. See superheater for the device used in steam engines.

In physics, **superheating** (sometimes referred to as **boiling retardation**, or **boiling delay**) is the phenomenon in which a liquid is heated to a temperature higher than its boiling point, without boiling. Superheating is achieved by heating a homogeneous substance in a clean container, free of nucleation sites, while taking care not to disturb the liquid.

Contents [hide]

1 Cause
2 Occurrence via microwave oven
3 Myth
4 See also
5 References
6 External links

Cause [edit]

Water is said to "boil" when bubbles of water vapor grow without bound, bursting at the surface. For a vapor bubble to expand, the temperature must be high enough that the vapor pressure exceeds the ambient pressure (the atmospheric pressure, primarily). Below that temperature, a water vapor bubble will shrink and vanish.

Superheating is an exception to this simple rule; a liquid is sometimes observed not to boil even though its vapor pressure does exceed the ambient pressure. The cause is an additional force, the surface tension, which suppresses the growth of bubbles.[1]

Surface tension makes the bubble act like a rubber balloon (more precisely, one that is under-inflated so that the rubber is still elastic). The pressure inside is raised slightly by the "skin" attempting to contract. For the bubble to expand, the temperature must be raised slightly above the boiling point to generate enough vapor pressure.

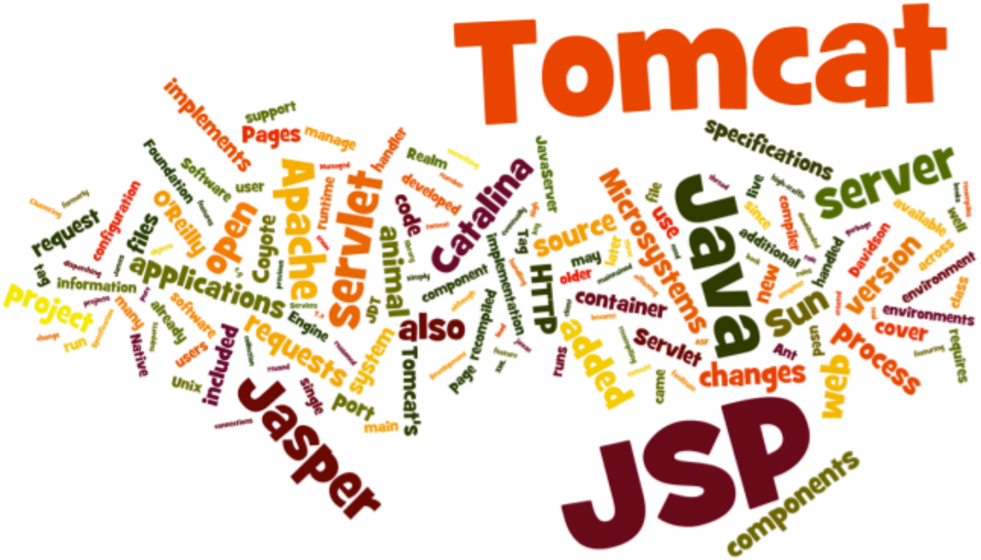
What makes superheating so explosive is that a larger bubble is easier to inflate than a small one, just as when blowing up a balloon, the hardest part is getting it started. It turns out the excess pressure due to surface tension is inversely proportional to the diameter of the bubble.[2] This means if the largest bubbles in a container are only a few micrometres in diameter, overcoming the surface tension may require exceeding the boiling point by several degrees Celsius. Once a bubble does begin to grow, the pressure due to the surface tension reduces, so it expands explosively. In practice, most containers have scratches or other imperfections which trap pockets of air that provide starting bubbles. But a container of liquid with only microscopic bubbles can superheat dramatically.

For boiling to occur, the vapor pressure must exceed the ambient pressure plus a small amount of pressure induced by surface tension

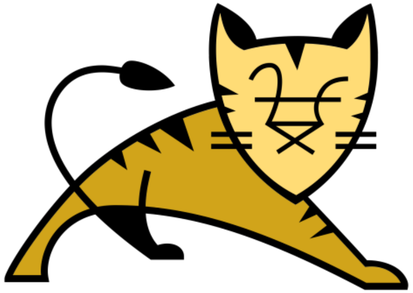
$\{c_1, c_2, c_3, \ldots c_N\}$ where N is the number of articles in Wikipedia

Gabrilovich and Markovitch (2007)

WIKIPEDIA *The Free Encyclopedia* articles for Tomcat



Tomcat
F-14
aircraft
Navy
Grumman
Air
U.S
also
Phoenix
missiles
Iranian
two
first
combat
F-14A
Iran
missile
system
upgraded
fighter
flight
Super
radar
engines
F-14D
F-14s
engine
avionics
Tomcats
reconnaissance
TARPS
LANTIRN
AIM-54



Explicit Semantic Analysis (ESA)

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

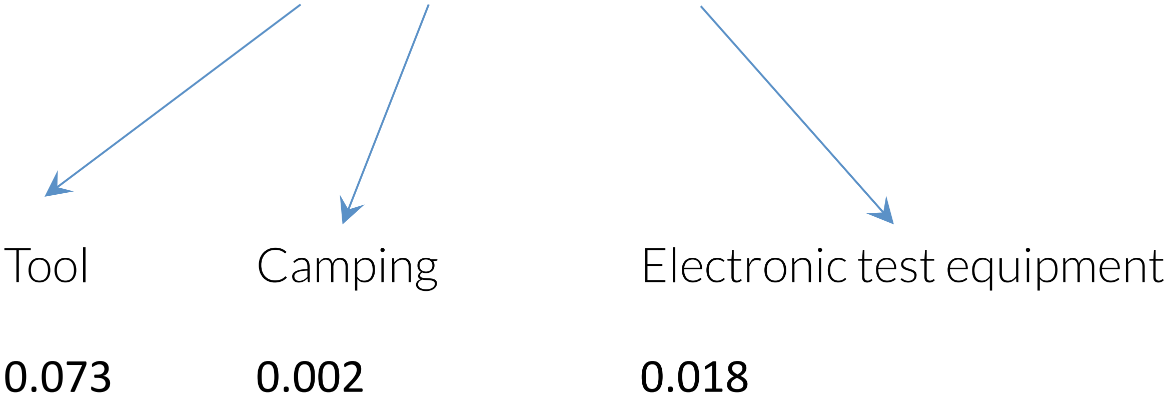
$$T[i,j] = tf(t_i, d_j) \cdot \log \frac{n}{df_i},$$

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

Explicit Semantic Analysis (ESA)

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

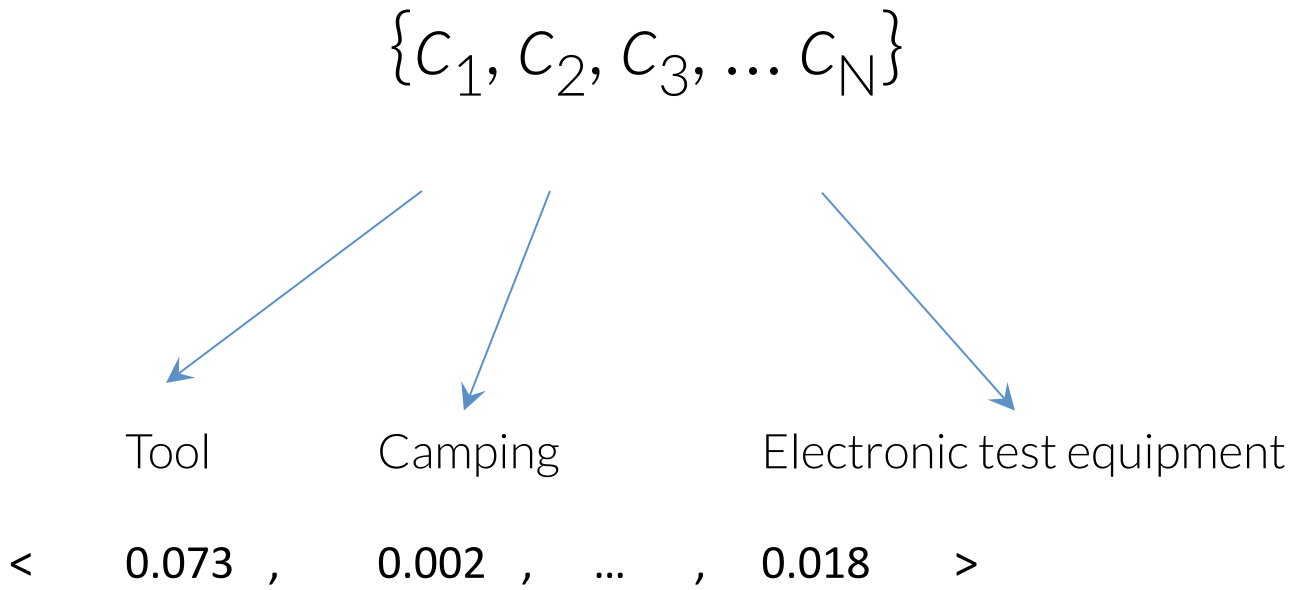
$$\{c_1, c_2, c_3, \ldots c_N\}$$



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:



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

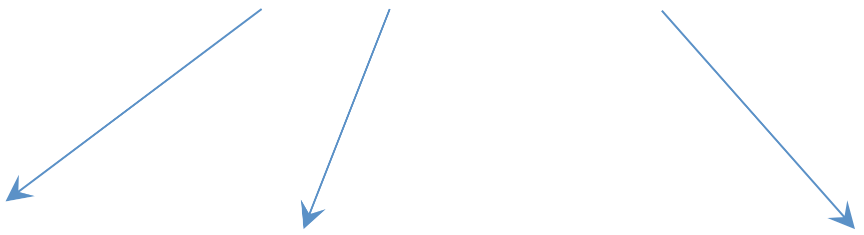
Tool Camping Electronic test equipment

< 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:

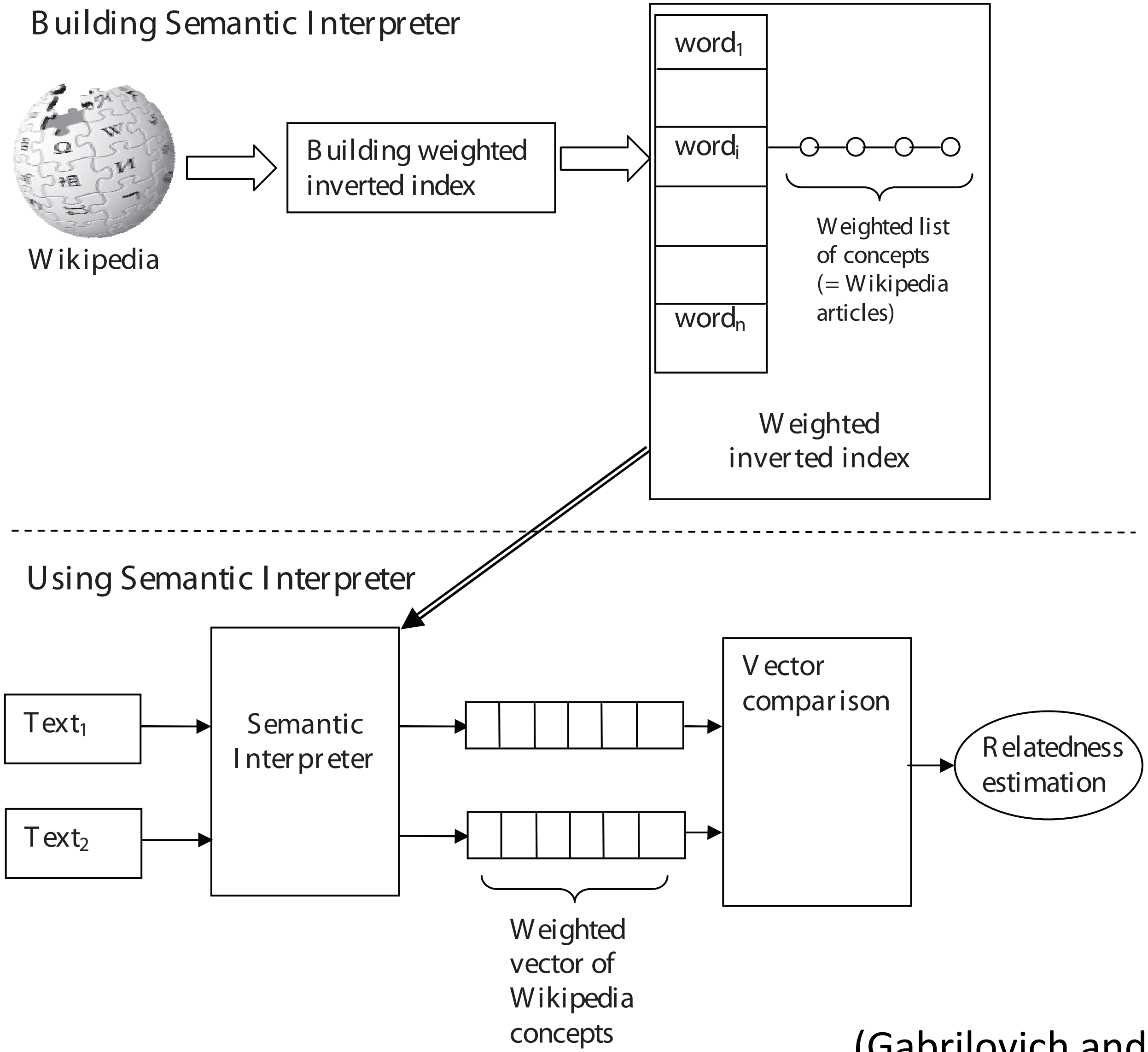
$$\{c_1, c_2, c_3, \ldots c_N\}$$



Tool Camping Electronic test equipment

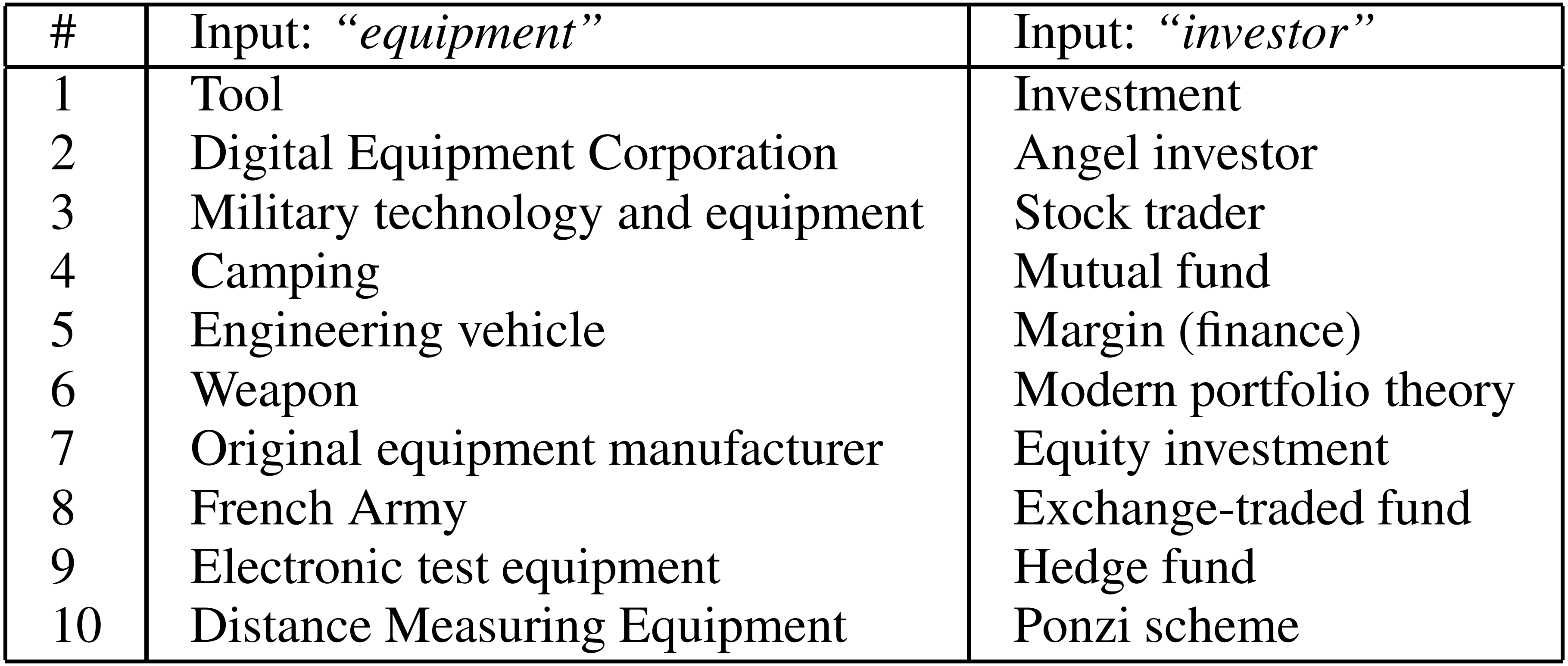
< 0.073 , 0.002 , ... , 0.018 > **Vector for equipment**

ESA pipeline


Building Semantic Interpreter
Wikipedia
Building weighted inverted index
word$_1$
word$_i$
word$_n$
Weighted list of concepts (= Wikipedia articles)
Weighted inverted index
Using Semantic Interpreter
Text$_1$
Text$_2$
Semantic Interpreter
Vector comparison
Relatedness estimation
Weighted vector of Wikipedia concepts

(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

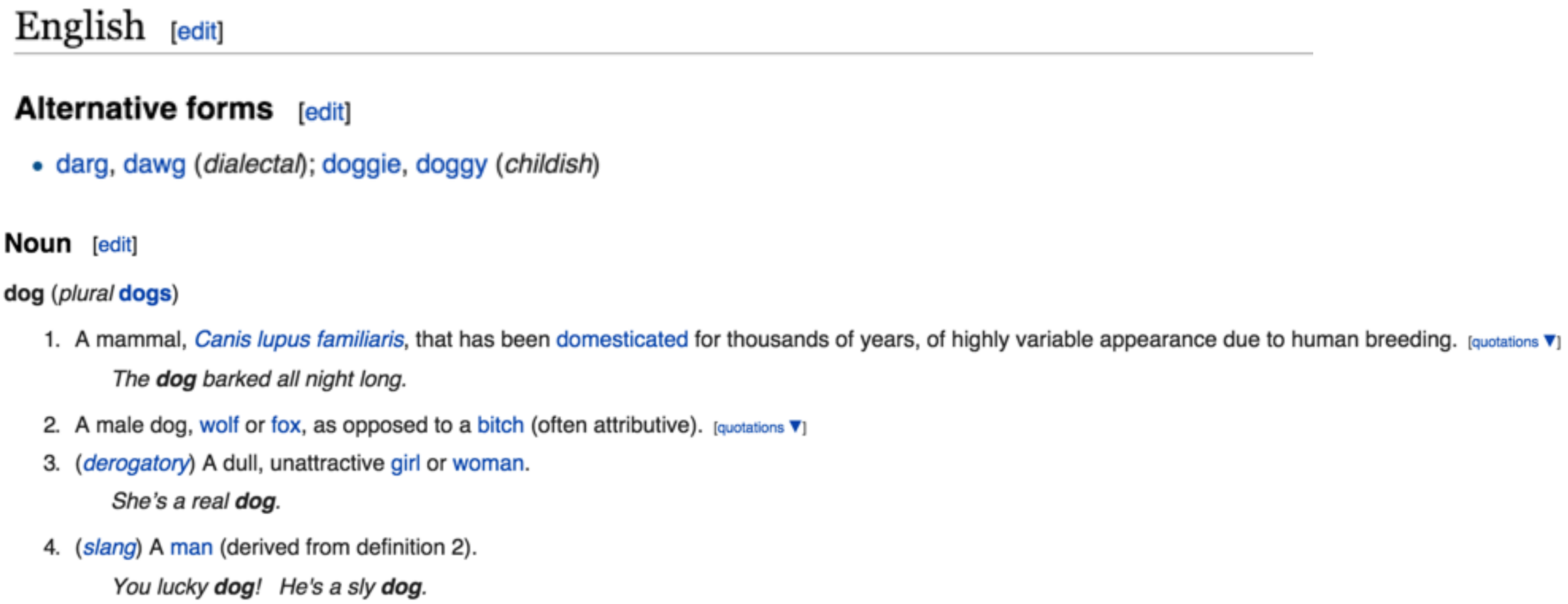
- Can be weighted:

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

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.*

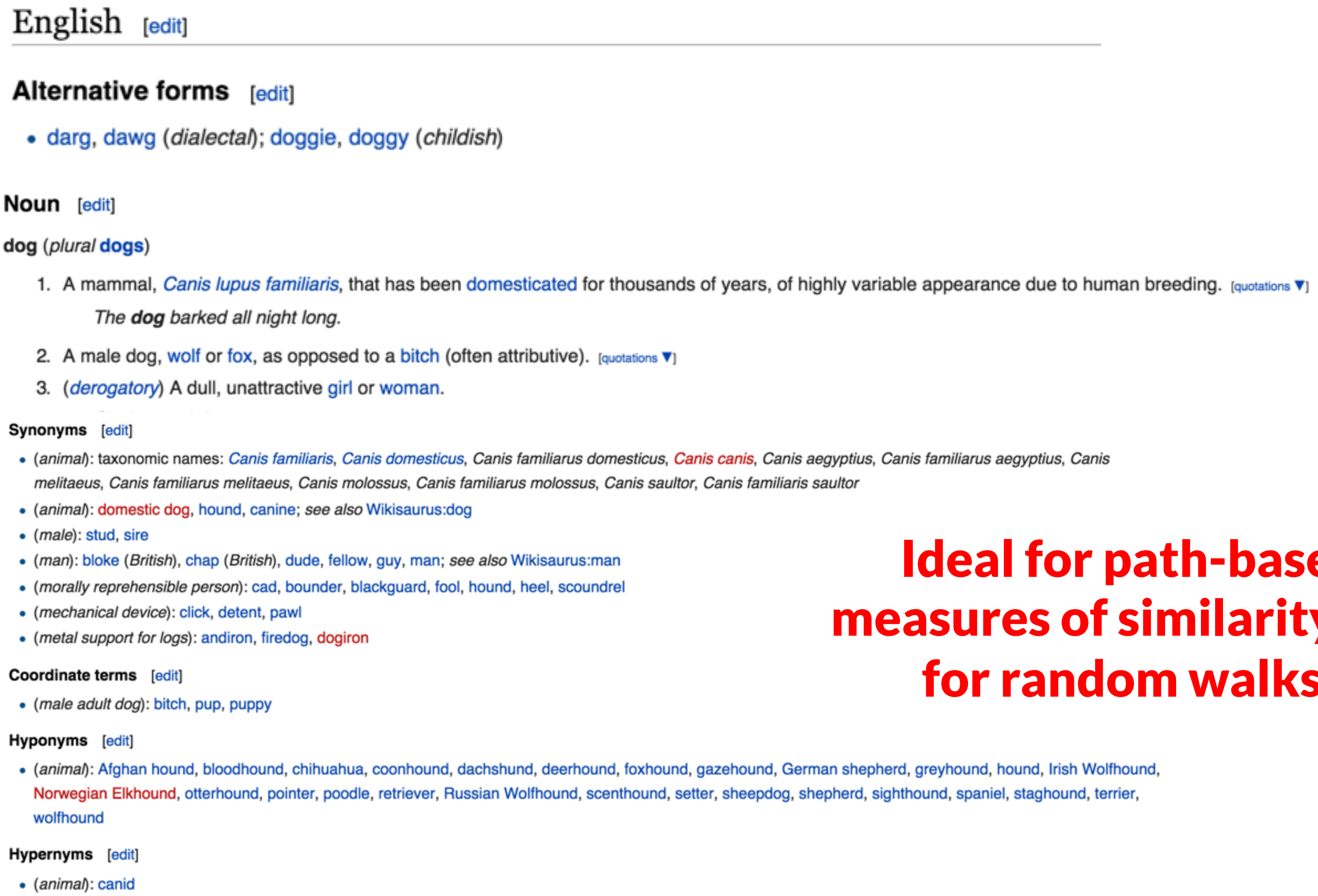
2. A male dog, wolf or fox, as opposed to a bitch (often attributive). [quotations ▼]
3. (*derogatory*) A dull, unattractive girl or woman.

   *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!

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: $p_i = \alpha u_i + \beta v_i + \gamma u_i v_i$

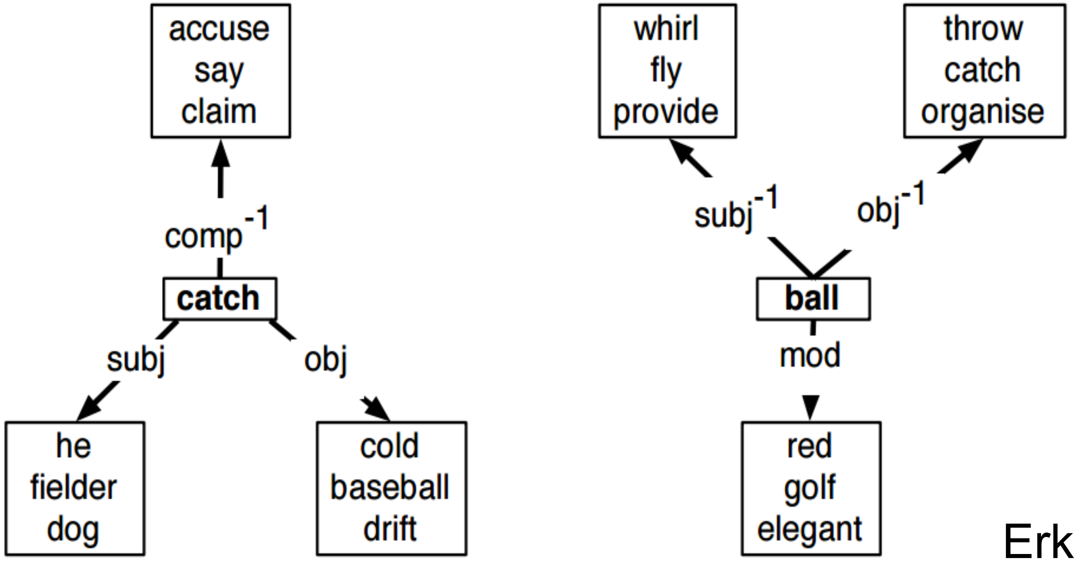
Better at distinguishing high and low semantic similarity

Mitchell and Lapata (2008)

Combine words by taking syntax into account

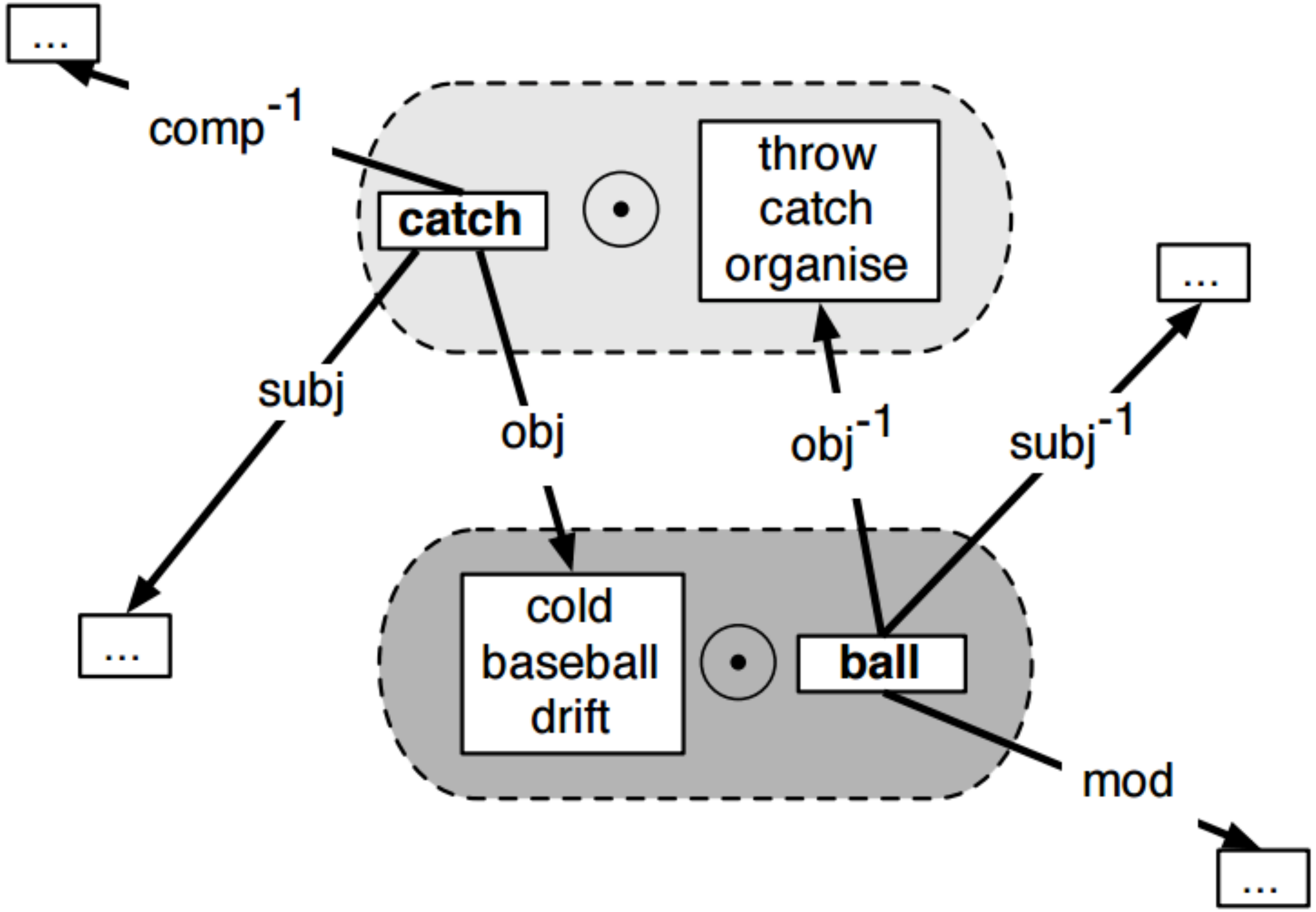
- Integrates lexical information with selectional preferences

- 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: $z_i = u_j + v_k$

Adjectives as matrices: $z_i = U_j v_k$

- Learn each adjective's U by comparing vectors when adjective is and isn't present.

(Baroni and Zamparelli, 2010; Guevara, 2010)

Moving beyond element-wise composition

Simple average: $\mathrm{z_i = u_j + v_k}$

Adjectives as matrices: $\mathrm{z_i = U_j v_k}$

- Learn each adjective's U by comparing vectors when adjective is and isn't present.

Composition as matrices: $\mathrm{z_i = Au_j + Bv_k}$

- 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: $z_i = u_j + v_k$

Adjectives as matrices: $z_i = U_j v_k$

- Learn each adjective's U by comparing vectors when adjective is and isn't present.

Key insight: composition is decoupled from word type!

Composition as matrices: $z_i = Au_j + Bv_k$

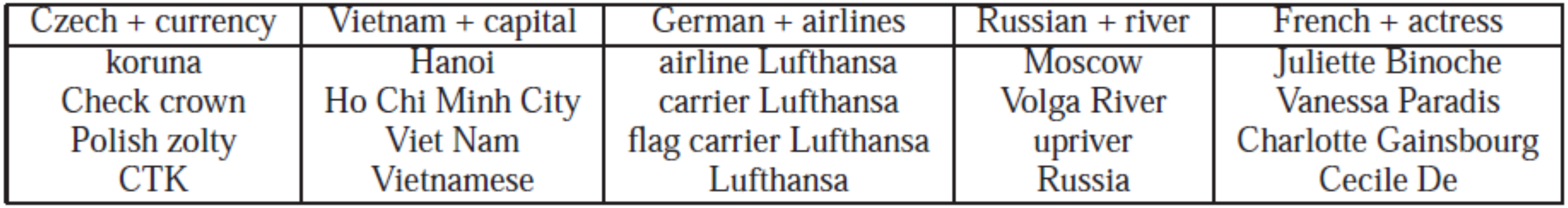
- Estimating A and B is a regression problem with multiple dependent variables. Use a dictionary to find training pairs (u, v, z)!

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

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

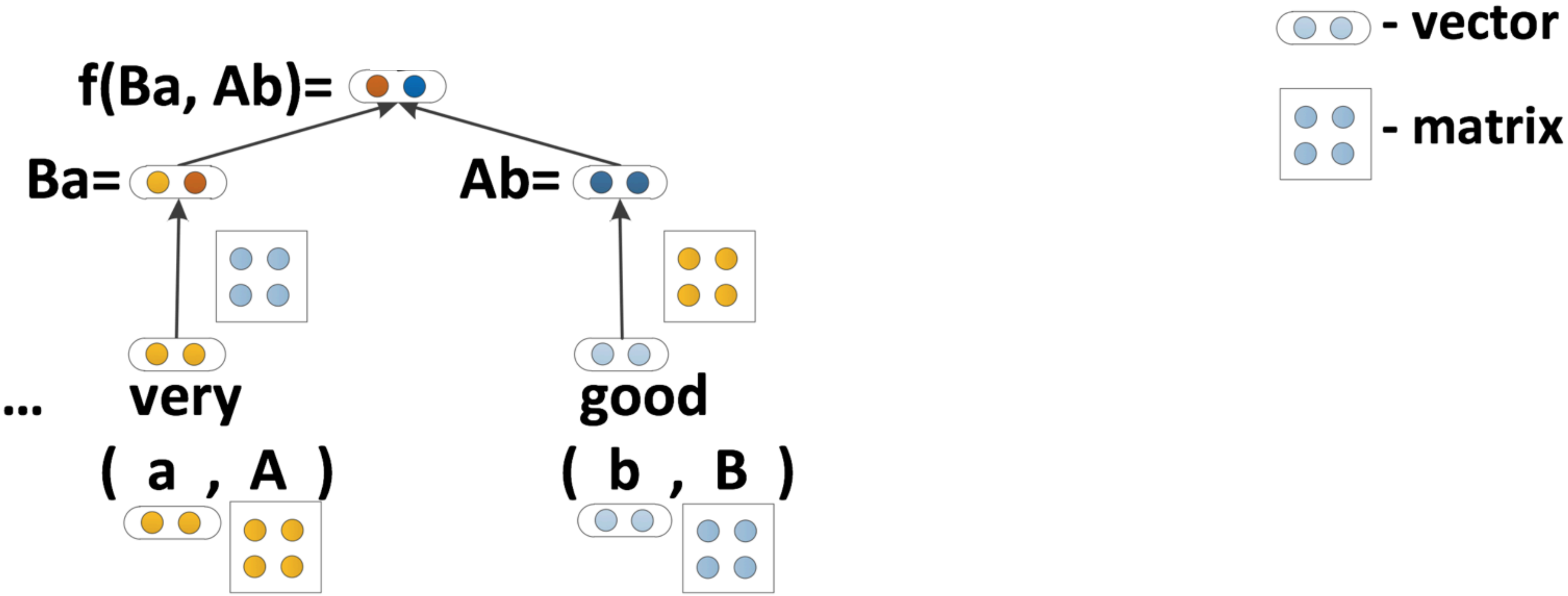


Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese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

Incapable of handling syntactic dependencies or related phrasal constructions

Mikolov et al (2013)

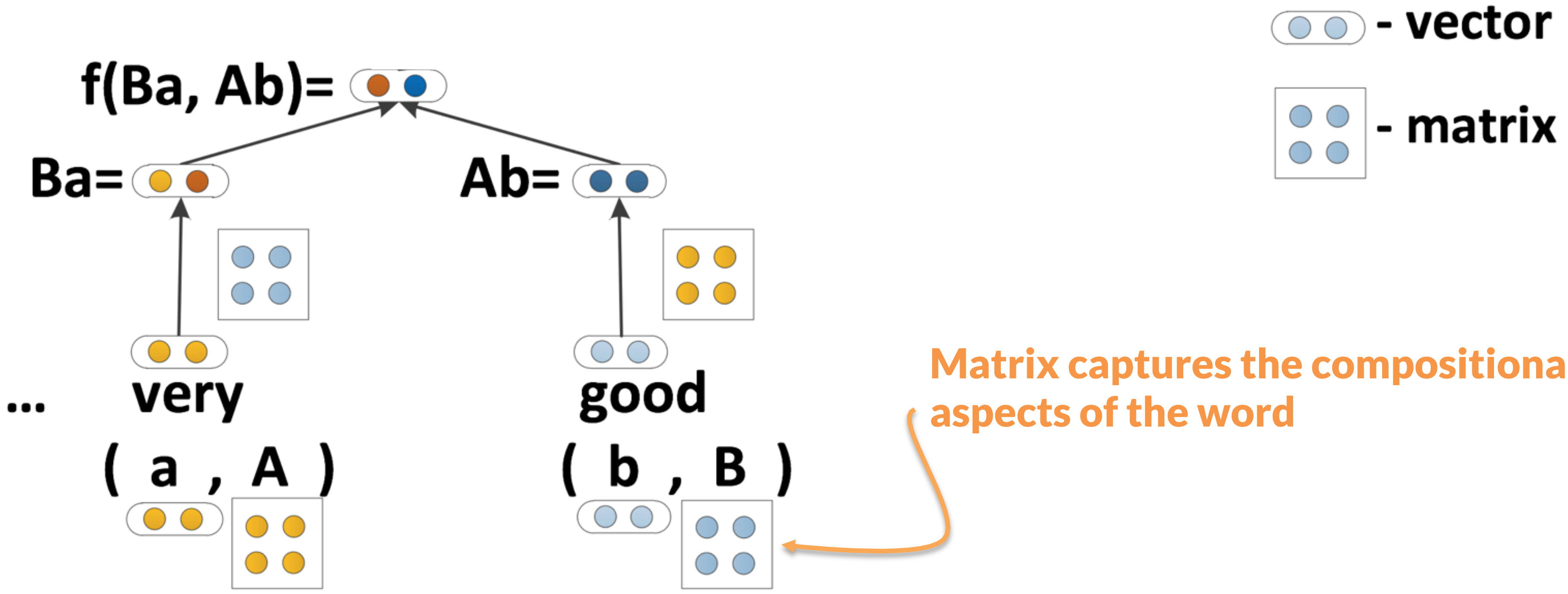
Compose with a recursive neural net



Note: Requires data be parsable.

(Socher et al., 2012)

Compose with a recursive neural net

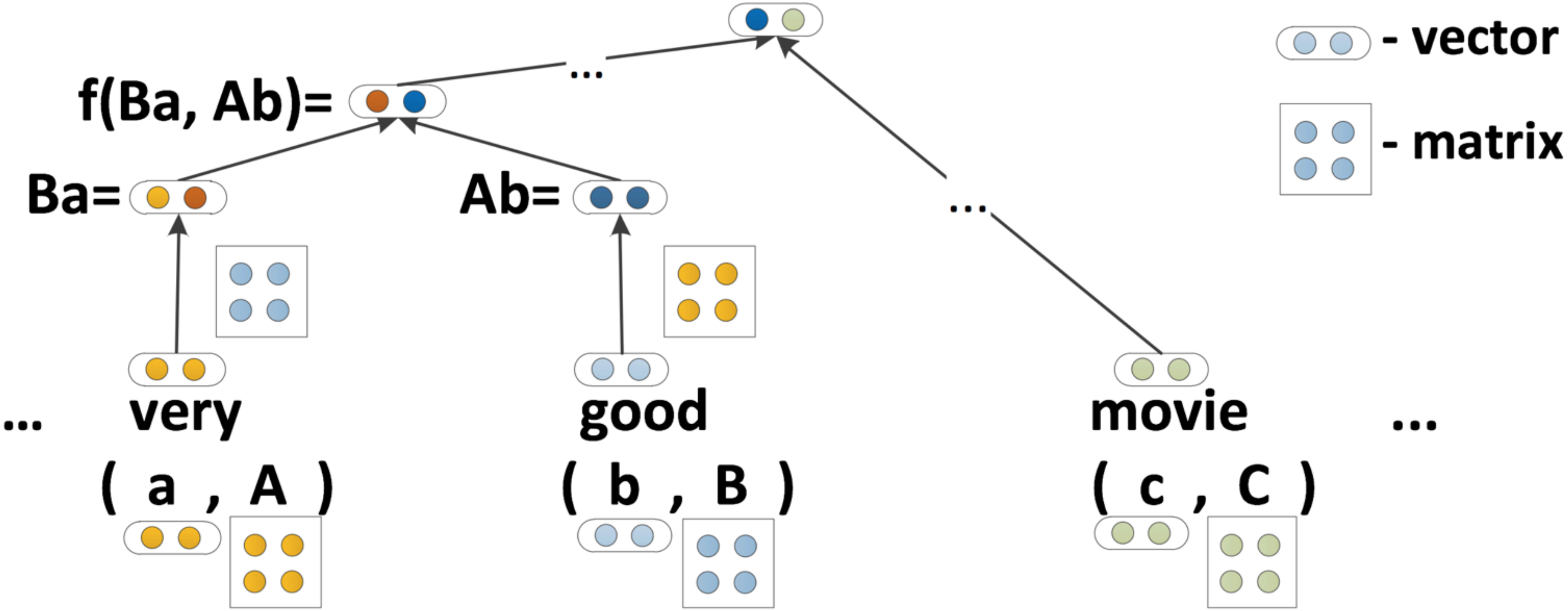


Matrix captures the compositional aspects of the word

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:

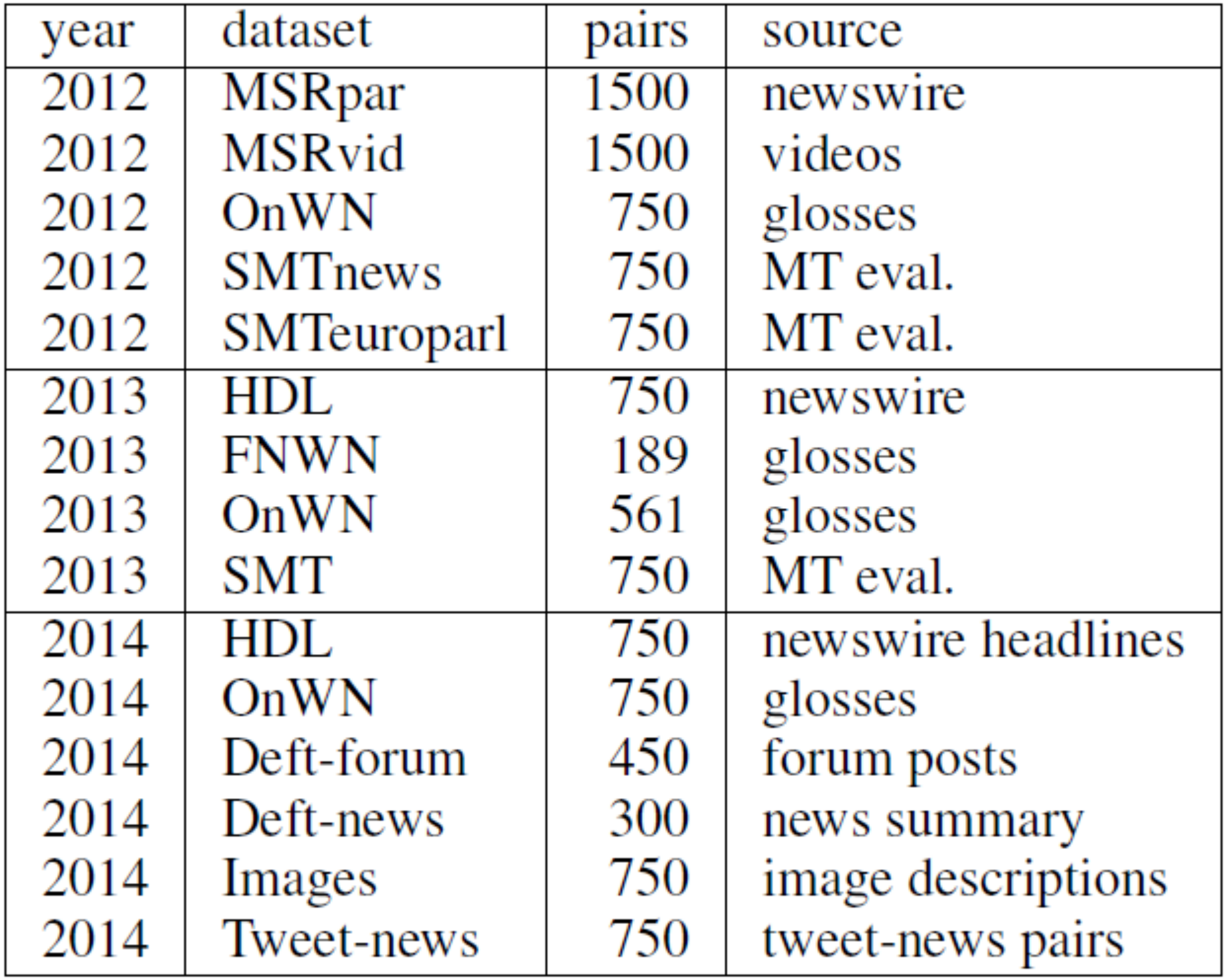
- Paraphrasing
- Textual entailment
- Machine translation
- Question Answering

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



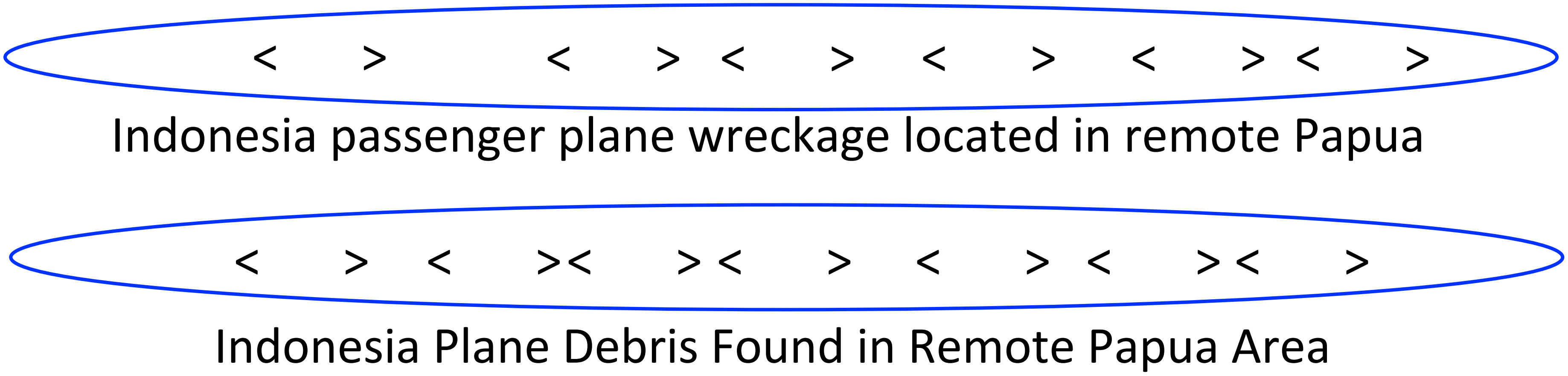
year	dataset	pairs	source
2012	MSRpar	1500	newswire
2012	MSRvid	1500	videos
2012	OnWN	750	glosses
2012	SMTnews	750	MT eval.
2012	SMTeuroparl	750	MT eval.
2013	HDL	750	newswire
2013	FNWN	189	glosses
2013	OnWN	561	glosses
2013	SMT	750	MT eval.
2014	HDL	750	newswire headlines
2014	OnWN	750	glosses
2014	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

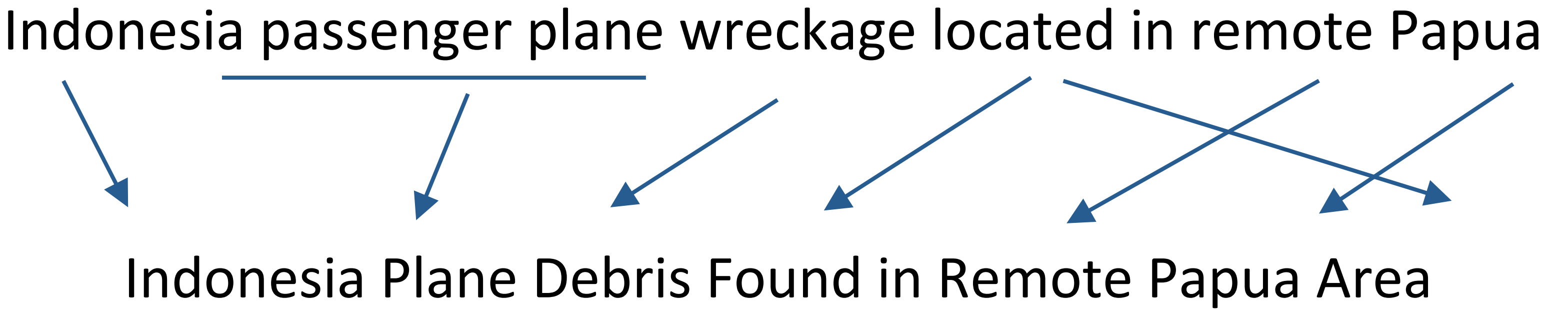
Indonesia Plane Debris Found in Remote Papua Area

Sentence Similarity Techniques

Alignment

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

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



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-
He-
e-i
-is
is-
s-t
…

<b>He-
He-
e-t
-ta
tal
alk
…

Sentence Similarity Techniques

Usually feature-based regression models

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

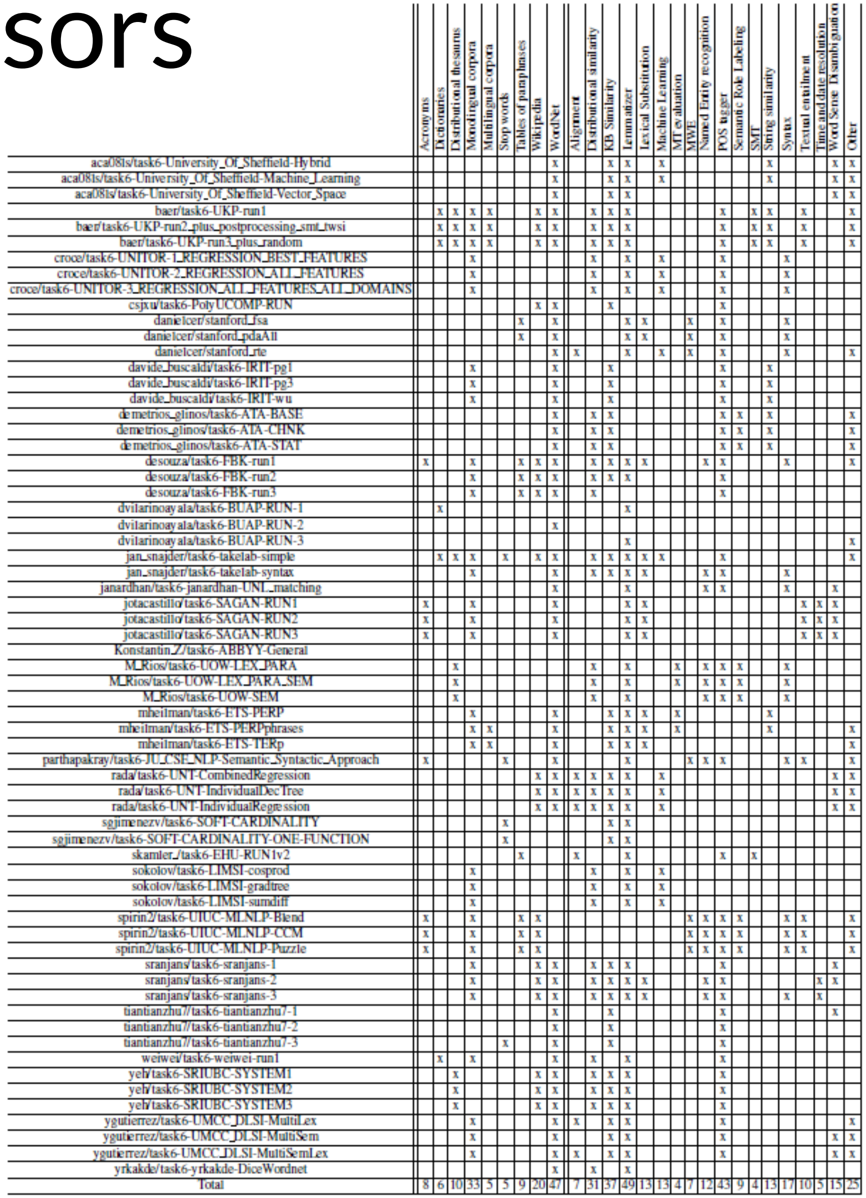
String-based similarity: character n-gram, GST, etc.

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

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

Most STS systems are multi-feature regressors

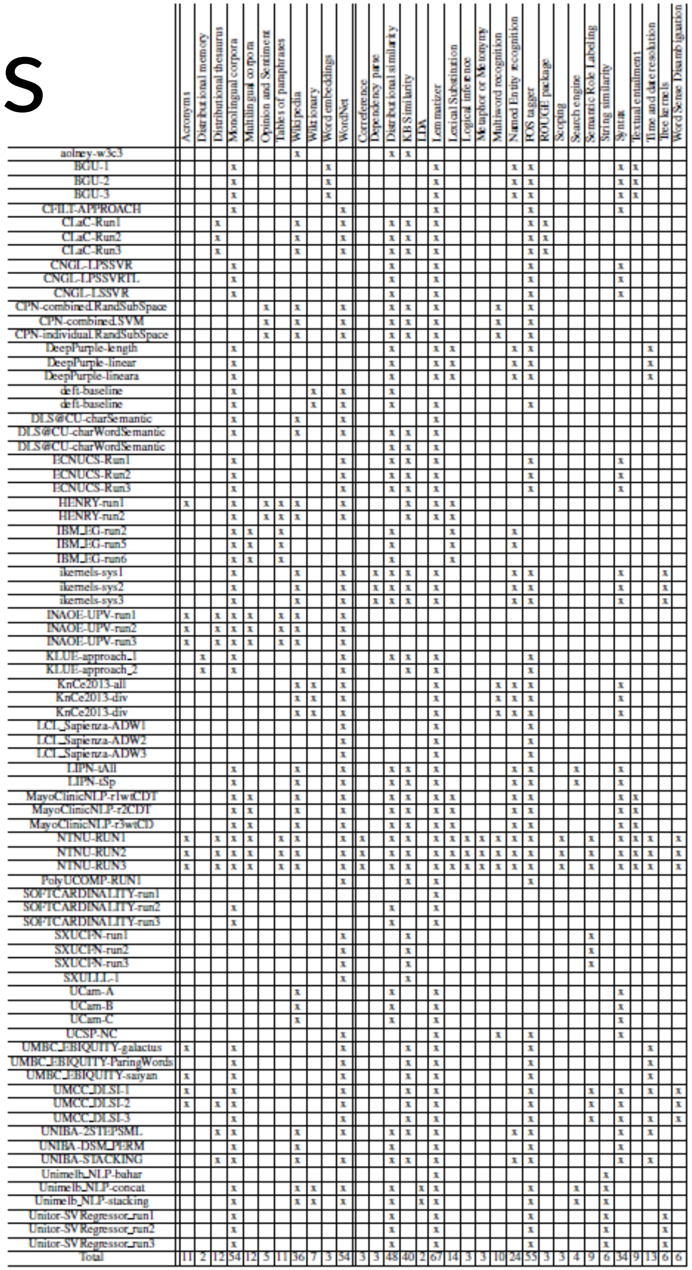
- STS-2012
- Resources and tools used by the systems (from the Task's paper)



	Acronyms	Dictionaries	Distributional thesaurus	Monolingual corpora	Multilingual corpora	Stop words	Tables of paraphrases	Wikipedia	WordNet	Alignment	Distributional similarity	KB Similarity	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_fsa							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						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							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	KB 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	x									
CLaC-Run2			x					x			x			x	x		x						x	x									
CLaC-Run3			x					x			x			x	x		x						x	x									
CNGL-LPSSVR				x										x			x						x						x				
CNGL-LPSSVRTL				x										x			x						x						x				
CNGL-LSSVR				x										x			x						x						x				
CPN-combined.RandSubSpace						x		x			x			x	x		x				x		x										
CPN-combined.SVM						x		x			x			x	x		x				x		x										
CPN-individual.RandSubSpace						x		x			x			x	x		x				x		x										
DeepPurple-length				x										x			x	x				x	x								x		
DeepPurple-linear				x										x			x	x				x	x								x		
DeepPurple-lineara				x										x			x	x				x	x								x		
deft-baseline				x					x		x			x																			
deft-baseline				x					x		x			x			x						x										
DLS@CU-charSemantic				x				x			x						x																
DLS@CU-charWordSemantic				x				x			x			x	x		x																
DLS@CU-charWordSemantic														x	x		x																
ECNUCS-Run1				x							x			x	x		x						x						x				
ECNUCS-Run2				x							x			x	x		x						x						x				
ECNUCS-Run3				x							x			x	x		x						x						x				
HENRY-run1	x			x		x	x	x			x				x		x	x															
HENRY-run2				x		x	x	x			x				x		x	x															
IBM_EG-run2				x	x		x							x				x				x											
IBM_EG-run5				x	x		x							x				x				x											
IBM_EG-run6				x	x		x							x				x															
ikernels-sys1				x				x			x		x	x	x		x					x	x						x			x	
ikernels-sys2				x				x			x		x	x	x		x					x	x						x			x	
ikernels-sys3				x				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						x										
KLUE-approach_2		x		x							x				x		x						x										
KnCe2013-all								x	x		x						x				x	x	x						x				
KnCe2013-div								x	x		x						x				x	x	x						x				
KnCe2013-div								x	x		x						x				x	x	x						x				
LCL_Sapienza-ADW1											x						x						x										
LCL_Sapienza-ADW2											x						x						x										
LCL_Sapienza-ADW3											x						x						x										
LIPN-tAll				x				x			x			x	x		x					x	x			x			x				
LIPN-tSp				x				x			x			x	x		x					x	x			x			x				
MayoClinicNLP-r1wtCDT				x	x			x			x			x	x		x	x				x	x						x	x			
MayoClinicNLP-r2CDT				x	x			x			x			x	x		x	x				x	x						x	x			
MayoClinicNLP-r3wtCD				x	x			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
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
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
PolyUCOMP-RUN1											x				x		x						x										
SOFTCARDINALITY-run1																	x																
SOFTCARDINALITY-run2				x										x			x																
SOFTCARDINALITY-run3				x										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		x						x				
UMBC_EBIQUITY-galactus	x			x							x				x		x						x								x		
UMBC_EBIQUITY-ParingWords				x							x				x		x						x								x		
UMBC_EBIQUITY-saiyan	x			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-2STEPSML			x	x				x			x			x	x		x					x	x						x		x		
UNIBA-DSM_PERM				x				x						x			x						x						x				
UNIBA-STACKING			x	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		x					
Unimelb_NLP-stacking				x				x	x		x			x		x	x						x			x		x					
Unitor-SVRegressor_run1				x										x			x						x					x				x	
Unitor-SVRegressor_run2				x										x			x						x					x				x	
Unitor-SVRegressor_run3				x										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

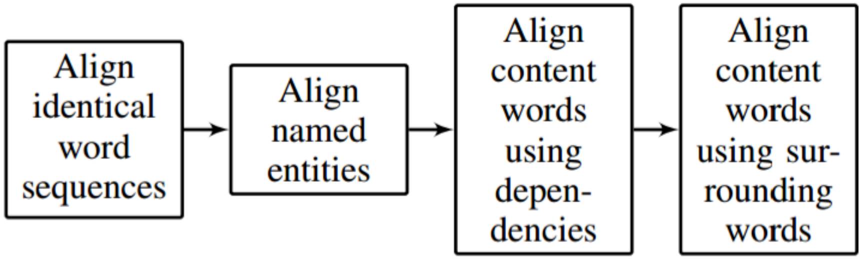
$$SIM(A,B) = \frac{|A \cap B| + bias}{\alpha \max\left(|A|,|B|\right) + (1-\alpha)\min\left(|A|,|B|\right)}$$

Sentence Similarity Techniques

Monolingual alignment

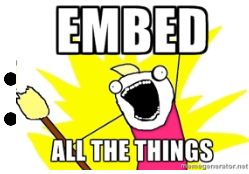
Sultan et al (2014): best system in STS-14 and -15

DLS@CU


Align identical word sequences
Align named entities
Align content words using dependencies
Align content words using surrounding words

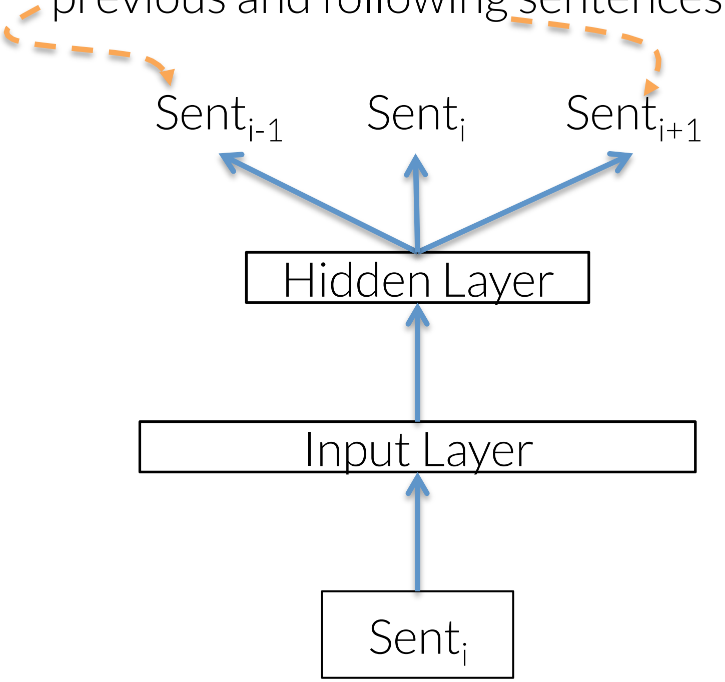
The similarity score is computed as a function of the proportions of aligned content words in the two input sentences.

Sentence Similarity Techniques: Skip-thought vectors



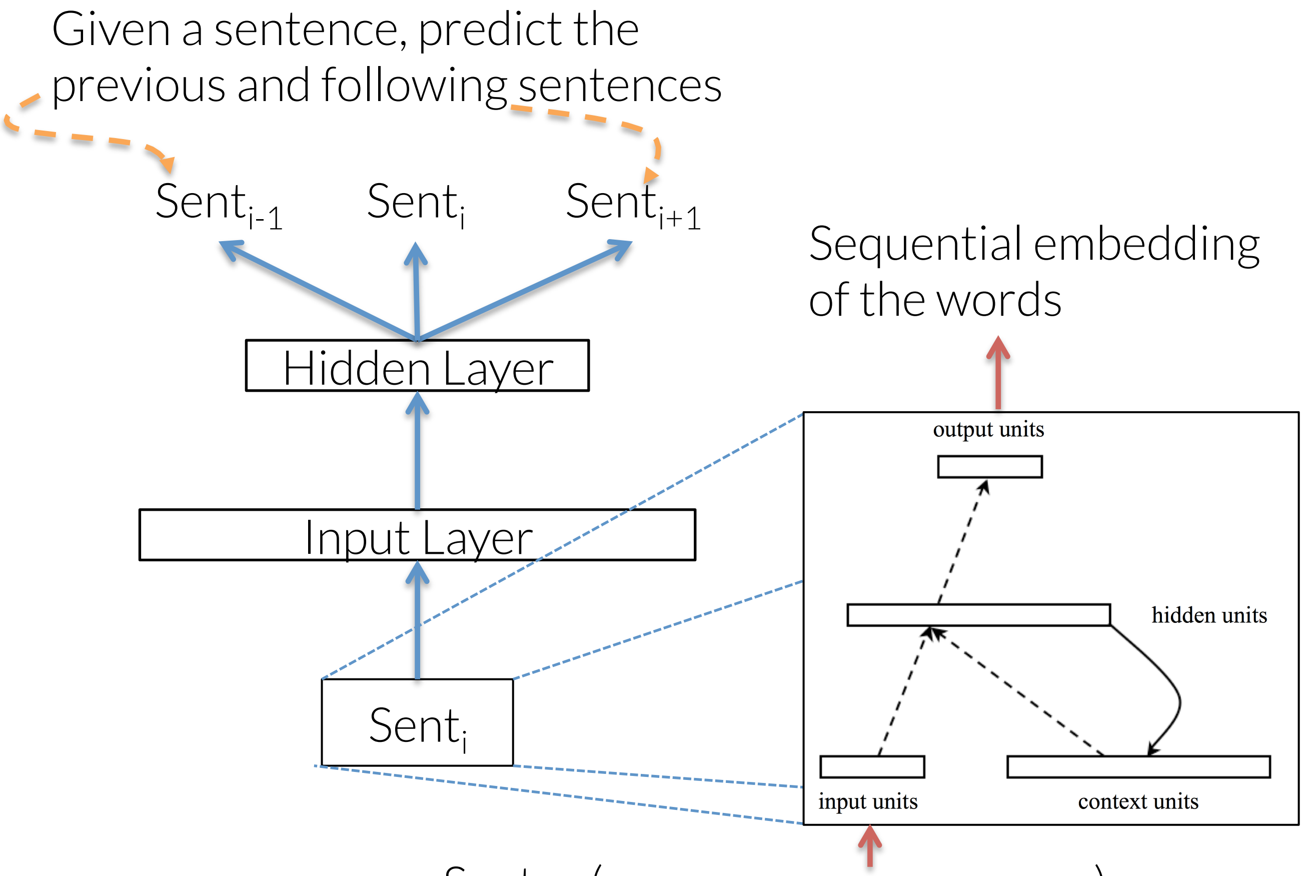
Embedding a sentence with unsupervised training

Given a sentence, predict the previous and following sentences



Sentence Similarity Techniques: Skip-thought vectors

Embedding a sentence with unsupervised training



Given a sentence, predict the previous and following sentences

Sent_{i-1} Sent_{i} Sent_{i+1}

Hidden Layer

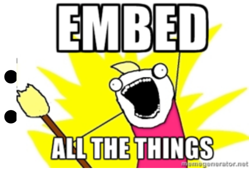
Input Layer

Sent_{i}

Sequential embedding of the words

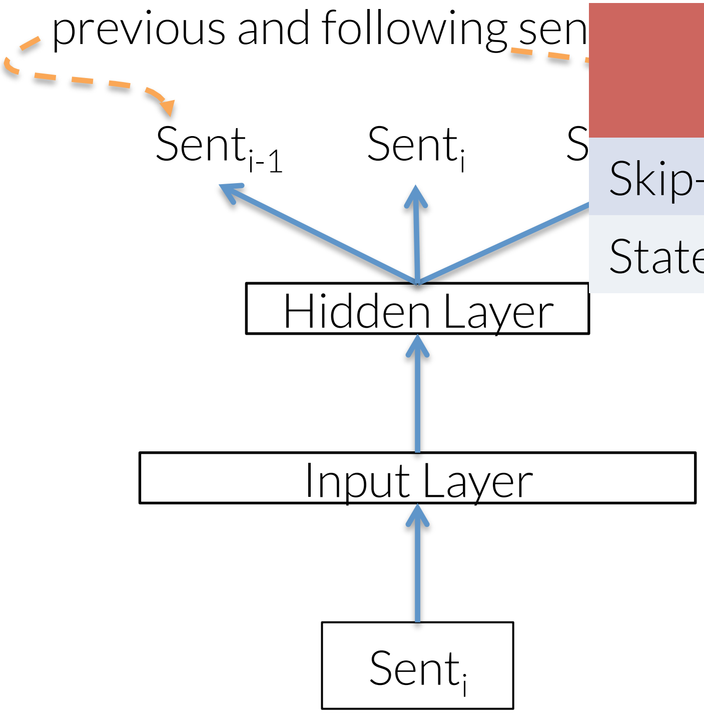
$\text{Sent}_i = (w_1, w_2, w_3, w_4, w_5, \dots w_{\underline{N}})$

Sentence Similarity Techniques: Skip-thought vectors



Embedding a sentence with unsupervised training

Given a sentence, predict the previous and following sen



	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

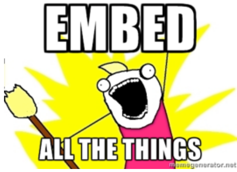
Simplest Idea: Model paragraphs as a bag of words (BoW)

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

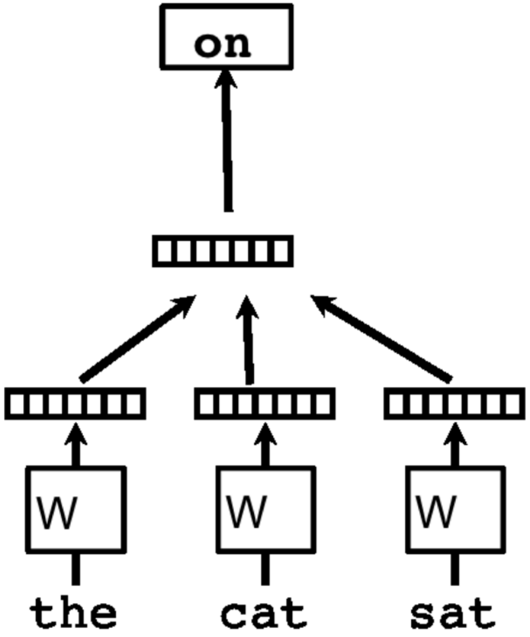
- The topic sentence can matter!



Current state of the art: doc2vec

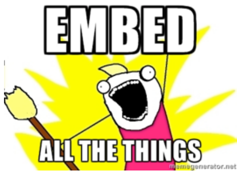
Tackles two problems with bag-of-word and topic modeling approaches:

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



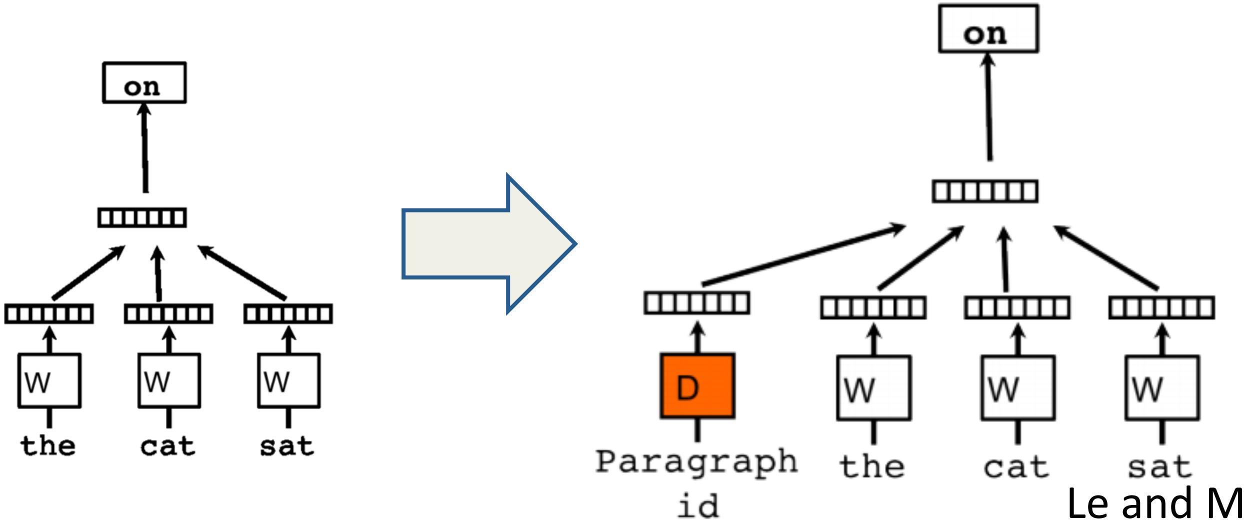
Base model is a prediction task to predict the next word in a sequence

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

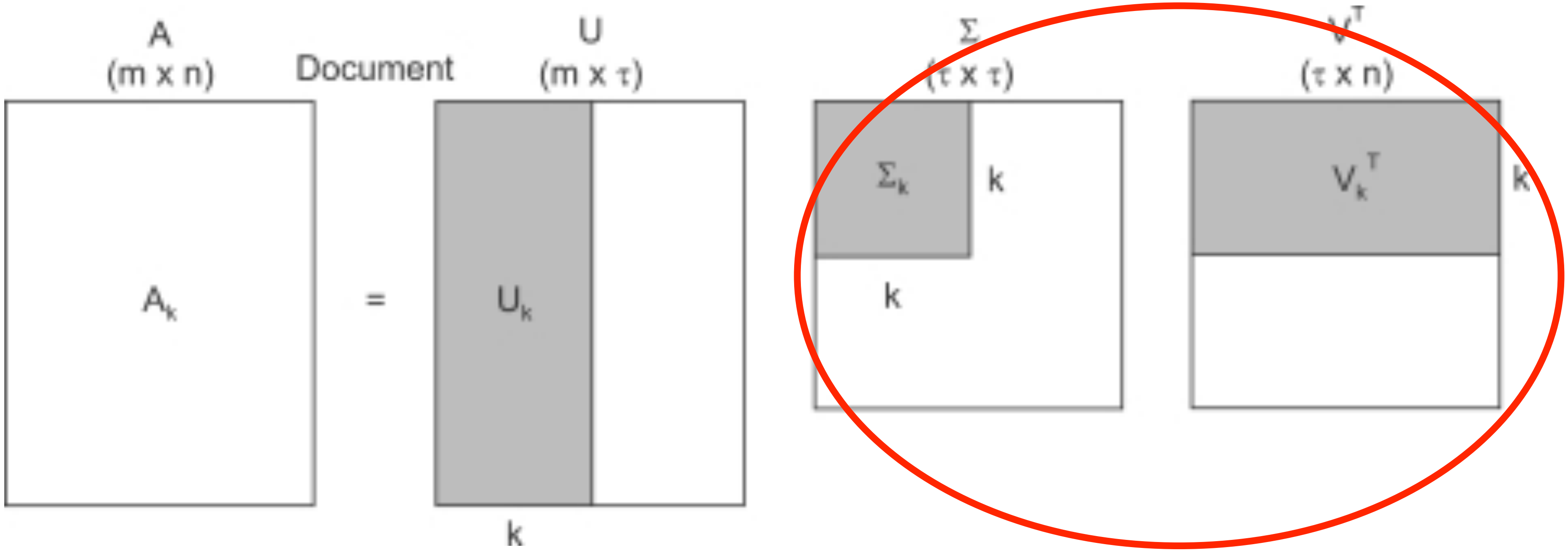


Le and Mikolov (2014)

Document Similarity

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

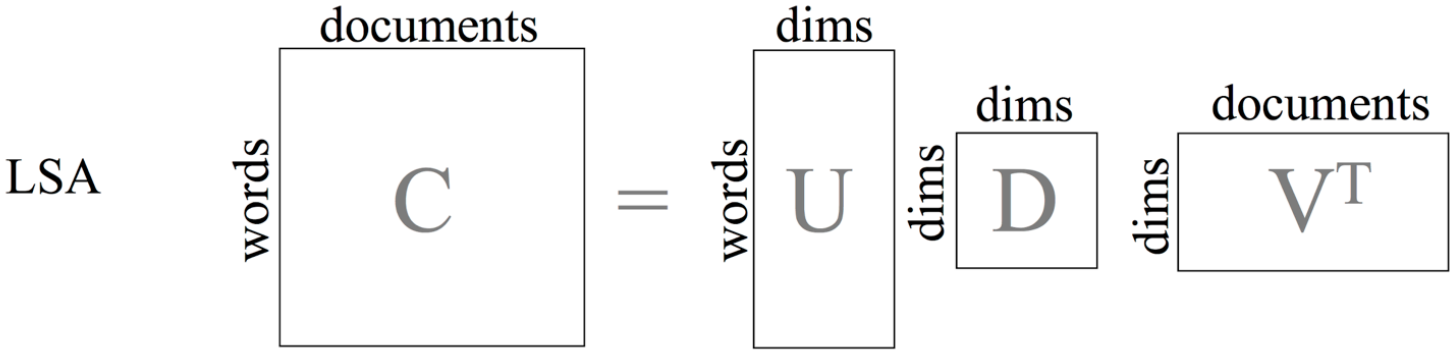


Image credit: Steyvers and Griffiths (2007)

Topic Modeling: Viewing document contents as a mixture of topics

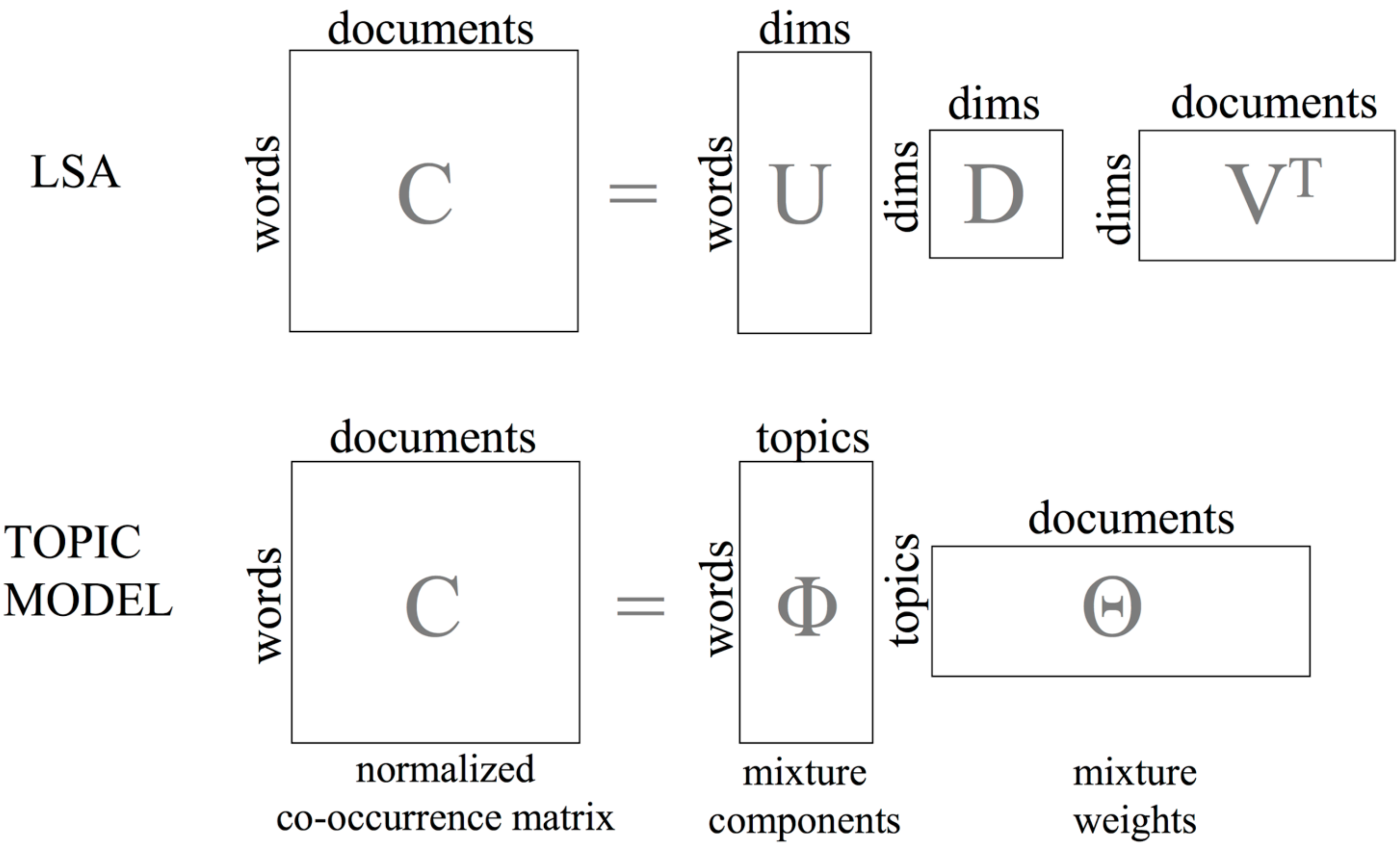
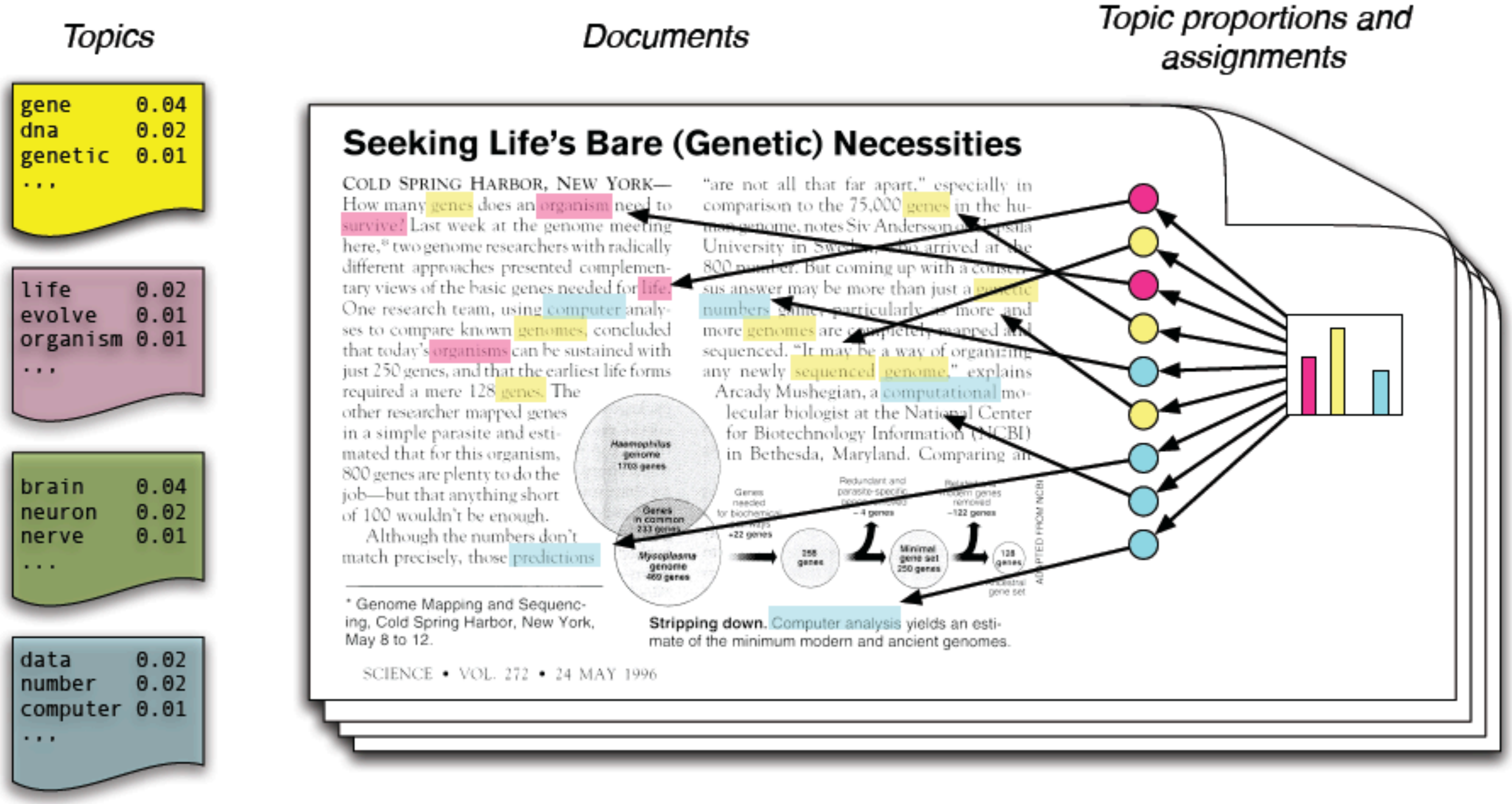


Image credit: Steyvers and Griffiths (2007)

Topic Modeling: Viewing document contents as a mixture of topics



gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related to modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Image credit: Blei (2012)

Document Similarity Techniques

Latent Dirichlet Allocation

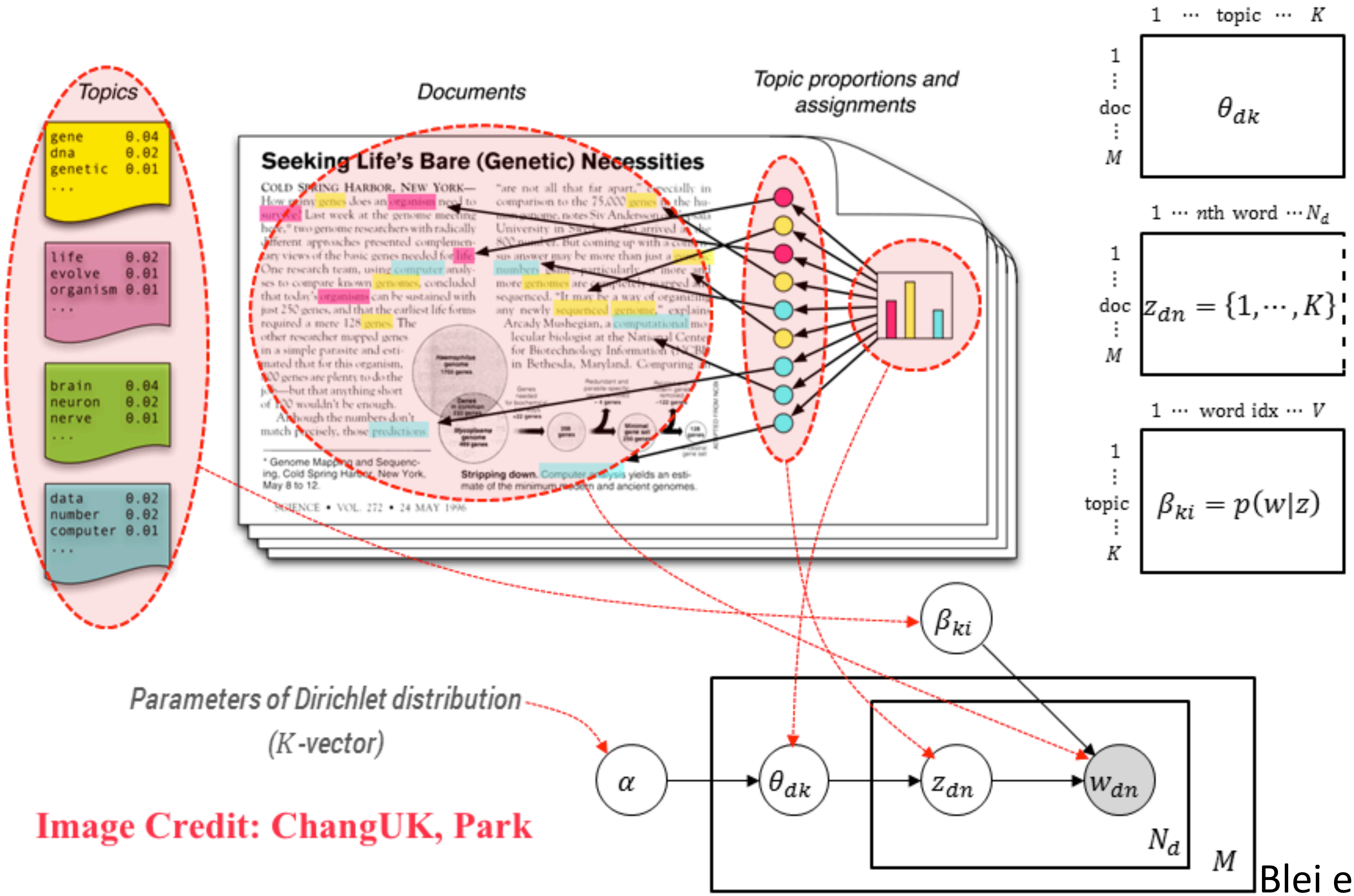


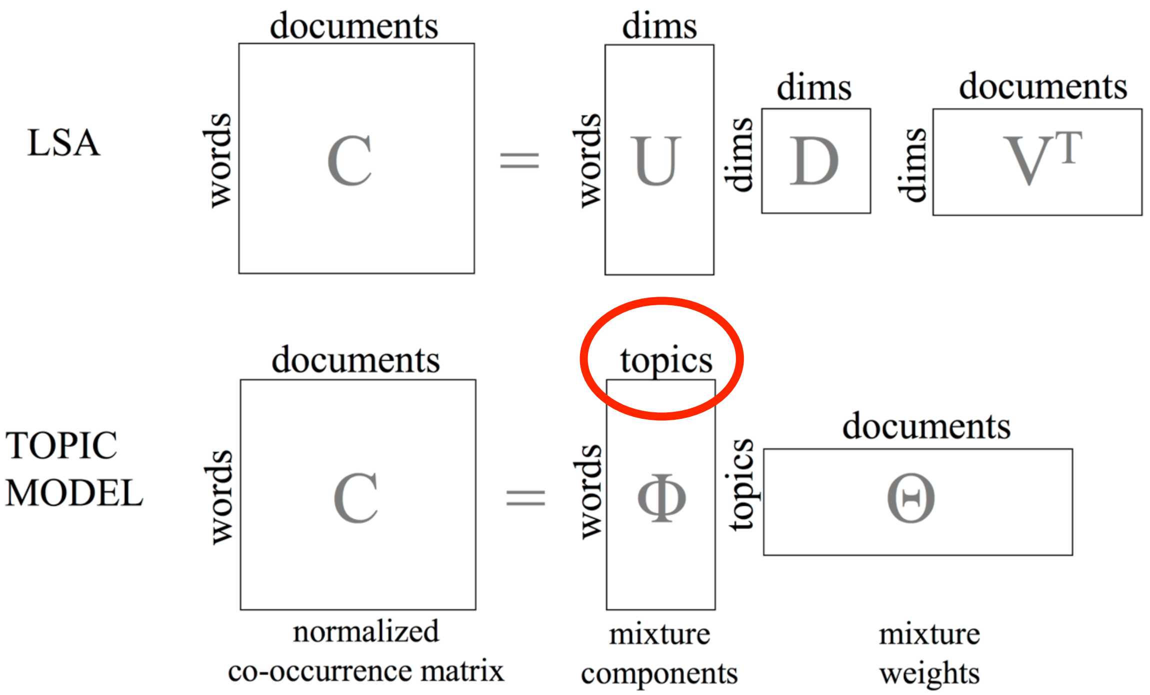
Image Credit: ChangUK, Park

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

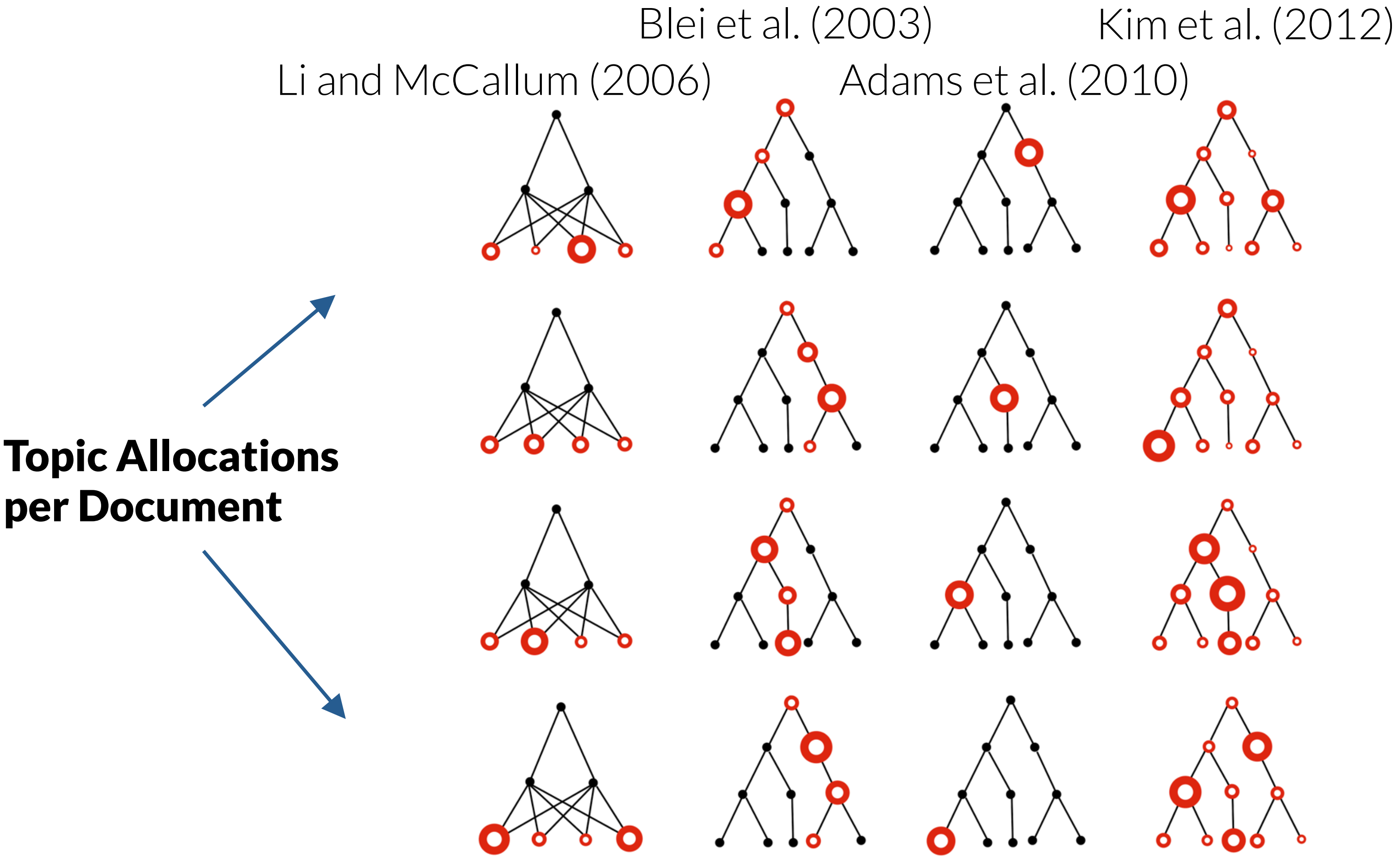
How many topics should you use?



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

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

Introducing structure into the topics



Blei et al. (2003)

Kim et al. (2012)

Li and McCallum (2006)

Adams et al. (2010)

Topic Allocations per Document

Hierarchical topic organizations can potentially yield more informative document representations

Image credit: Kim et al. (2012)

Incorporating Multi-Word Expressions into topics

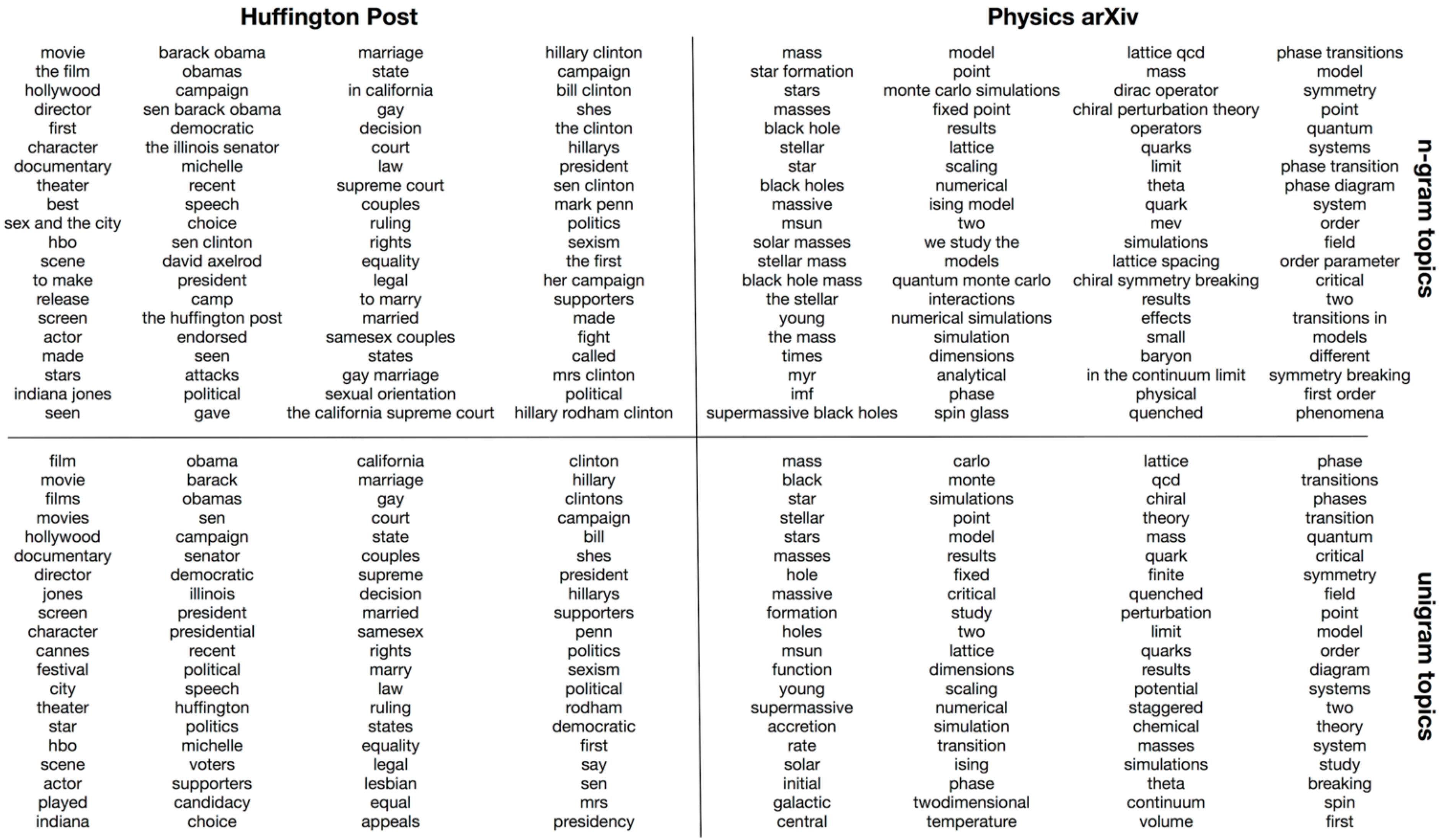
- Pre-process the corpus to glob MWEs together, e.g., “white house” -> white_house
  - –Not feasible for domain-specific MWEs

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



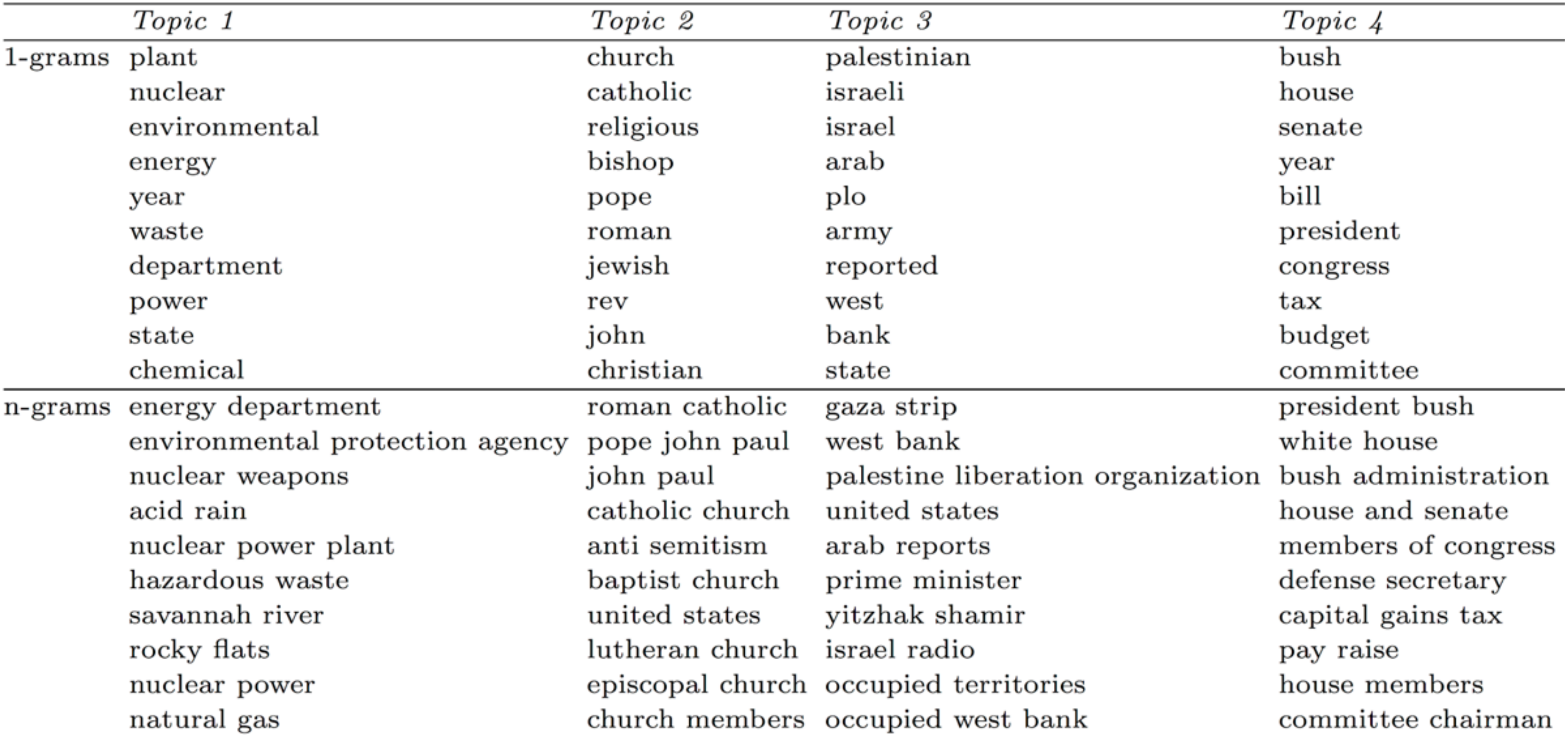
Huffington Post				Physics arXiv				
movie	barack obama	marriage	hillary clinton	mass	model	lattice qcd	phase transitions	n-gram topics
the film	obamas	state	campaign	star formation	point	mass	model	
hollywood	campaign	in california	bill clinton	stars	monte carlo simulations	dirac operator	symmetry	
director	sen barack obama	gay	shes	masses	fixed point	chiral perturbation theory	point	
first	democratic	decision	the clinton	black hole	results	operators	quantum	
character	the illinois senator	court	hillarys	stellar	lattice	quarks	systems	
documentary	michelle	law	president	star	scaling	limit	phase transition	
theater	recent	supreme court	sen clinton	black holes	numerical	theta	phase diagram	
best	speech	couples	mark penn	massive	ising model	quark	system	
sex and the city	choice	ruling	politics	msun	two	mev	order	
hbo	sen clinton	rights	sexism	solar masses	we study the	simulations	field	
scene	david axelrod	equality	the first	stellar mass	models	lattice spacing	order parameter	
to make	president	legal	her campaign	black hole mass	quantum monte carlo	chiral symmetry breaking	critical	
release	camp	to marry	supporters	the stellar	interactions	results	two	
screen	the huffington post	married	made	young	numerical simulations	effects	transitions in	
actor	endorsed	samesex couples	fight	the mass	simulation	small	models	
made	seen	states	called	times	dimensions	baryon	different	
stars	attacks	gay marriage	mrs clinton	myr	analytical	in the continuum limit	symmetry breaking	
indiana jones	political	sexual orientation	political	imf	phase	physical	first order	
seen	gave	the california supreme court	hillary rodham clinton	supermassive black holes	spin glass	quenched	phenomena	
film	obama	california	clinton	mass	carlo	lattice	phase	unigram topics
movie	barack	marriage	hillary	black	monte	qcd	transitions	
films	obamas	gay	clintons	star	simulations	chiral	phases	
movies	sen	court	campaign	stellar	point	theory	transition	
hollywood	campaign	state	bill	stars	model	mass	quantum	
documentary	senator	couples	shes	masses	results	quark	critical	
director	democratic	supreme	president	hole	fixed	finite	symmetry	
jones	illinois	decision	hillarys	massive	critical	quenched	field	
screen	president	married	supporters	formation	study	perturbation	point	
character	presidential	samesex	penn	holes	two	limit	model	
cannes	recent	rights	politics	msun	lattice	quarks	order	
festival	political	marry	sexism	function	dimensions	results	diagram	
city	speech	law	political	young	scaling	potential	systems	
theater	huffington	ruling	rodham	supermassive	numerical	staggered	two	
star	politics	states	democratic	accretion	simulation	chemical	theory	
hbo	michelle	equality	first	rate	transition	masses	system	
scene	voters	legal	say	solar	ising	simulations	study	
actor	supporters	lesbian	sen	initial	phase	theta	breaking	
played	candidacy	equal	mrs	galactic	twodimensional	continuum	spin	
indiana	choice	appeals	presidency	central	temperature	volume	first	

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)

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

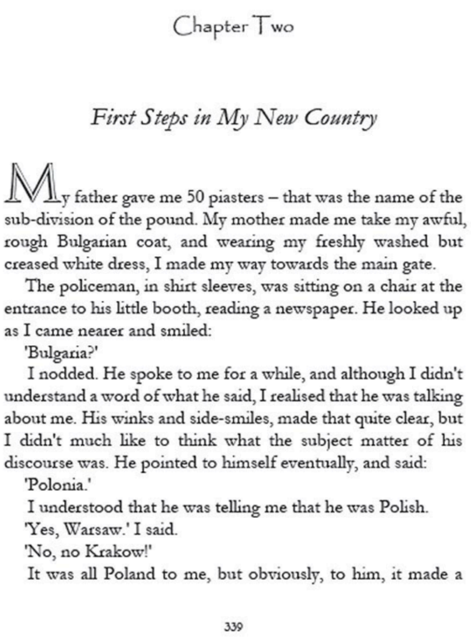
Social network lays emphasis on erpersonal interaction and reflects user's essential need for internet service. In recent years, SNS has become an increasing concern because it allows users to create and manage things independently by themselves. This concern not only comes from the need to analyze network 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 © Springer-Verlag Berlin Heidelberg 2012



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.'

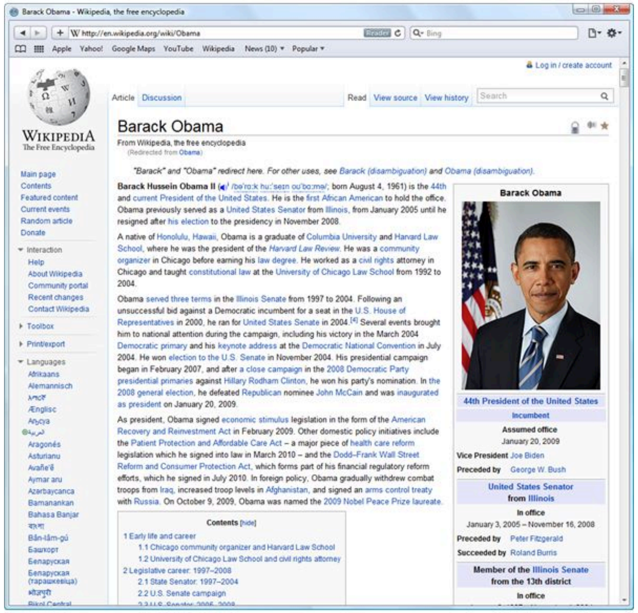
I understood that he was telling me that he was Polish.

'Yes, Warsaw.' I said.

'No, no Krakow!'

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

339



Barack Obama

From Wikipedia, the free encyclopedia
(Redirected from Obama)

"Barack" and "Obama" redirect here. For other uses, see Barack (disambiguation) and Obama (disambiguation).

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/; born August 4, 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama previously served as a United States Senator from Illinois, from January 2005 until he resigned after his election to the presidency in November 2008.

A native of Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the *Harvard Law Review*. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004.

Obama served three terms in the Illinois Senate from 1997 to 2004. Following an unsuccessful bid against a Democratic incumbent for a seat in the U.S. House of Representatives in 2000, he ran for United States Senate in 2004.[4] Several events brought him to national attention during the campaign, including his victory in the March 2004 Democratic primary and his keynote address at the Democratic National Convention in July 2004. He won election to the U.S. Senate in November 2004. His presidential campaign began in February 2007, and after a close campaign in the 2008 Democratic Party presidential primaries against Hillary Rodham Clinton, he won his party's nomination. In the 2008 general election, he defeated Republican nominee John McCain and was inaugurated as president on January 20, 2009.

As president, Obama signed economic stimulus legislation in the form of the American Recovery and Reinvestment Act in February 2009. Other domestic policy initiatives include the Patient Protection and Affordable Care Act – a major piece of health care reform legislation which he signed into law in March 2010 – and the Dodd–Frank Wall Street Reform and Consumer Protection Act, which forms part of his financial regulatory reform efforts, which he signed in July 2010. In foreign policy, Obama gradually withdrew combat troops from Iraq, increased troop levels in Afghanistan, and signed an arms control treaty with Russia. On October 9, 2009, Obama was named the 2009 Nobel Peace Prize laureate.

Barack Obama

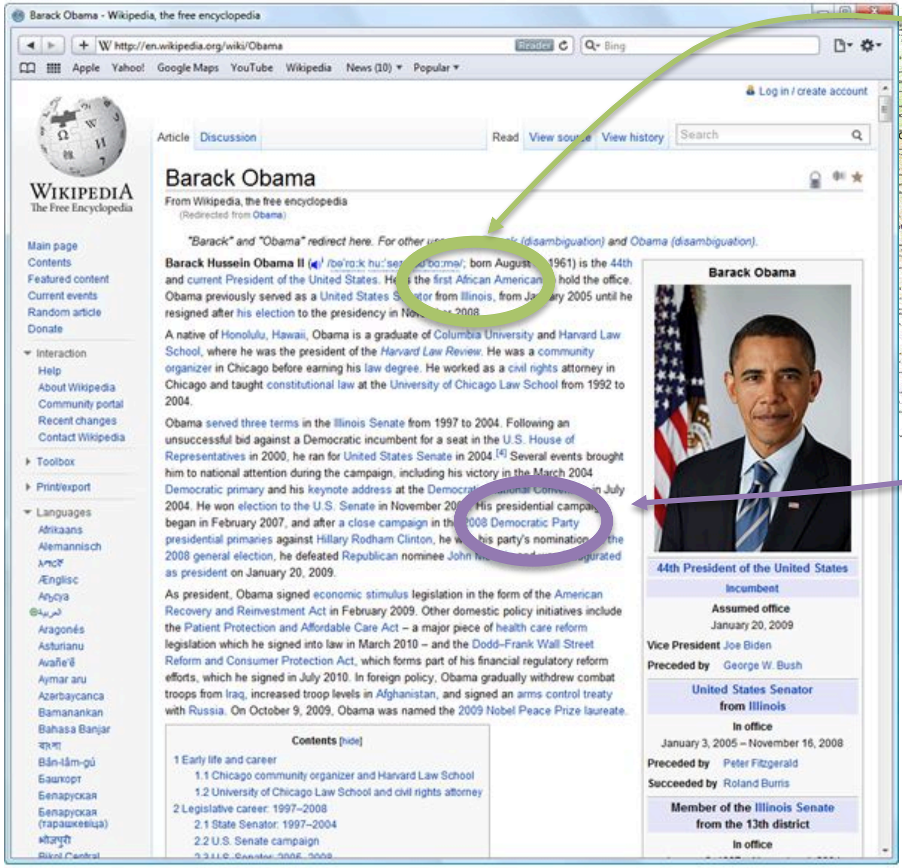
44th President of the United States
Incumbent
Assumed office
January 20, 2009
Vice President Joe Biden
Preceded by George W. Bush

United States Senator from Illinois
In office
January 3, 2005 – November 16, 2008
Preceded by Peter Fitzgerald
Succeeded by Roland Burris

Member of the Illinois Senate from the 13th district
In office

Lots of work on structured document similarity

Adding knowledge to the document representation



Barack Obama - Wikipedia, the free encyclopedia

http://en.wikipedia.org/wiki/Obama

Apple Yahoo! Google Maps YouTube Wikipedia News (10) ▾ Popular ▾

Log in / create account

Article Discussion Read View source View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate

▾ Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

▸ Toolbox

▸ Print/export

▾ Languages
Afrikaans
Alemannisch
አማርኛ
Ænglisc
Аҧсуа
العربية
Aragonés
Asturianu
Avañe'ẽ
Aymar aru
Azərbaycanca
Bamanankan
Bahasa Banjar
বাংলা
Bân-lâm-gú
Башҡорт
Беларуская
Беларуская (тарашкевіца)
भोजपुरी
Bikol Central

Barack Obama

From Wikipedia, the free encyclopedia
(Redirected from Obama)

"Barack" and "Obama" redirect here. For other u... (disambiguation) and Obama (disambiguation).

Barack Hussein Obama II (/bəˈrɑːk huːˈse... bɑːmə/; born August ... 1961) is the 44th and current President of the United States. He ... the first African American ... hold the office. Obama previously served as a United States S...tor from Illinois, from Ja...ary 2005 until he resigned after his election to the presidency in Nov... 2008.

A native of Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the *Harvard Law Review*. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004.

Obama served three terms in the Illinois Senate from 1997 to 2004. Following an unsuccessful bid against a Democratic incumbent for a seat in the U.S. House of Representatives in 2000, he ran for United States Senate in 2004.[4] Several events brought him to national attention during the campaign, including his victory in the March 2004 Democratic primary and his keynote address at the Democrat... in July 2004. He won election to the U.S. Senate in November 20... His presidential campa... began in February 2007, and after a close campaign in th...008 Democratic Party presidential primaries against Hillary Rodham Clinton, he w... his party's nomination ... the 2008 general election, he defeated Republican nominee John ... ugurated as president on January 20, 2009.

As president, Obama signed economic stimulus legislation in the form of the American Recovery and Reinvestment Act in February 2009. Other domestic policy initiatives include the Patient Protection and Affordable Care Act – a major piece of health care reform legislation which he signed into law in March 2010 – and the Dodd–Frank Wall Street Reform and Consumer Protection Act, which forms part of his financial regulatory reform efforts, which he signed in July 2010. In foreign policy, Obama gradually withdrew combat troops from Iraq, increased troop levels in Afghanistan, and signed an arms control treaty with Russia. On October 9, 2009, Obama was named the 2009 Nobel Peace Prize laureate.

Contents [hide]

Barack Obama

44th President of the United States

Incumbent

Assumed office
January 20, 2009

Vice President	Joe Biden
Preceded by	George W. Bush

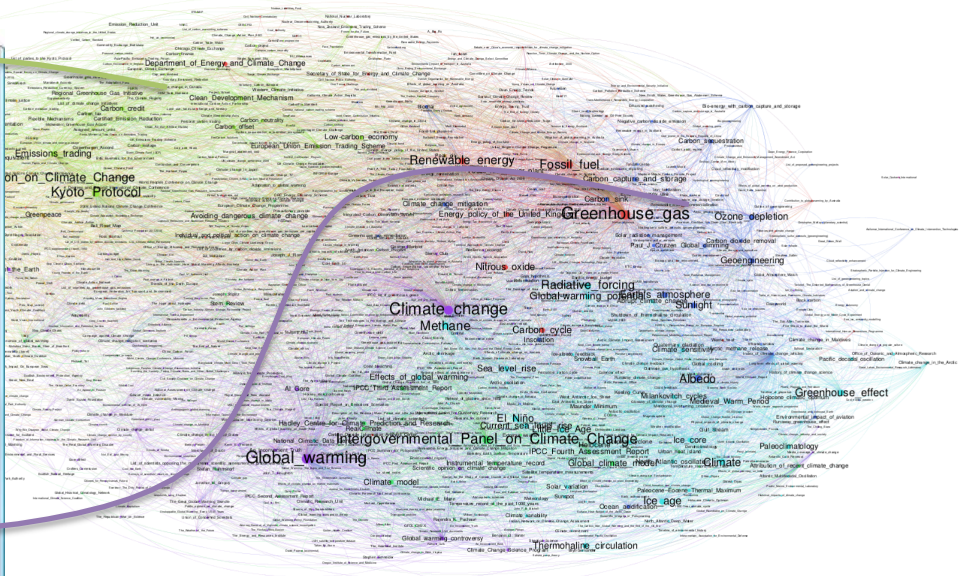
United States Senator
from Illinois

In office
January 3, 2005 – November 16, 2008

Preceded by	Peter Fitzgerald
Succeeded by	Roland Burris

Member of the Illinois Senate
from the 13th district

In office

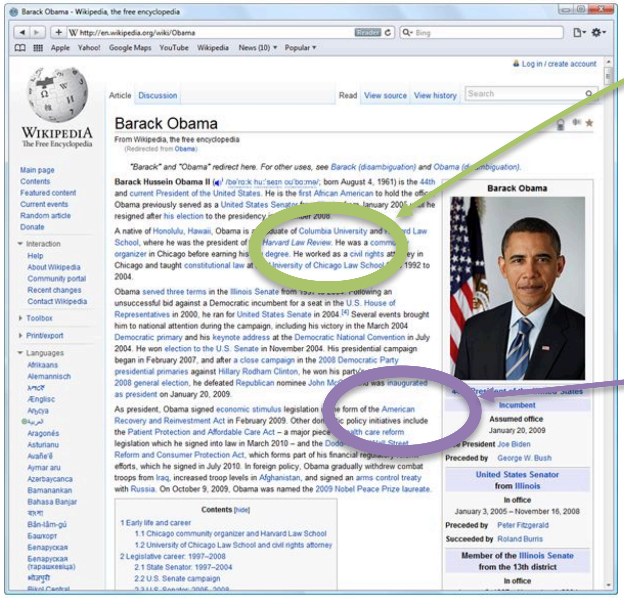


Entities that can be linked in a document become connected to Wikipedia's semantic network

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

Schuhmacher and Ponzetto (2014)

Adding knowledge to the document representation



Barack Obama

From Wikipedia, the free encyclopedia
(Redirected from Obama)

"Barack" and "Obama" redirect here. For other uses, see Barack (disambiguation) and Obama (disambiguation).

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/; born August 4, 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama previously served as a United States Senator from Illinois from January 2005 until he resigned after his election to the presidency in November 2008.

A native of Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004.

Obama served three terms in the Illinois Senate from 1997 to 2004. Following an unsuccessful bid against a Democratic incumbent for a seat in the U.S. House of Representatives in 2000, he ran for United States Senate in 2004.[4] Several events brought him to national attention during the campaign, including his victory in the March 2004 Democratic primary and his keynote address at the Democratic National Convention in July 2004. He won election to the U.S. Senate in November 2004. His presidential campaign began in February 2007, and after a close campaign in the 2008 Democratic Party presidential primaries against Hillary Rodham Clinton, he won his party's nomination. In the 2008 general election, he defeated Republican nominee John McCain and was inaugurated as president on January 20, 2009.

As president, Obama signed economic stimulus legislation in the form of the American Recovery and Reinvestment Act in February 2009. Other domestic policy initiatives include the Patient Protection and Affordable Care Act – a major piece of health care reform legislation which he signed into law in March 2010 – and the Dodd–Frank Wall Street Reform and Consumer Protection Act, which forms part of his financial regulatory reform efforts, which he signed in July 2010. In foreign policy, Obama gradually withdrew combat troops from Iraq, increased troop levels in Afghanistan, and signed an arms control treaty with Russia. On October 9, 2009, Obama was named the 2009 Nobel Peace Prize laureate.

Barack Obama

44th President of the United States

Incumbent

Assumed office
January 20, 2009

Vice President Joe Biden
Preceded by George W. Bush

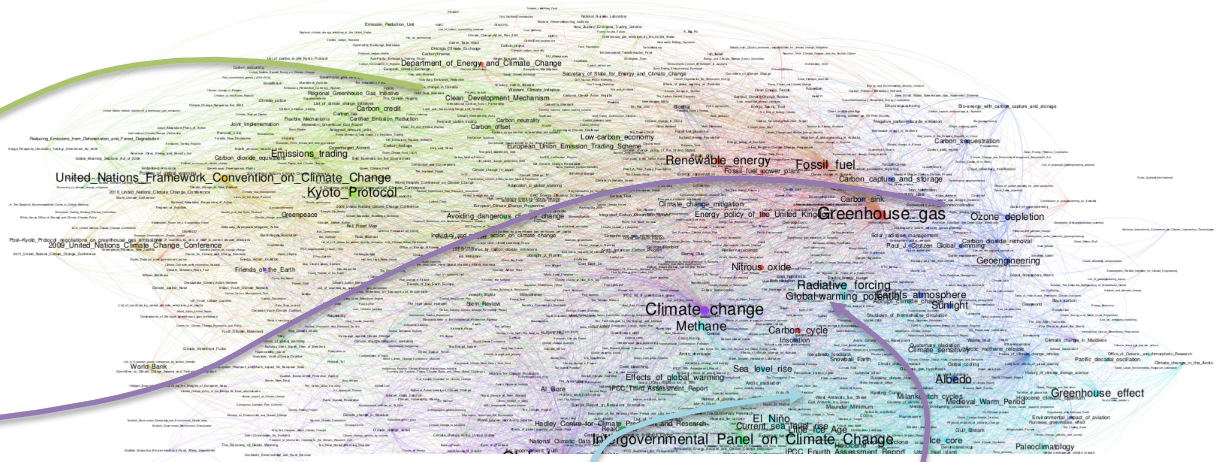
United States Senator from Illinois

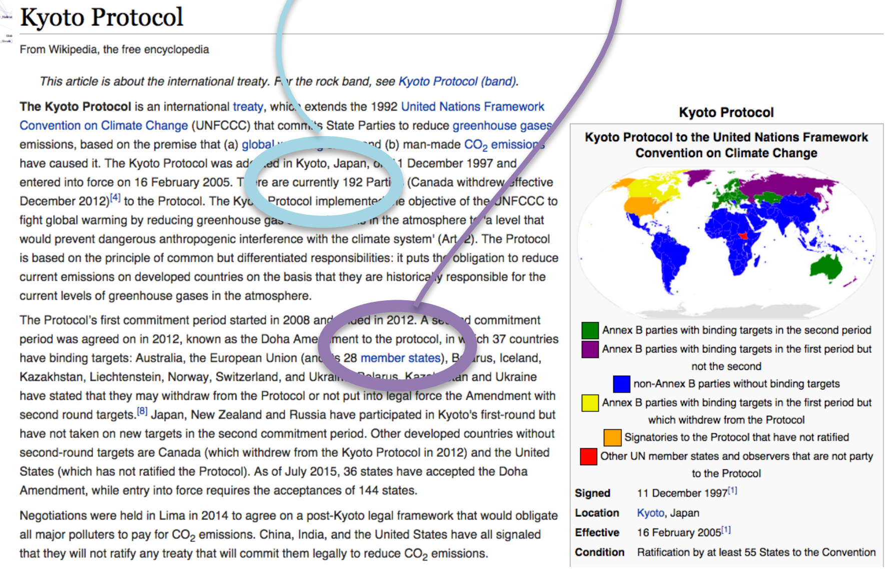
In office
January 3, 2005 – November 16, 2008

Preceded by Peter Fitzgerald
Succeeded by Roland Burris

Member of the Illinois Senate from the 13th district

In office





Kyoto Protocol

From Wikipedia, the free encyclopedia

This article is about the international treaty. For the rock band, see Kyoto Protocol (band).

The Kyoto Protocol is an international treaty, which extends the 1992 United Nations Framework Convention on Climate Change (UNFCCC) that commits State Parties to reduce greenhouse gases emissions, based on the premise that (a) global warming exists and (b) man-made CO_2 emissions have caused it. The Kyoto Protocol was adopted in Kyoto, Japan, on 11 December 1997 and entered into force on 16 February 2005. There are currently 192 Parties (Canada withdrew effective December 2012)[4] to the Protocol. The Kyoto Protocol implements the objective of the UNFCCC to fight global warming by reducing greenhouse gas concentrations in the atmosphere to 'a level that would prevent dangerous anthropogenic interference with the climate system' (Art. 2). The Protocol is based on the principle of common but differentiated responsibilities: it puts the obligation to reduce current emissions on developed countries on the basis that they are historically responsible for the current levels of greenhouse gases in the atmosphere.

The Protocol's first commitment period started in 2008 and ended in 2012. A second commitment period was agreed on in 2012, known as the Doha Amendment to the protocol, in which 37 countries have binding targets: Australia, the European Union (and its 28 member states), Belarus, Iceland, Kazakhstan, Liechtenstein, Norway, Switzerland, and Ukraine. Belarus, Kazakhstan and Ukraine have stated that they may withdraw from the Protocol or not put into legal force the Amendment with second round targets.[8] Japan, New Zealand and Russia have participated in Kyoto's first-round but have not taken on new targets in the second commitment period. Other developed countries without second-round targets are Canada (which withdrew from the Kyoto Protocol in 2012) and the United States (which has not ratified the Protocol). As of July 2015, 36 states have accepted the Doha Amendment, while entry into force requires the acceptances of 144 states.

Negotiations were held in Lima in 2014 to agree on a post-Kyoto legal framework that would obligate all major polluters to pay for CO_2 emissions. China, India, and the United States have all signaled that they will not ratify any treaty that will commit them legally to reduce CO_2 emissions.

Kyoto Protocol

Kyoto Protocol to the United Nations Framework Convention on Climate Change

🟩 Annex B parties with binding targets in the second period
🟪 Annex B parties with binding targets in the first period but not the second
🟦 non-Annex B parties without binding targets
🟨 Annex B parties with binding targets in the first period but which withdrew from the Protocol
🟧 Signatories to the Protocol that have not ratified
🟥 Other UN member states and observers that are not party to the Protocol

Signed	11 December 1997[1]
Location	Kyoto, Japan
Effective	16 February 2005[1]
Condition	Ratification by at least 55 States to the Convention

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)

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

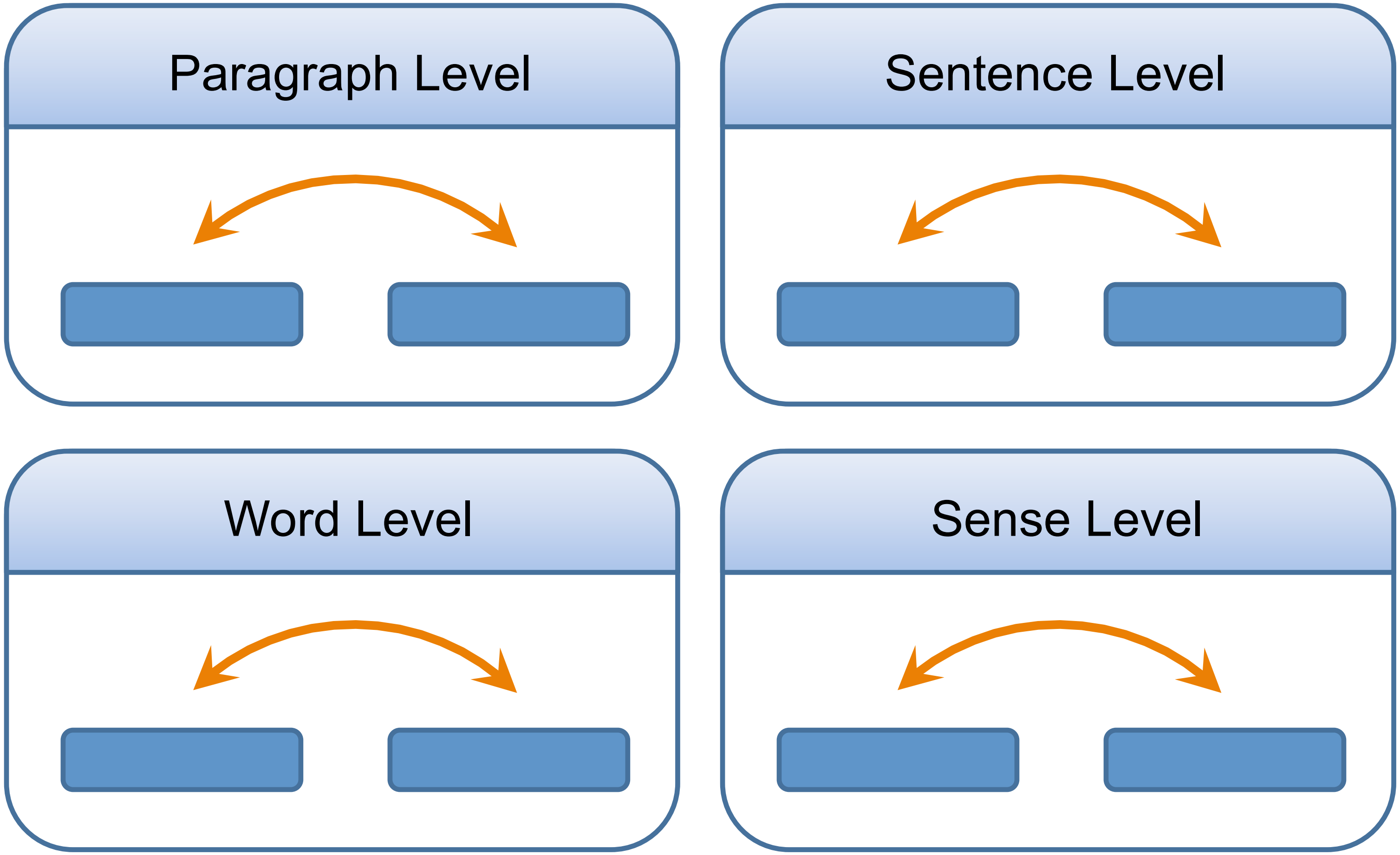
- Non-negative Matrix Factorization (Xu et al., 2003)
- Concept Factorization (Xu and Gong, 2004)
- Locally-Consistent Concept Factorization (Cai et al. 2011)
  - Non-linear dimensionality reduction

Main issues are computational complexity and representational opaqueness

Cross-Level Semantic Similarity

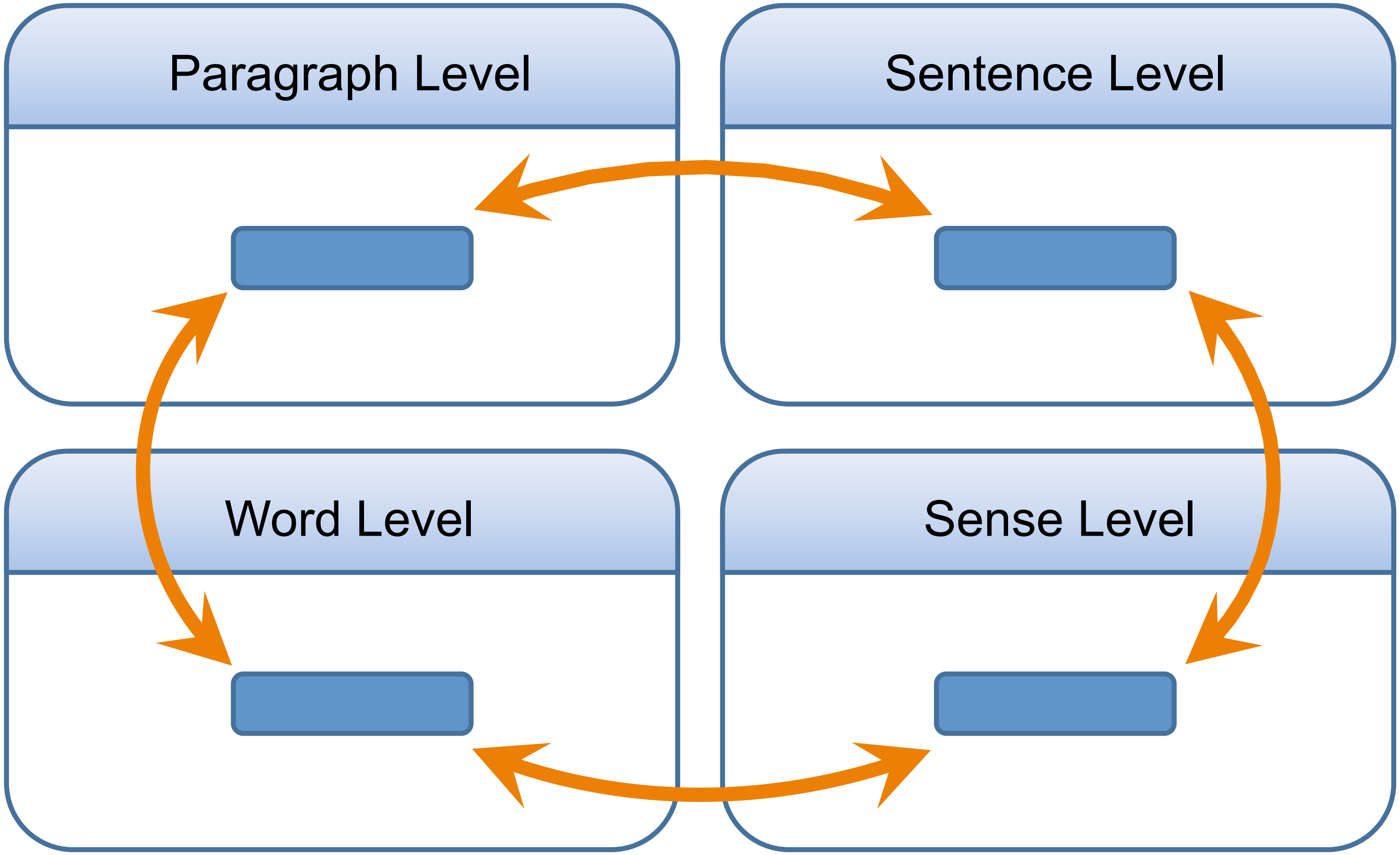
Semantic Similarity

Mostly focused on similar types of lexical items



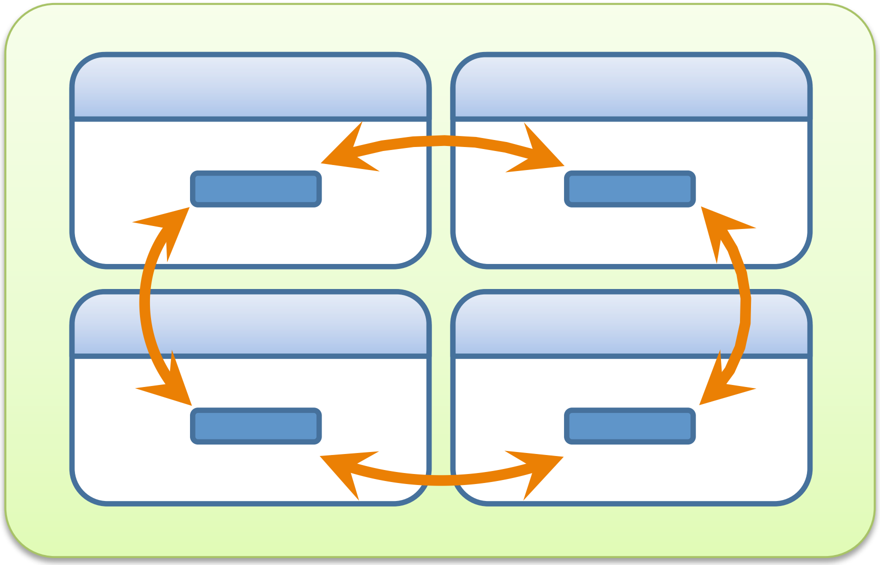
Semantic Similarity

What if we have different types of inputs?

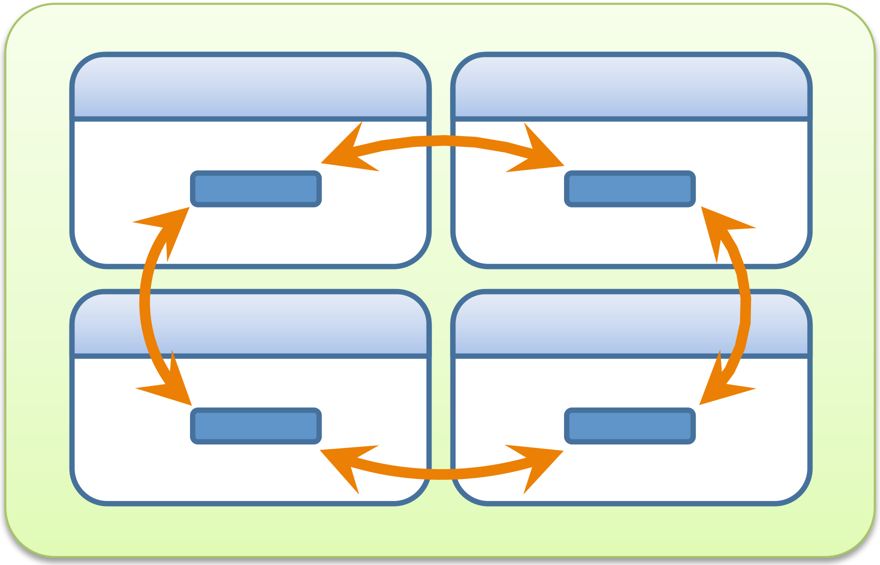


CLSS: Cross-Level Semantic Similarity

A new type of similarity task



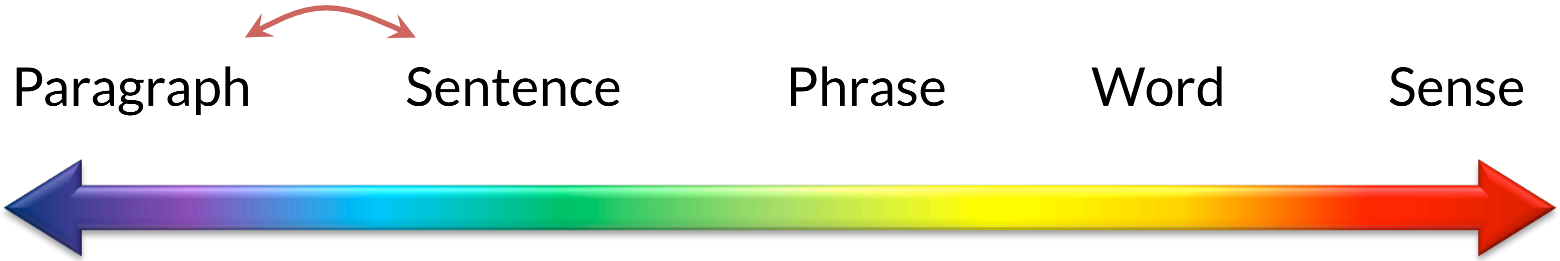
CLSS: Cross-Level Semantic Similarity



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

CLSS: Comparison Types

Paragraph to Sentence



Paragraph Sentence Phrase Word Sense

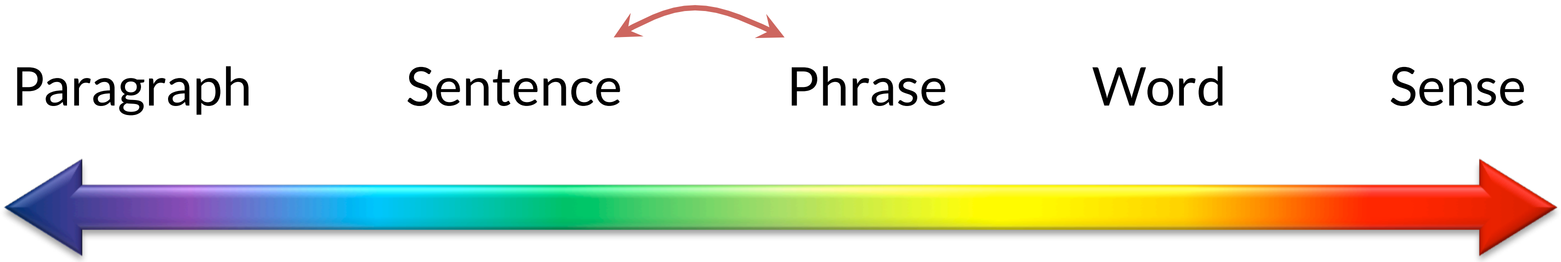
CLSS: Comparison Types

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

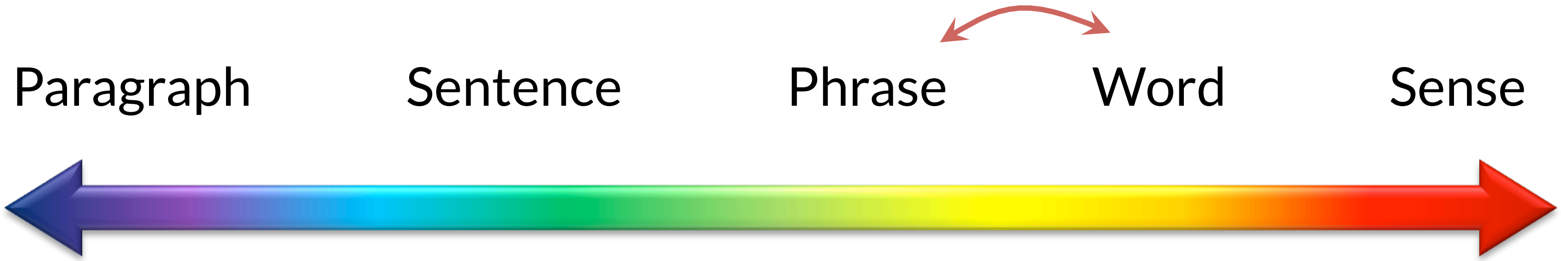
Paragraph to Sentence

Sentence to Phrase

Phrase to Word

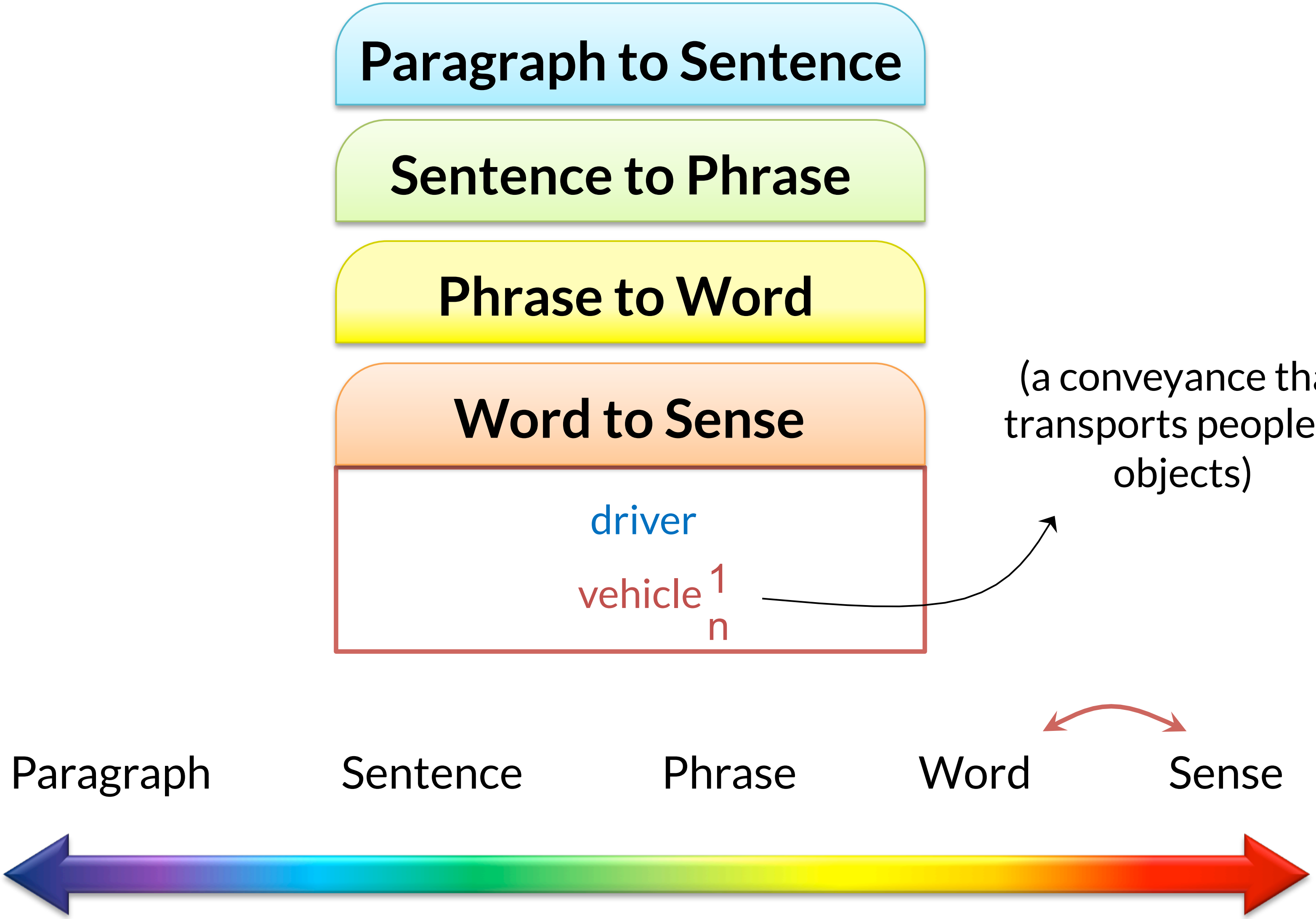
a large, expensive house

mansion



Paragraph Sentence Phrase Word Sense

CLSS: Comparison Types



Paragraph to Sentence

Sentence to Phrase

Phrase to Word

Word to Sense

driver

vehicle$^{1}_{n}$

(a conveyance that transports people or objects)

Paragraph Sentence Phrase Word Sense

Task Data

4000 pairs in total

500 pairs per type

- Paragraph to Sentence
- Sentence to Phrase
- Phrase to Word
- Word to Sense

Training set

500 pairs per type

- Paragraph to Sentence
- Sentence to Phrase
- Phrase to Word
- Word to Sense

Test set

Task Data

A wide range of domains and text styles

Paragraph to Sentence

- Newswire
- Travel
- Question Answering
- Scientific
- Metaphoric
- Review

Sentence to Phrase

- Newswire
- Travel
- Question Answering
- Scientific
- Slang
- Idiomatic

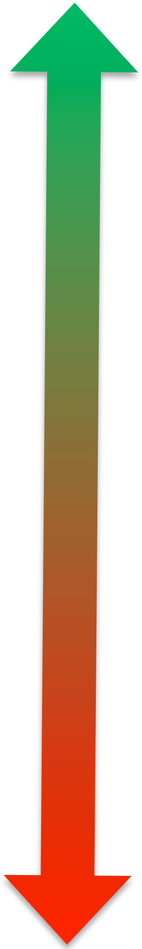
Phrase to Word

- Newswire
- Slang
- Idiomatic
- Lexicographic
- Descriptive
- 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

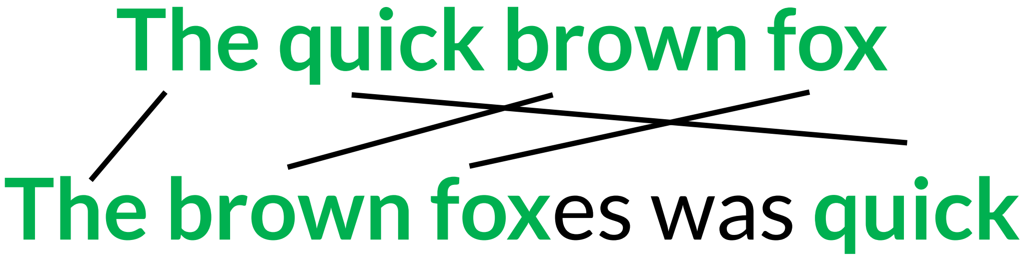
Comparison Baselines

- Longest Common Substring (LCS)

The quick **brown fox**

The **brown fox** was quick

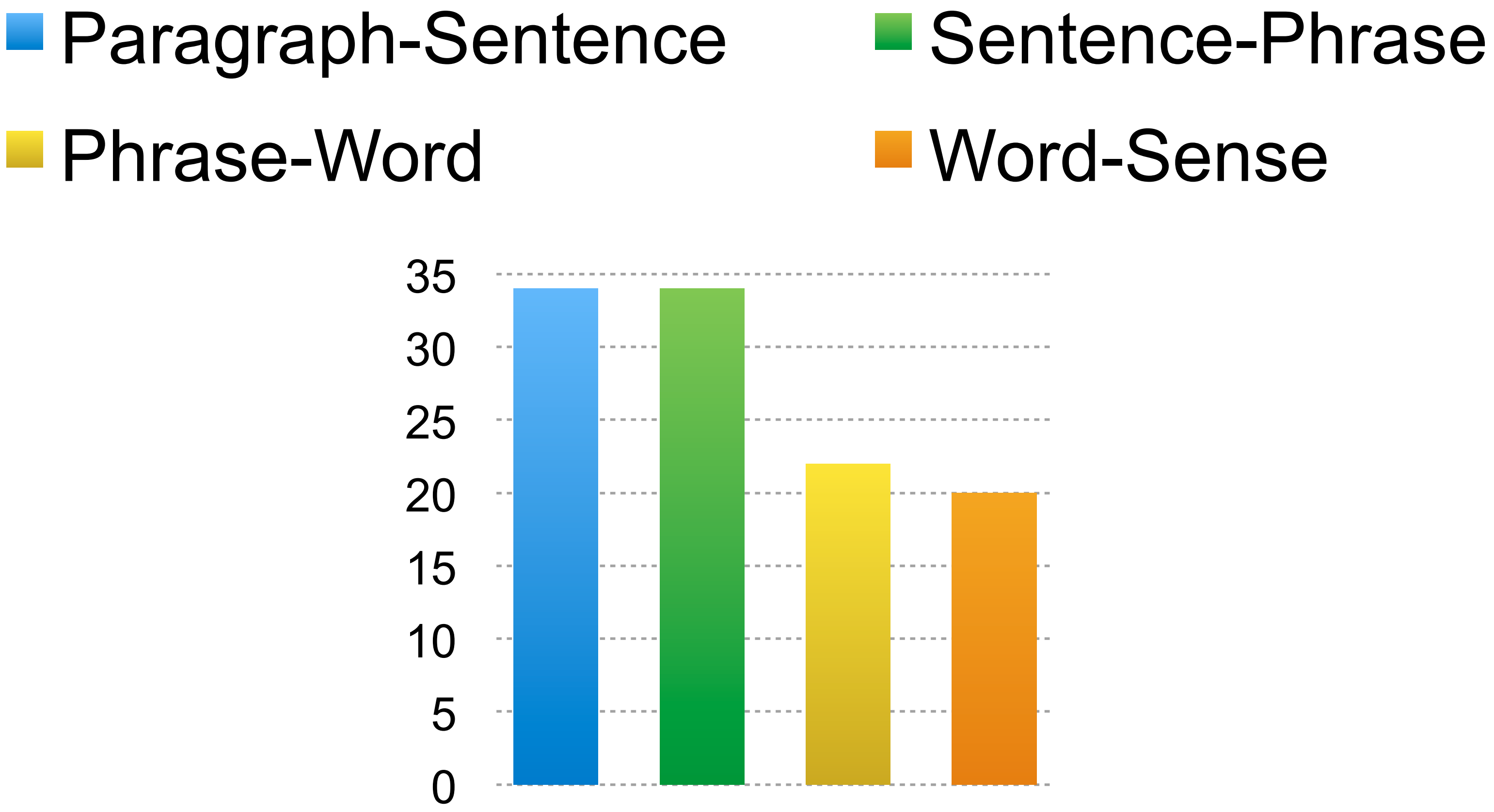
- Greedy String Tiling (GST)



The quick brown fox

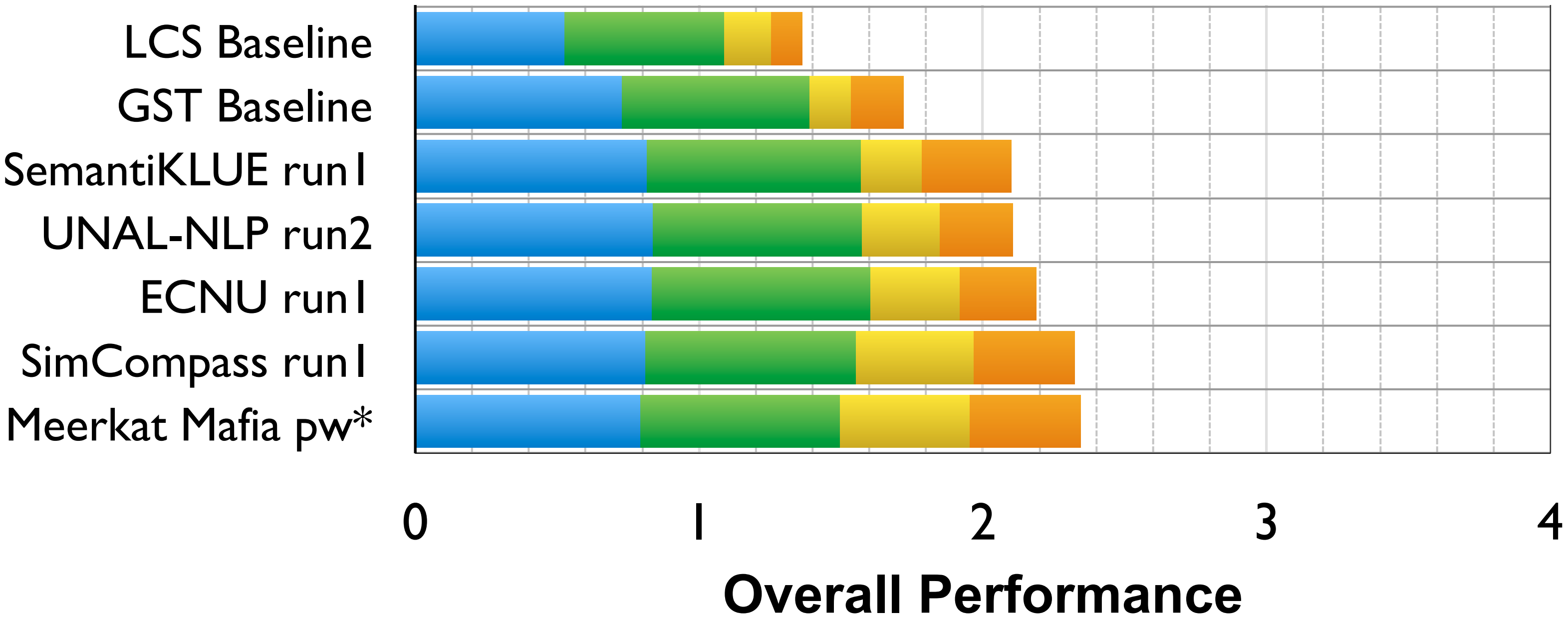
The brown foxes was **quick**

Number of participants



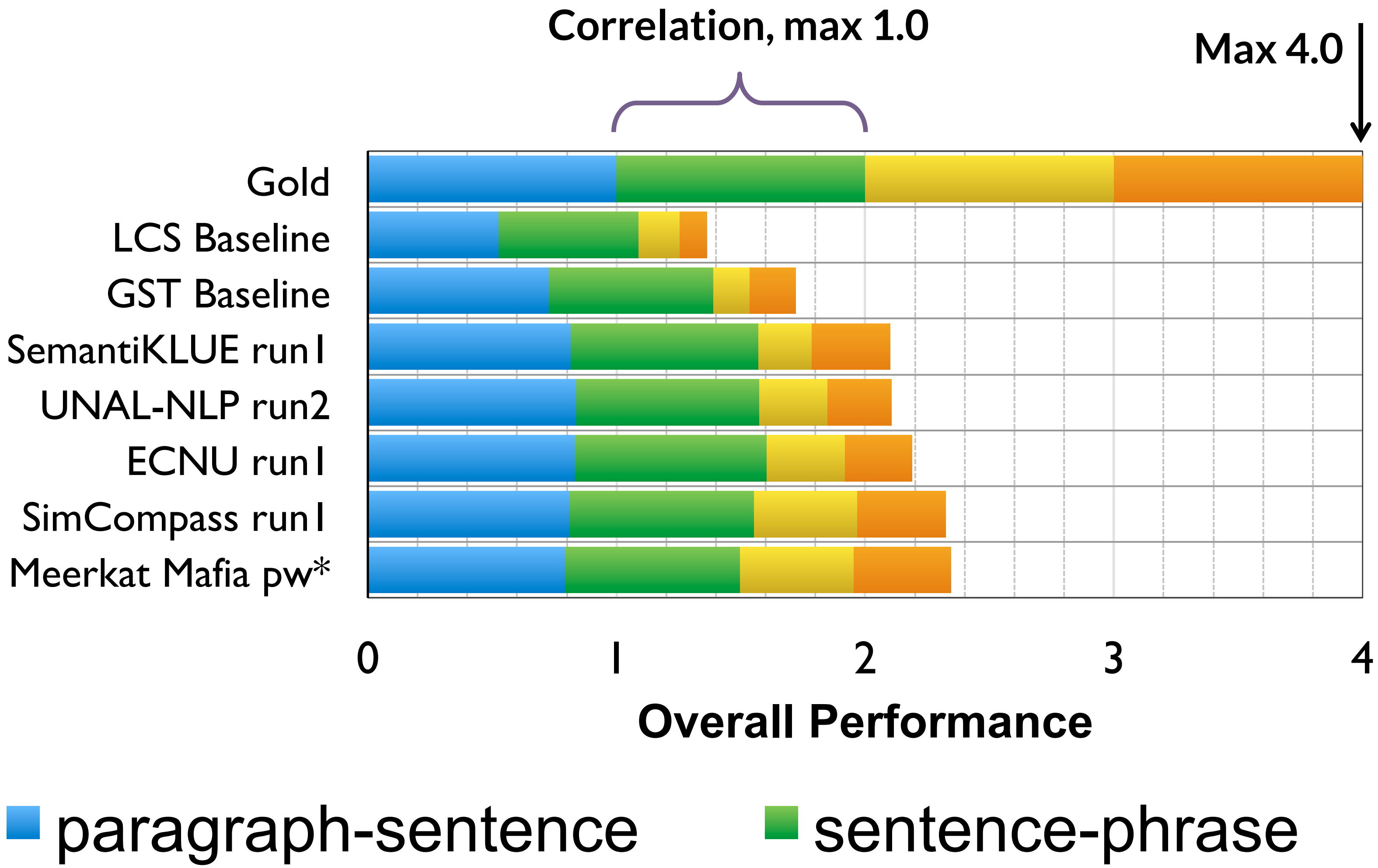
38 Systems total from 19 teams

Top 5 Systems and Baselines



🟦 paragraph-sentence 🟩 sentence-phrase

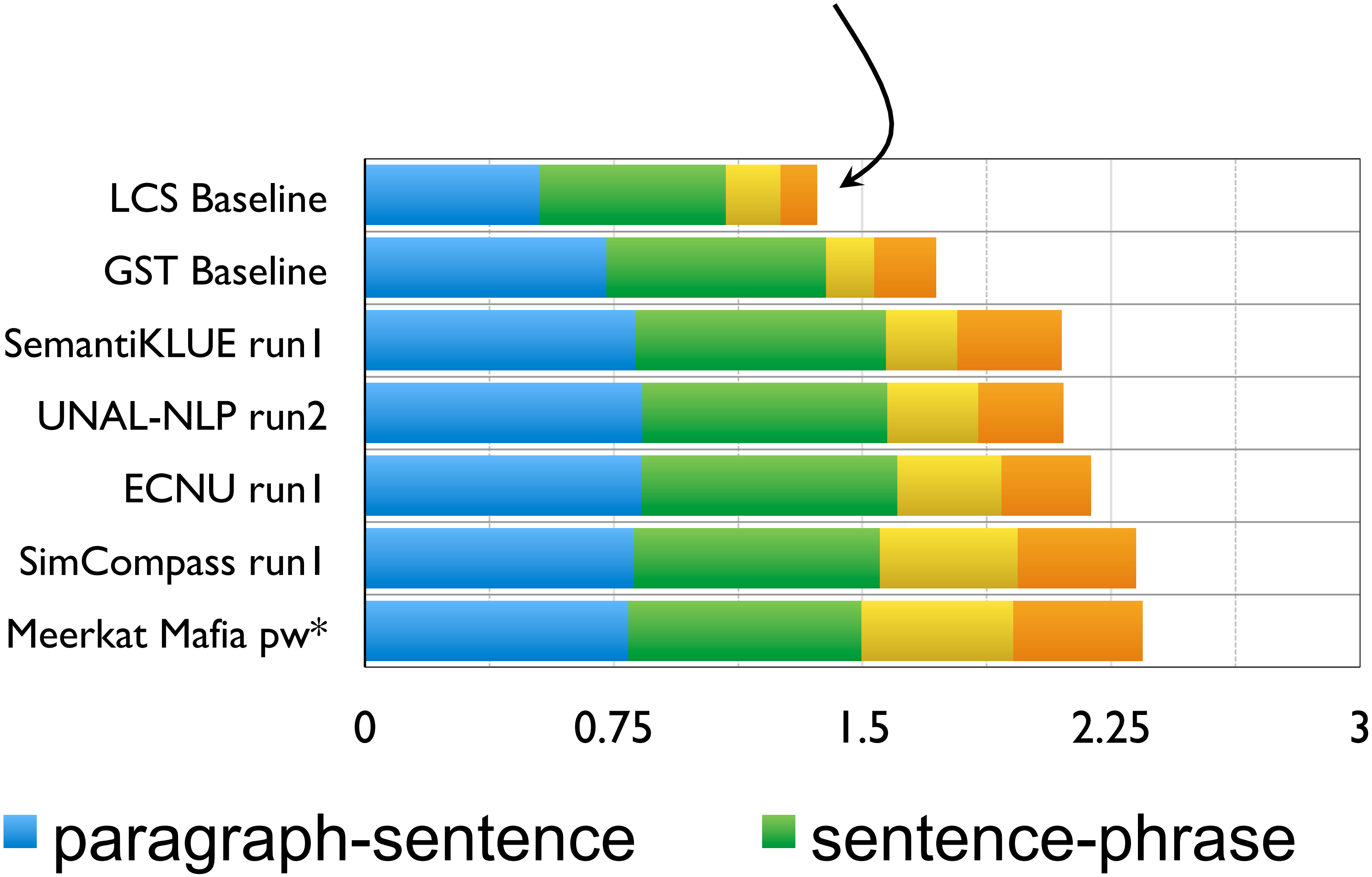
Top 5 Systems and Baselines



Correlation, max 1.0

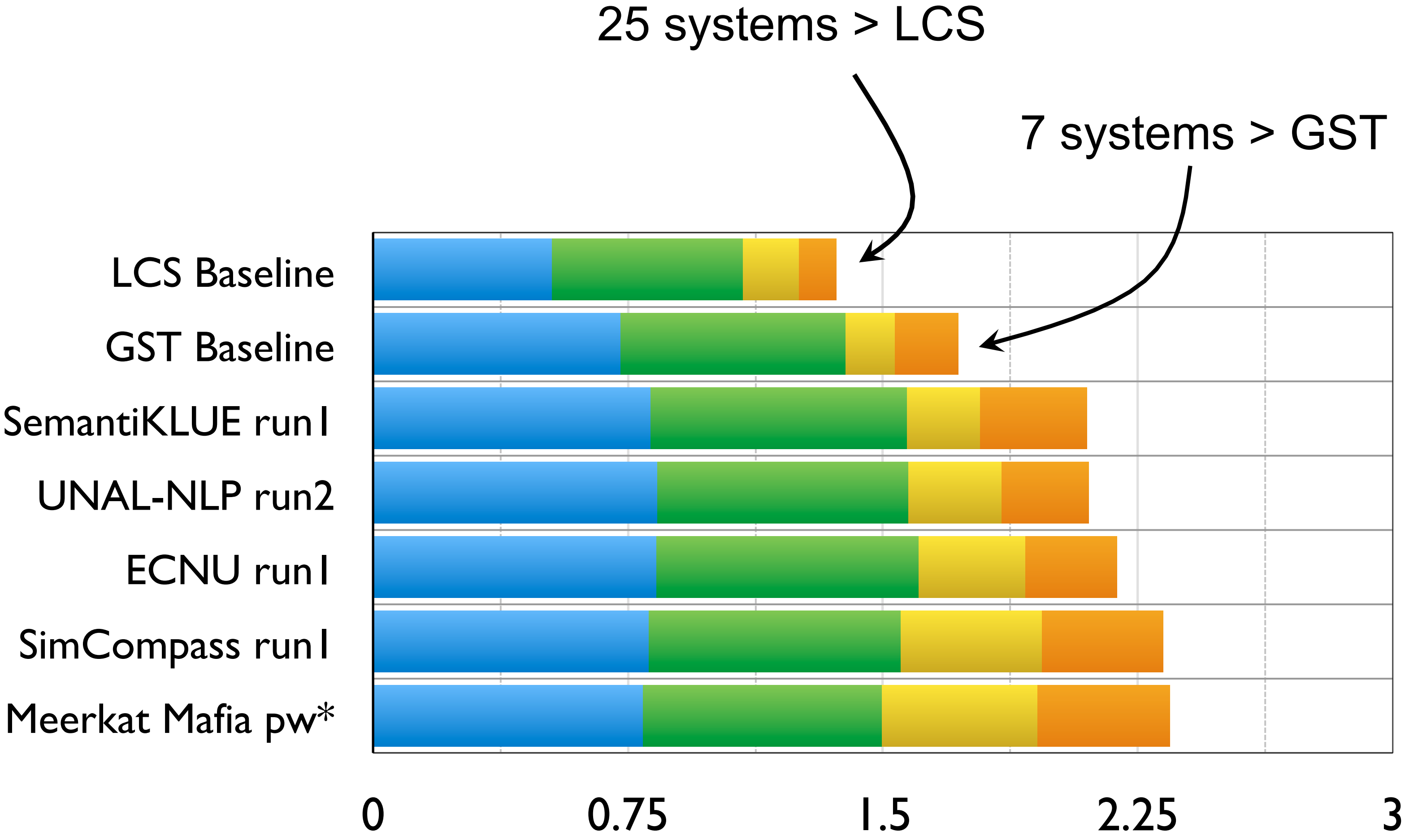
Max 4.0

Where do the baselines stand?



25 systems > LCS

Where do the baselines stand?



25 systems > LCS

7 systems > GST

🟦 paragraph-sentence 🟩 sentence-phrase

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.

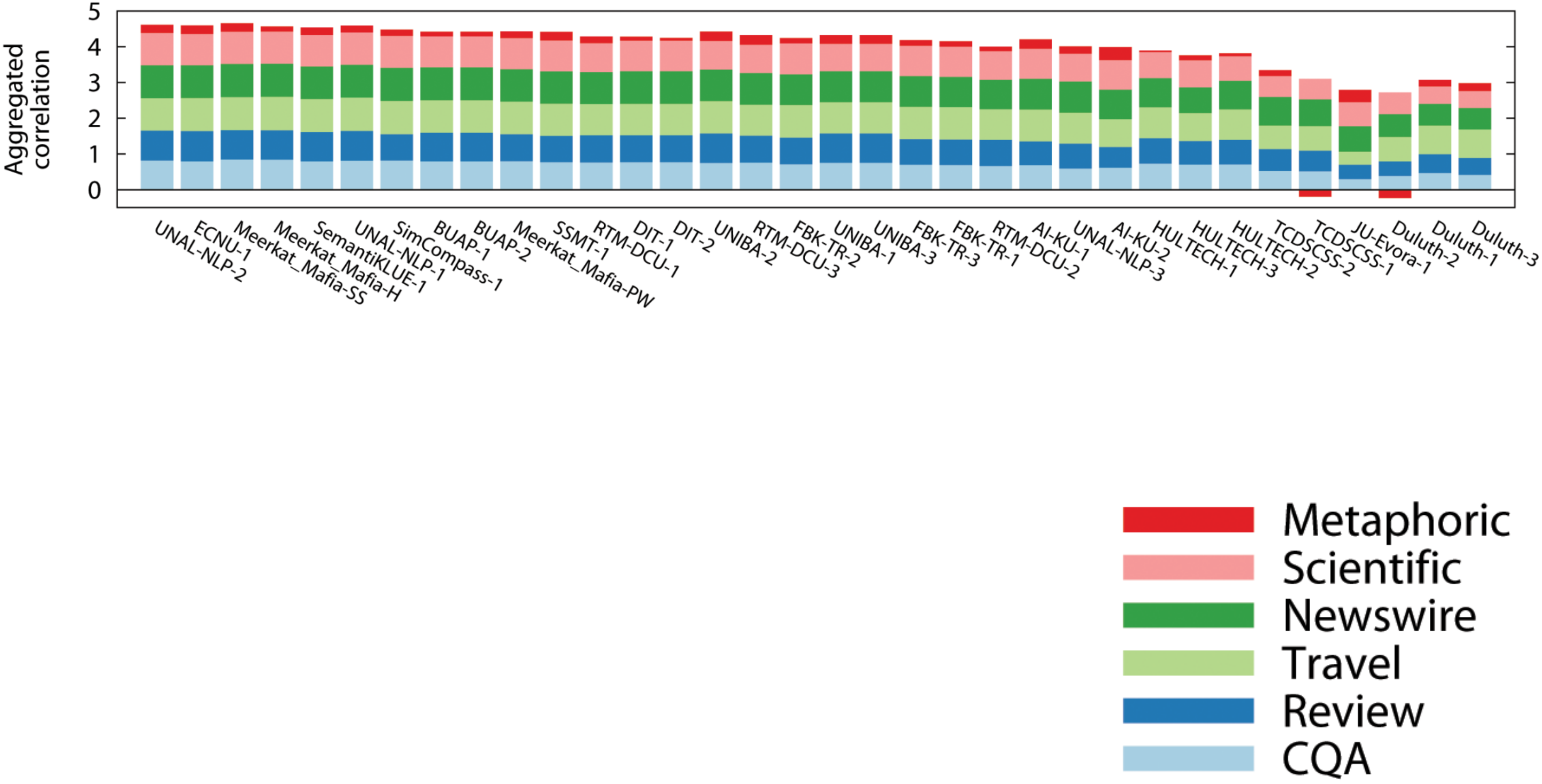
UNAL-NLP - Jimenez et al (2014)

Third best system overall

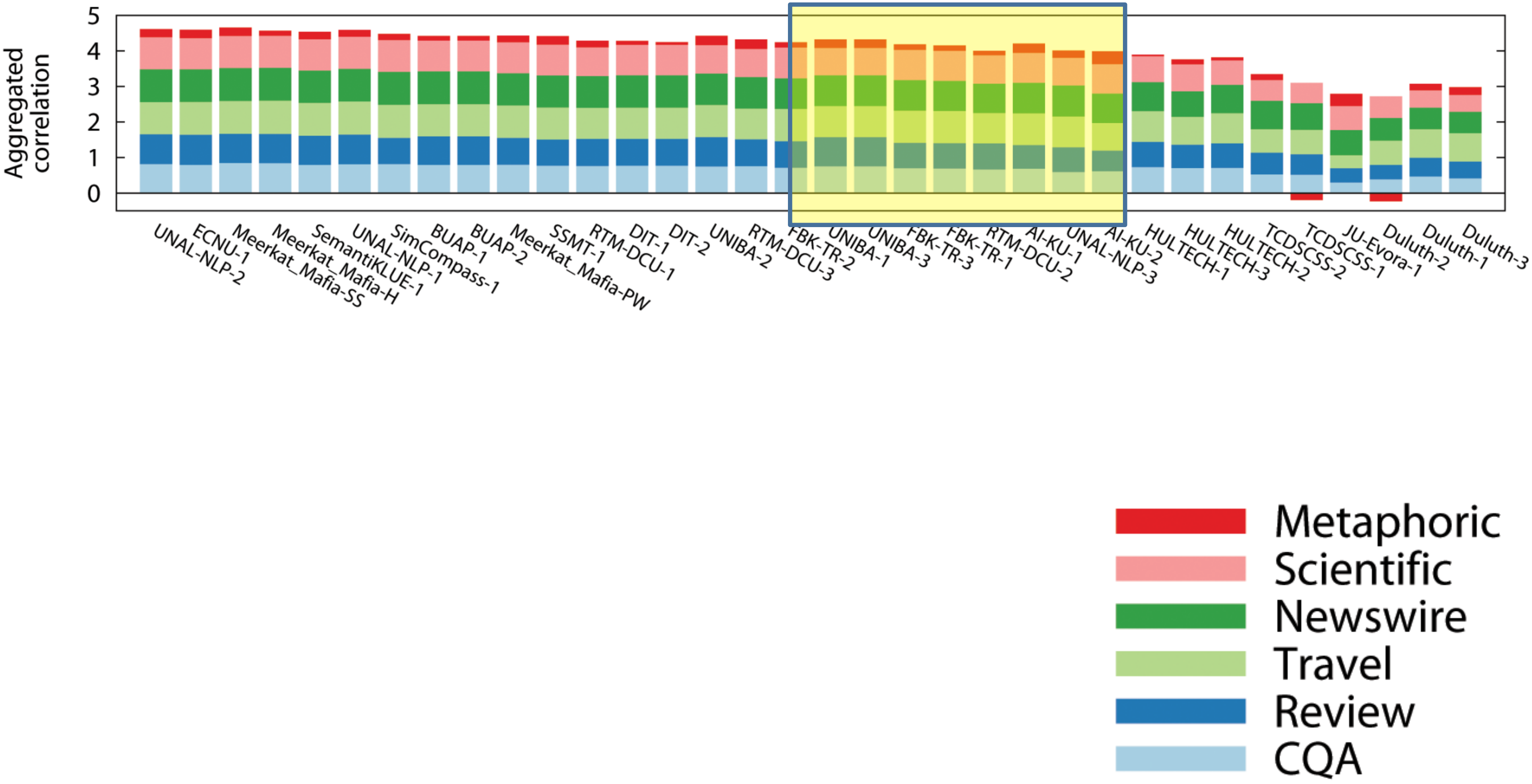
Utilizes only a set of simple **string-similarity features** based on soft cardinality (Jimenez et al, 2010).

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).

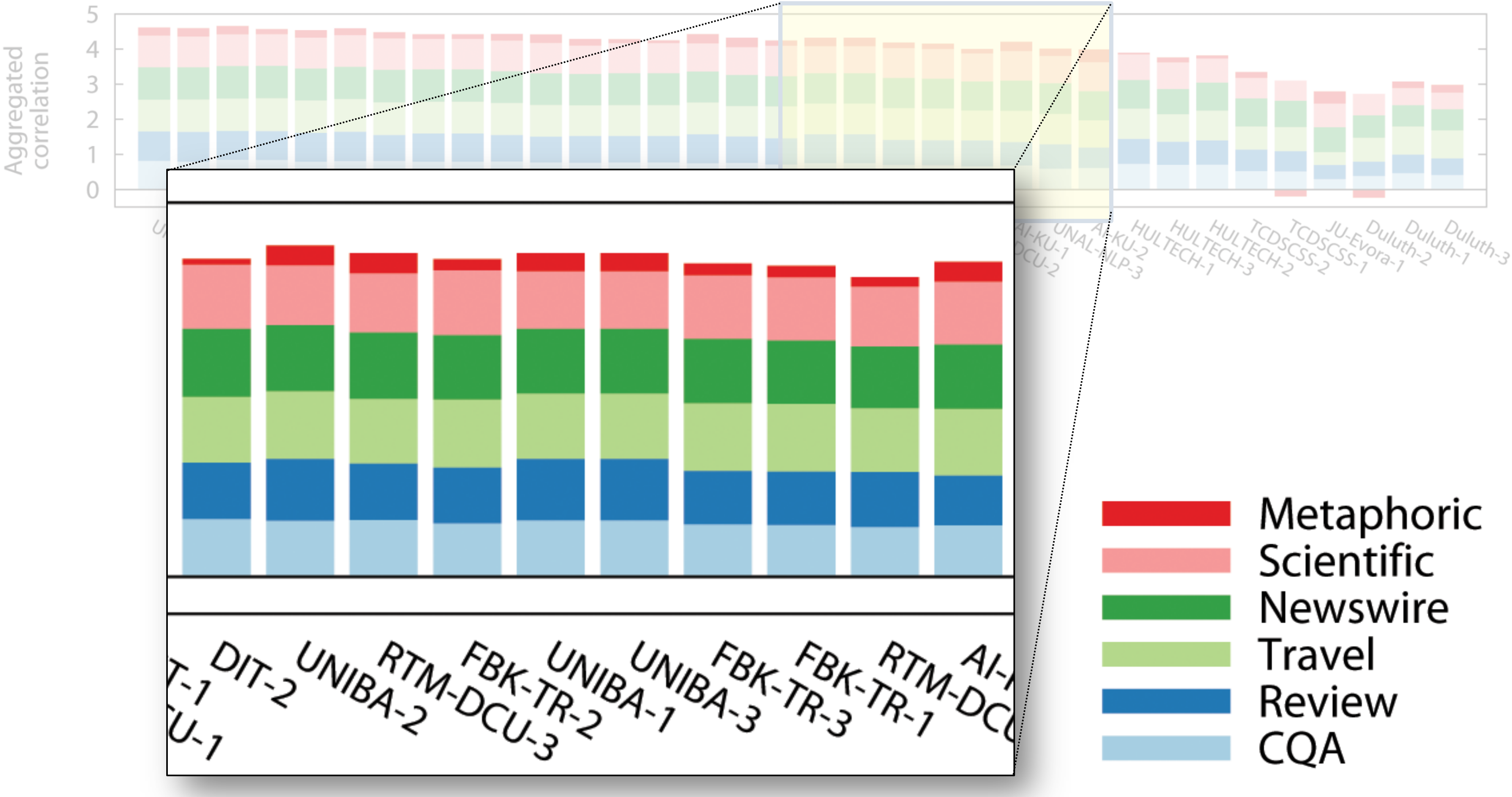
Correlation per genre
paragraph-to-sentence



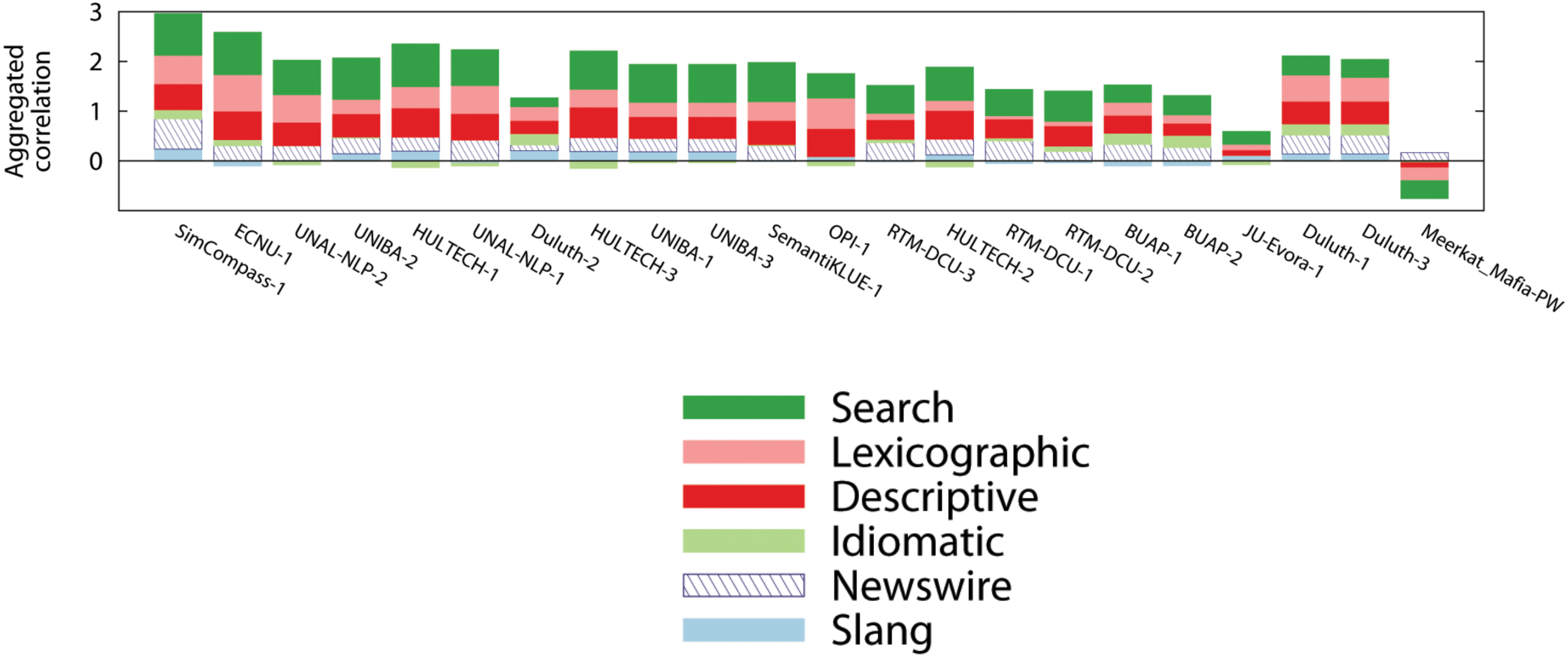
Correlation per genre
paragraph-to-sentence



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?

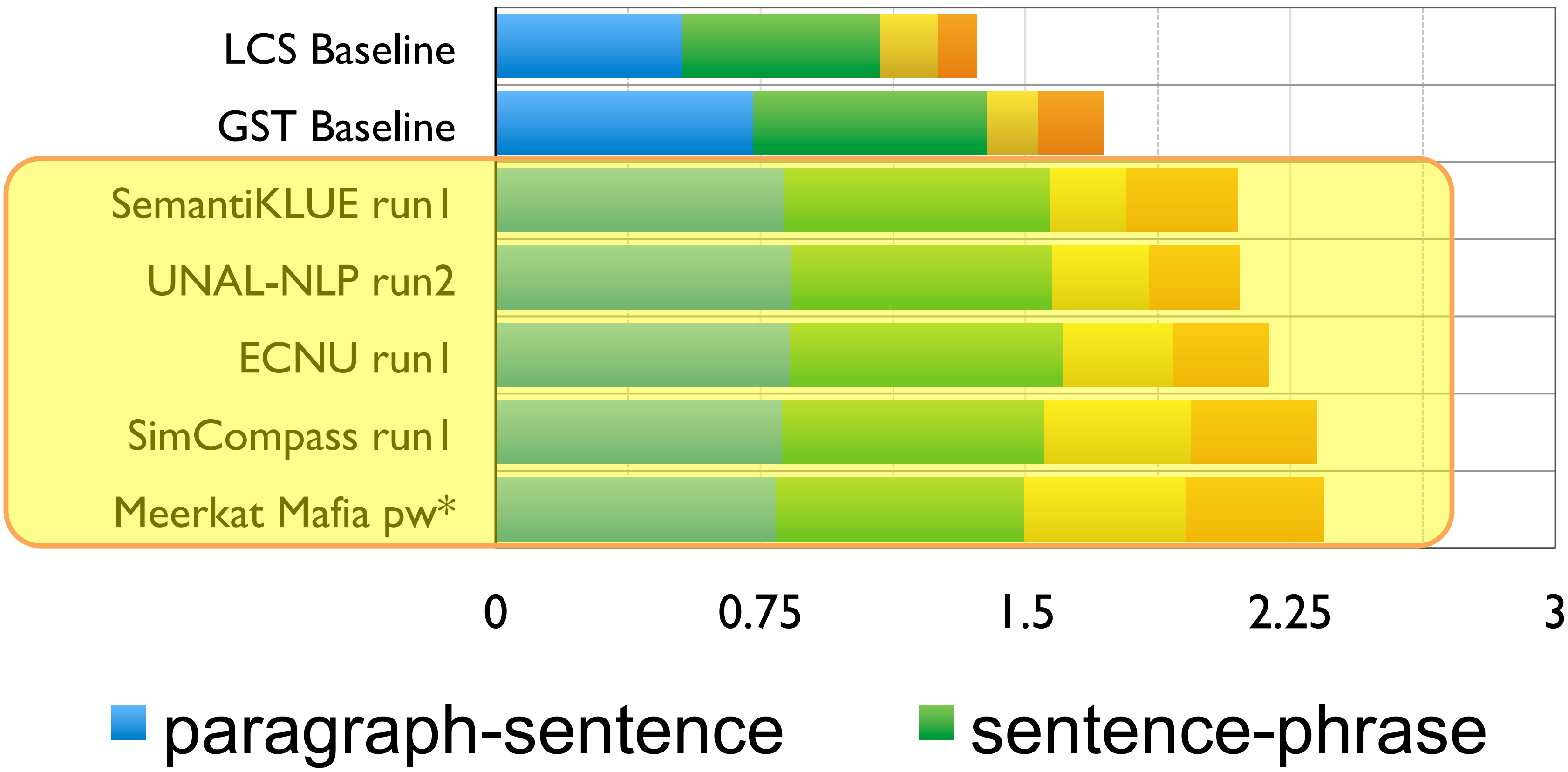
frequency of draik eggs in Merifoods

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

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).

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

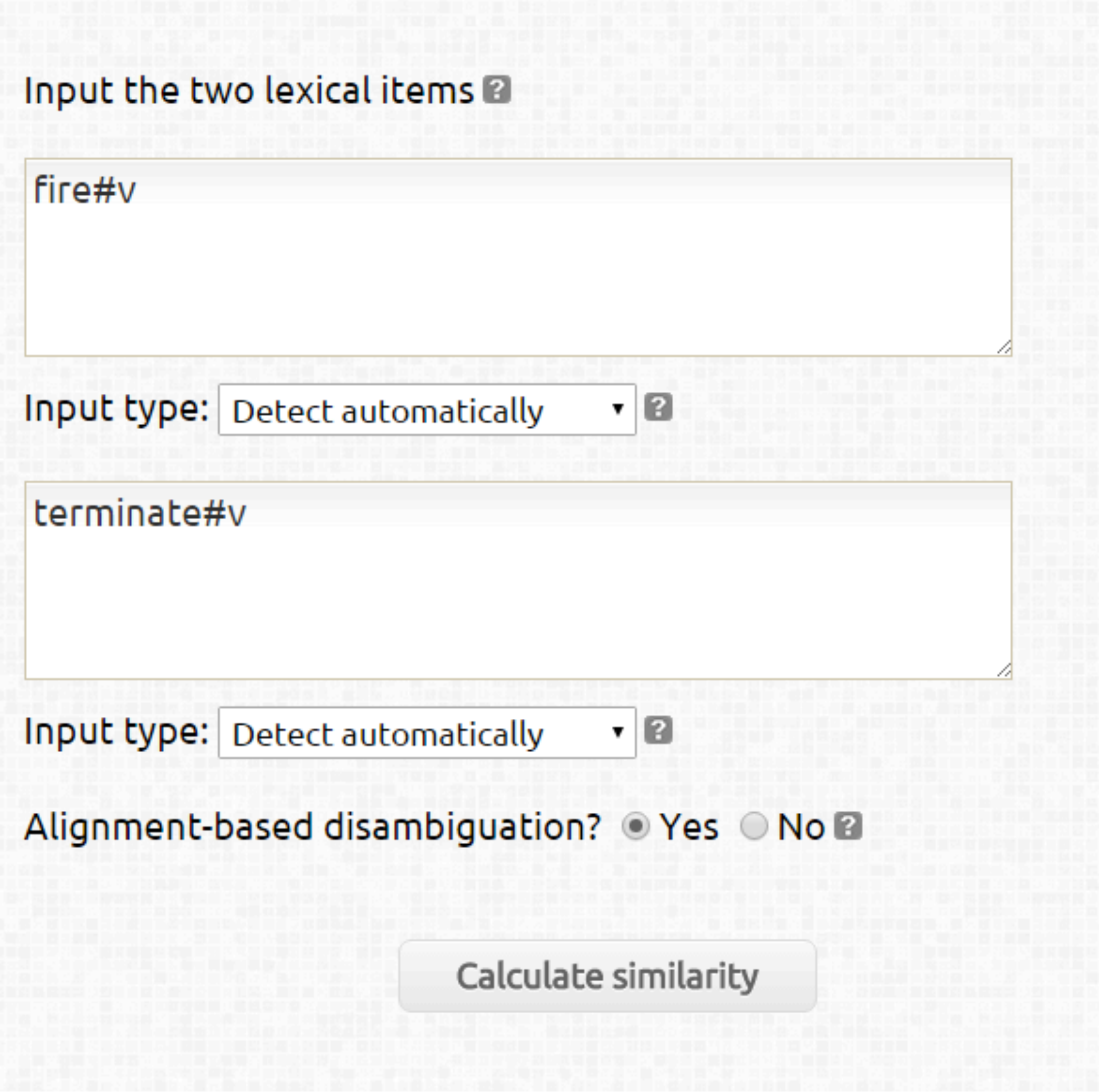
https://github.com/pilehvar/adw

Tools

Align, Disambiguate and Walk: ADW

Online demo at

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



Input the two lexical items

fire#v

Input type: Detect automatically

terminate#v

Input type: Detect automatically

Alignment-based disambiguation? ● Yes ○ No

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

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

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

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

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/

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

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

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

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

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

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

- Highlights include:
  - LDA in C (fast!)
  - HDP in C
  - TurboTopics in Python

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/

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

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)

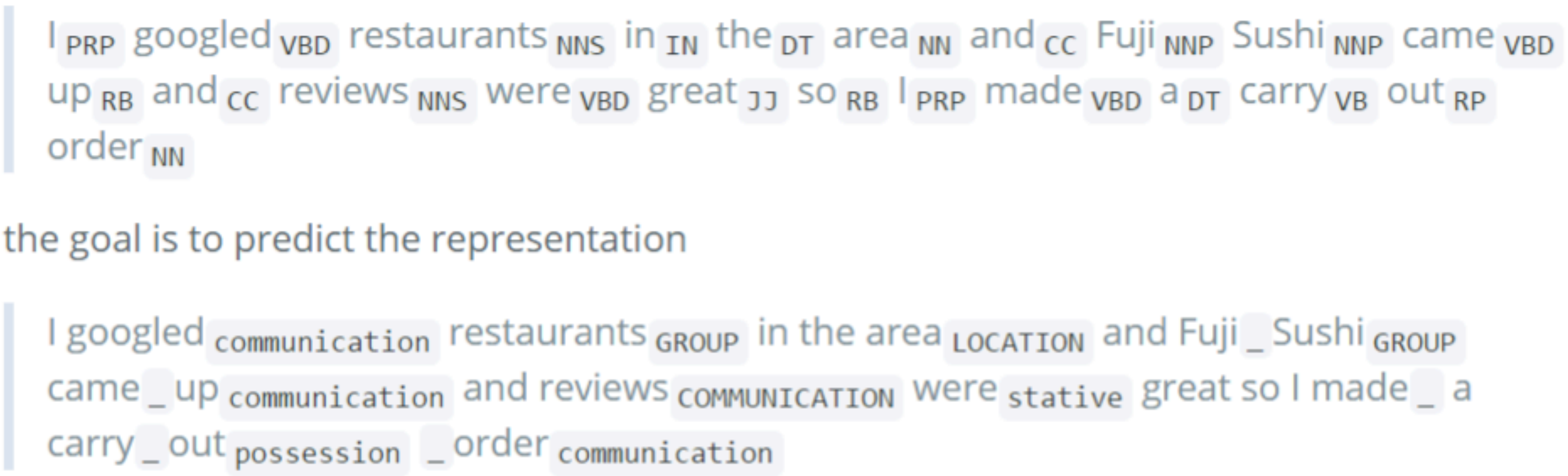
- Most approaches either ignore MWEs or recognize those from fixed lists of MWEs
  - Problematic unless lemmatizing
  - Even more problematic with syntactic rearrangement

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

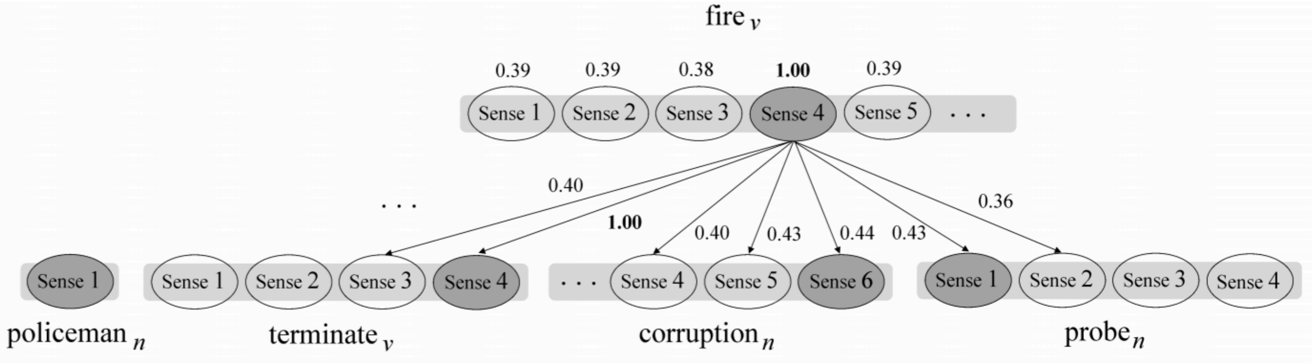
The boss **fired** his worker.

An employee was
terminated
from work by his boss.

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

Open Problem: Ambiguity

- Alignment-based disambiguation of ADW

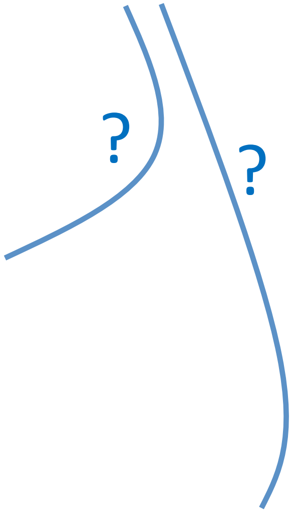


- WSD is a solution, but is still a long way off

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

EMNLP 2015
CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING
LISBON

erc
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