A Word at a Time: Computing Word Relatedness using

Temporal Semantic Analysis

Kira Radinsky (Technion)
Eugene Agichtein (Emory)
Evgeniy Gabrilovich (Yahoo! Research)
Shaul Markovitch (Technion)

Presented by **Andrei Broder**

Semantic relatedness of texts

Given two texts, quantify their semantic relatedness

Used in many NLP tasks:

- Information retrieval
- Word-sense disambiguation
- Text clustering
- Error correction

Ontologies and concepts

An ontology is a collection of concepts, for example:

1. **Wikipedia** as an **ontology**
   - Every Wikipedia article represents a **concept**
   - A **word** (or longer text fragment) can be represented as a vector of related Wikipedia **concepts** (using ESA)
2. **Flickr** as an **ontology**
   - Every Flickr tag represents a **concept**
   - A **word** can be represented as a vector of co-occurring Flickr **tags**

Current state of the art

(Concept-based representations)

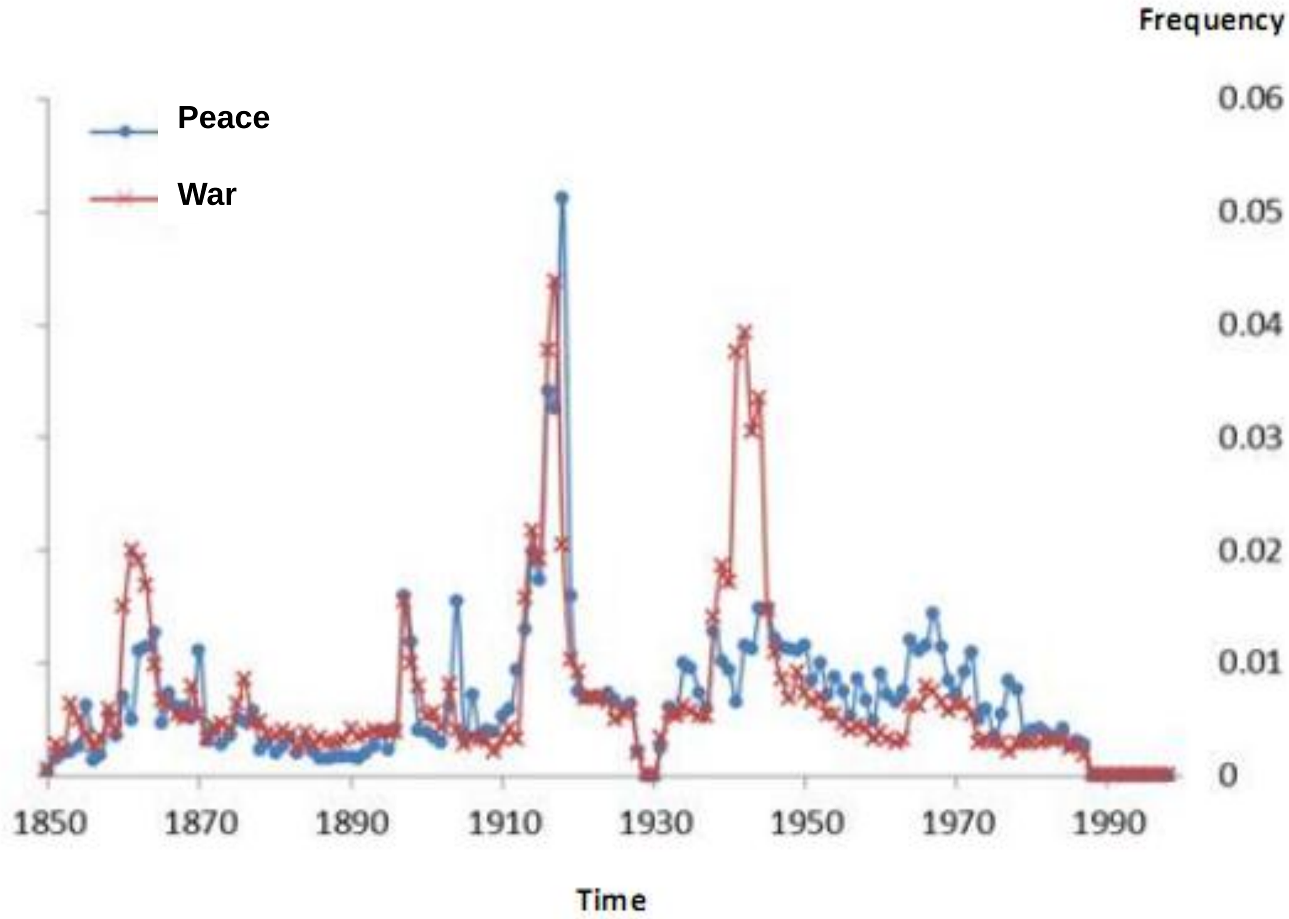
- Path-based measures using Wikipedia categories
  WikiRelate! (Strube,2006)
- Co-occurrence based measures
  Latent Semantic Analysis (Deerwester et al., 1990)
- WordNet-based measures
  Multiple measures formulated in the literature(see Budanitsky & Hirst, 2001, for a comprehensive review)
- Vector space models
  Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007)
  In ESA, a fragment of text is represented as a weighted vector of Wikipedia concepts.

All these approaches are based on a **static** corpus.

Can the **temporal** dynamics observed in a corpus be used to enhance text relatedness models?

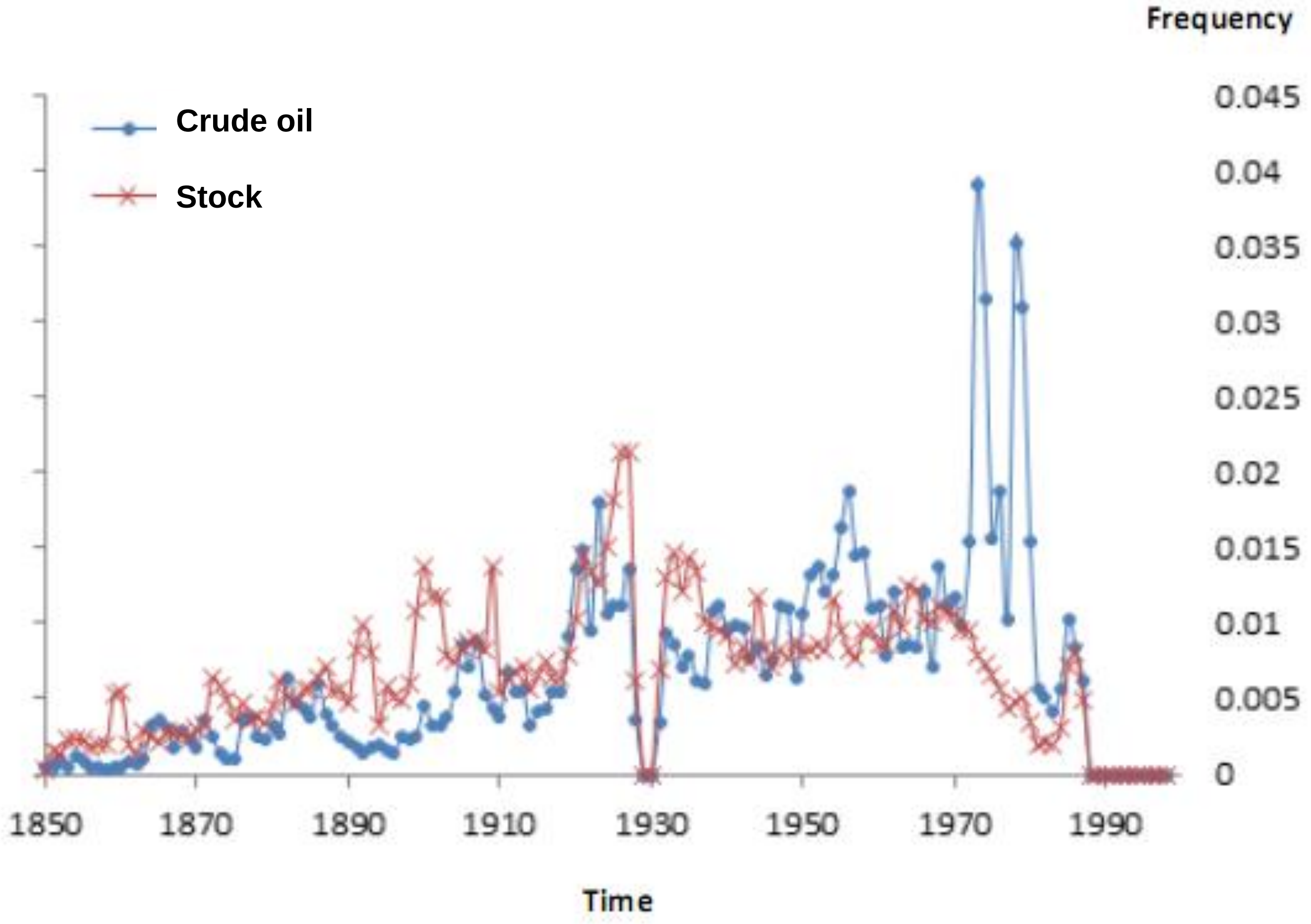
Intuition

Temporal co-appearance of "war" and "peace" in NYT archives 1850-2009



Intuition

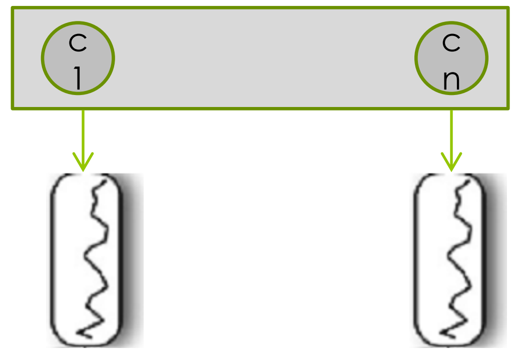
Temporal co-appearance of "crude oil" and "stock" in NYT archives 1850-2009



Overview: Temporal semantic analysis

1. Novel temporal representation of text

Word



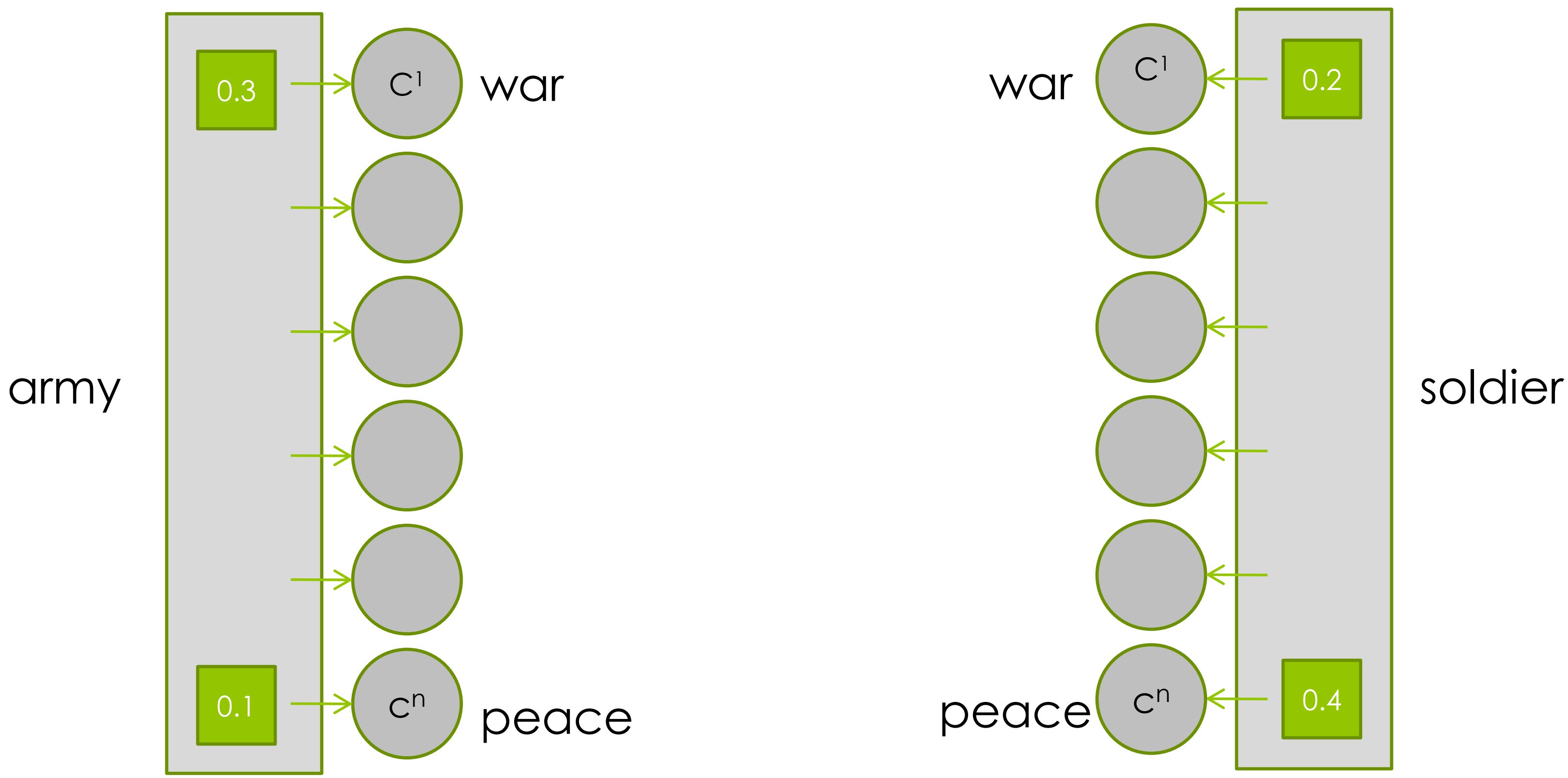
Represent words as concept vectors

Extend static representation with temporal dynamics

2. Novel temporal text-similarity measurement

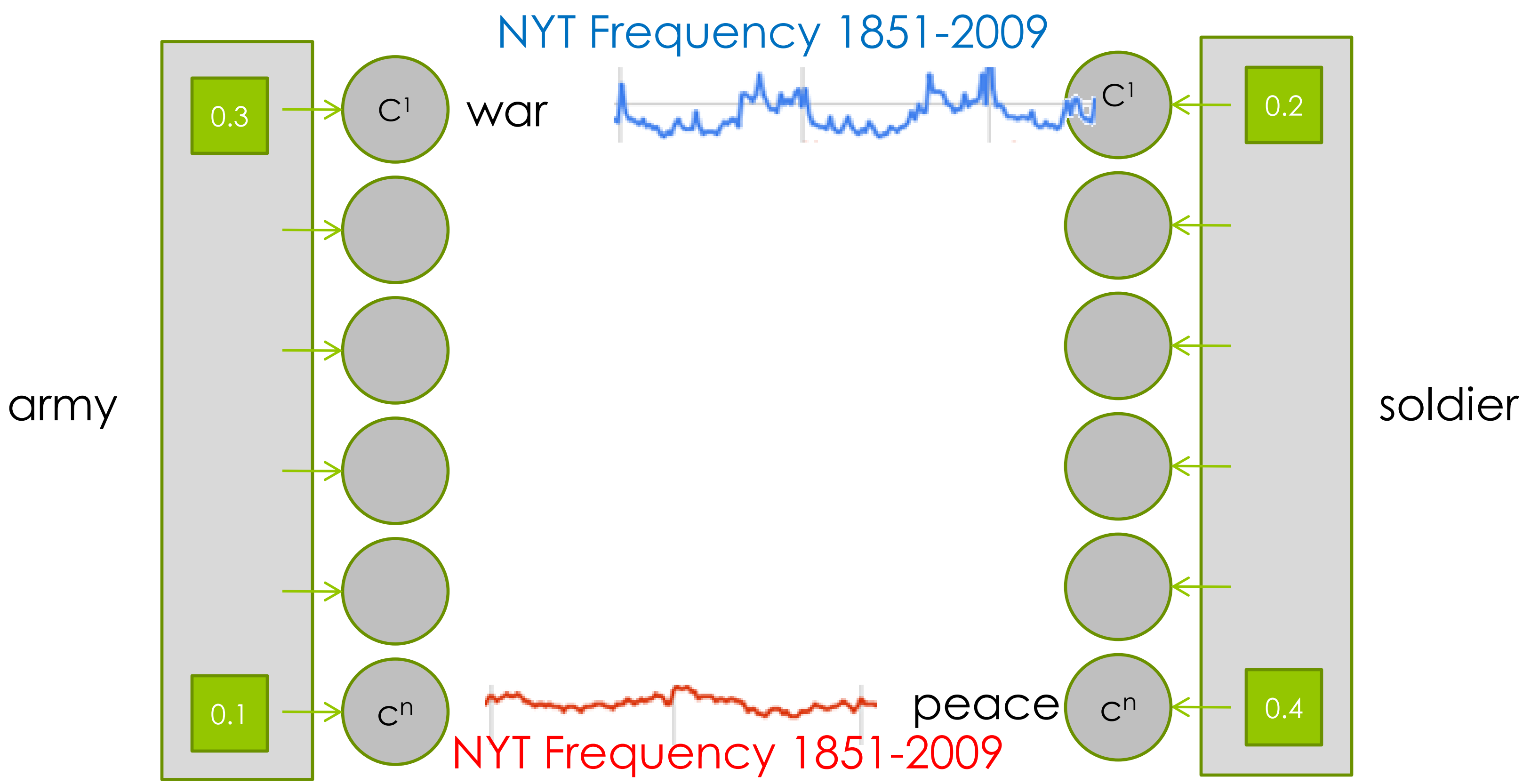
Method for computing semantic relatedness using the temporal representation

Static vector space representation



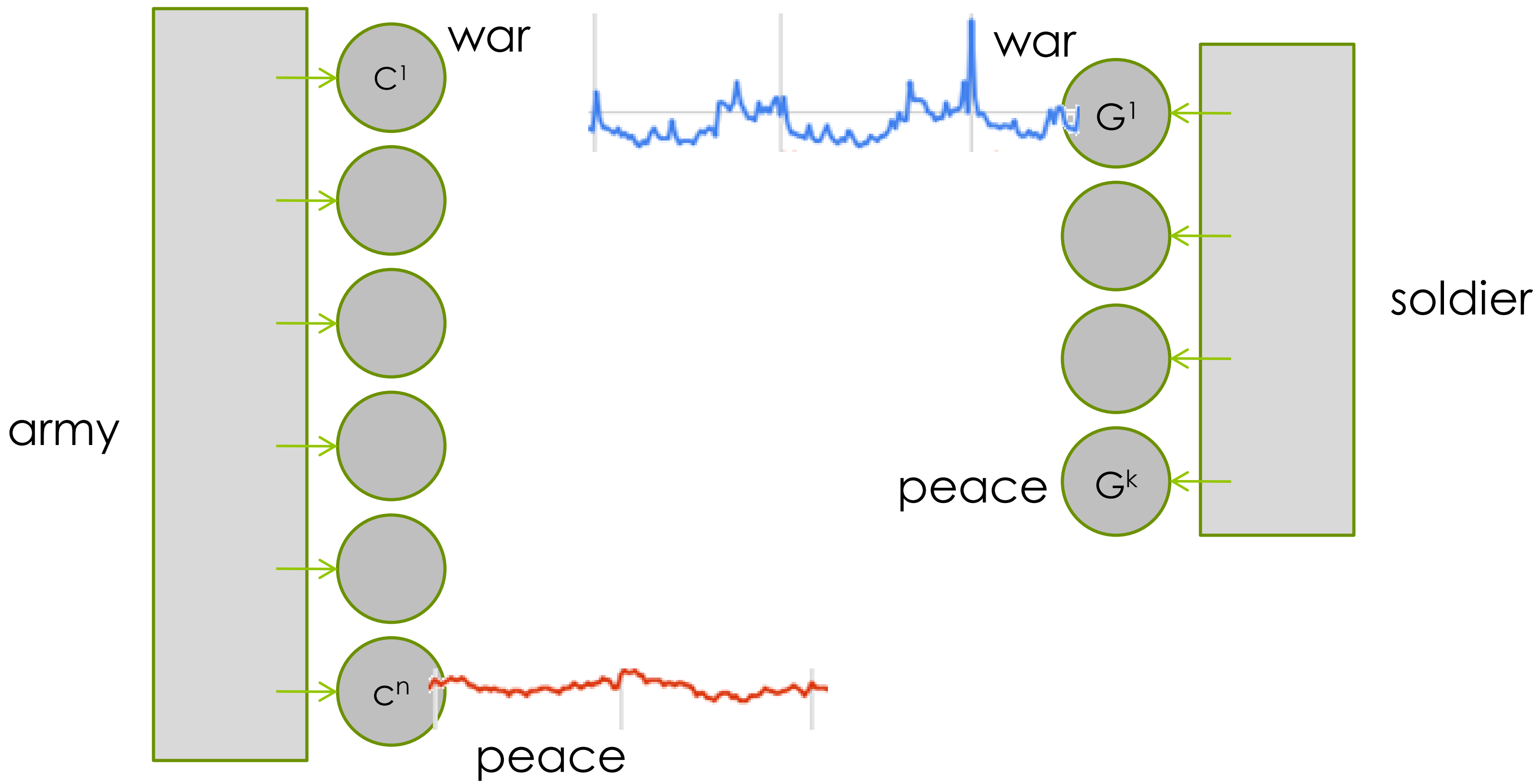
Words are represented as concept vectors: using a concept repository of choice (e.g., Wikipedia or Flickr image tags)

Temporal vector space representation



Extract temporal dynamics for each concept

Temporal vector space representation



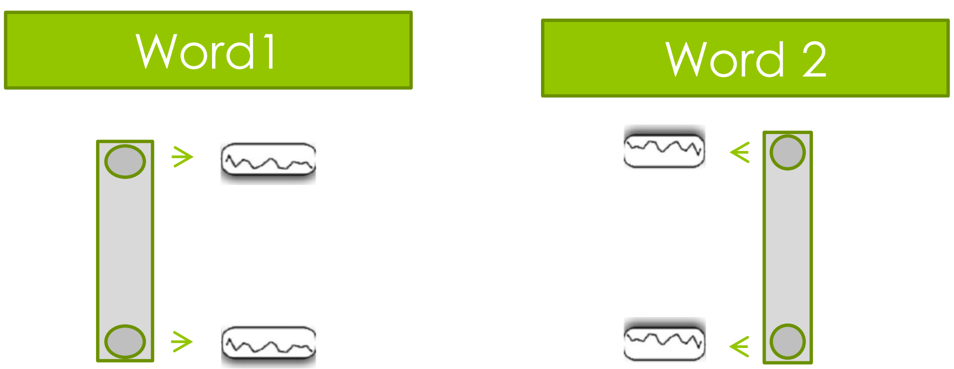
Temporal representations of words can be different, but related words tend to have similar temporal representations

Overview: Temporal semantic analysis

1. Novel temporal representation of text

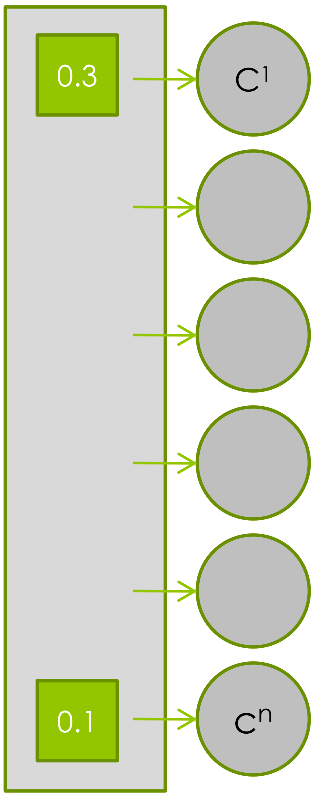
Extend static representation with temporal dynamics

2. Novel temporal text-similarity measurement

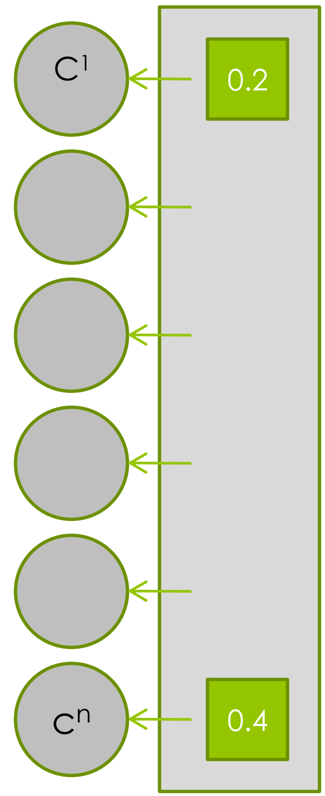


Method for computing semantic relatedness using the temporal representation

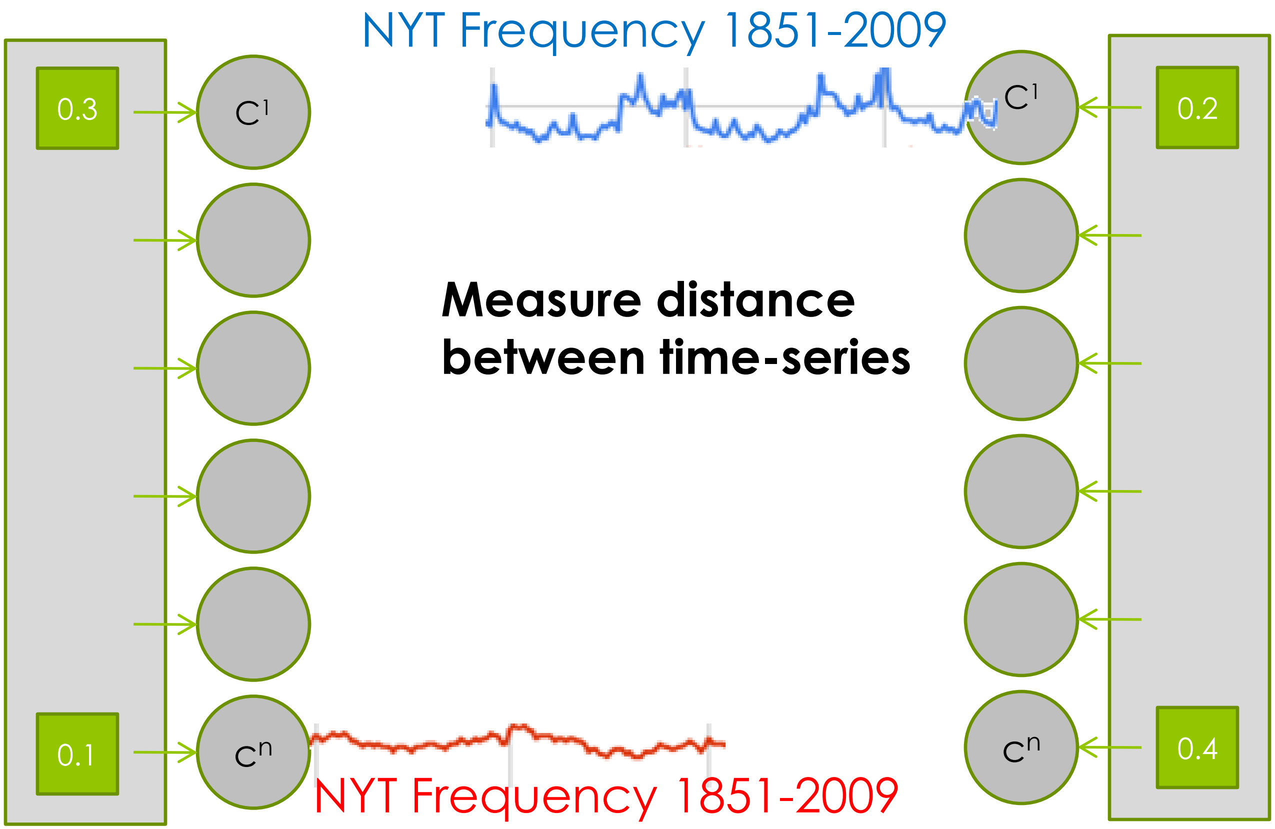
Static semantic similarity (as in ESA)



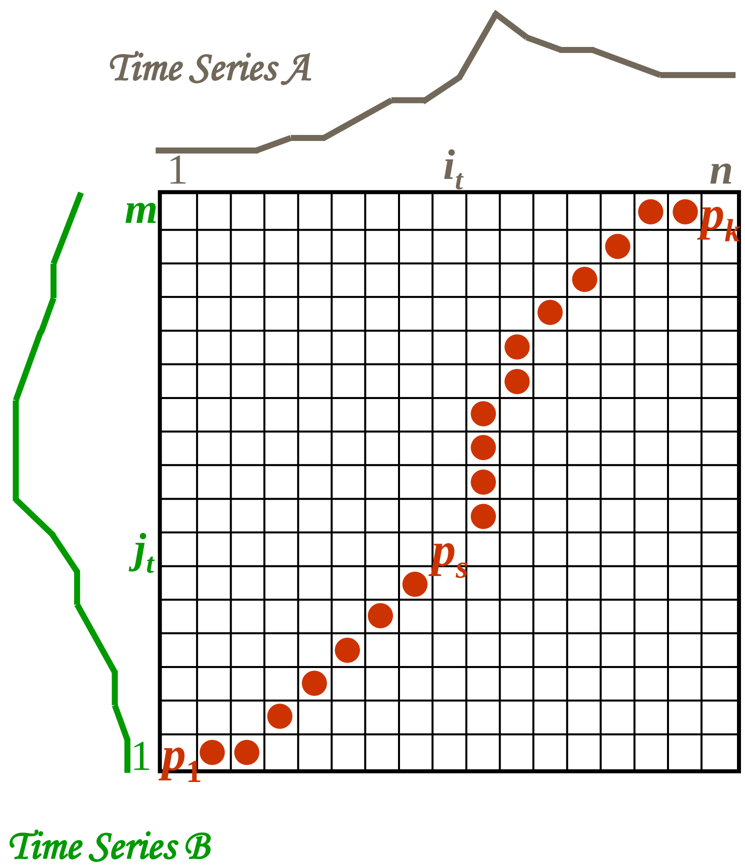
$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$



Temporal semantic similarity (TSA)



Temporal distances (Method 1):
Temporal-weighted dynamic time warping (DTW)



<u>*Time-weighted distance*</u> between $\mathcal{A}$ and $\mathcal{B}$:

$$D(\mathcal{A}, \mathcal{B}) = \sum_{t=1}^{k} d(p_t) \cdot w(t)$$

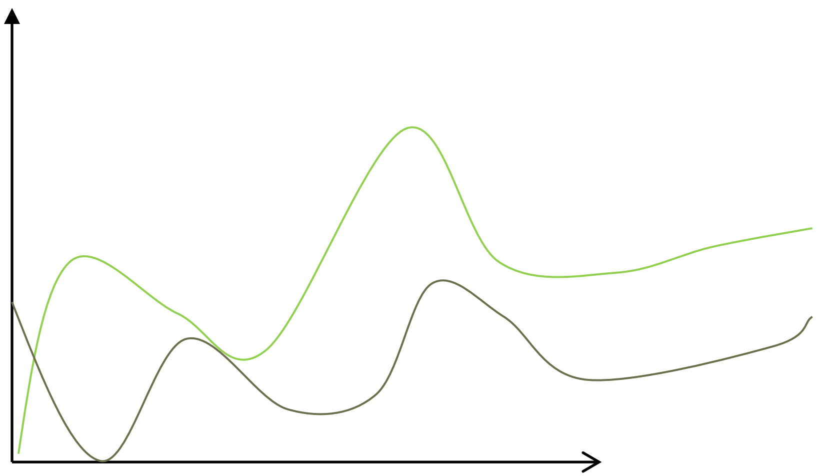
$d(p_s)$: distance between i_t and j_t

$w(t) > 0$: weighting coefficient (with decay over time)

<u>*Best alignment path*</u> between $\mathcal{A}$ and $\mathcal{B}$:

$$P_0 = \arg\min_P (D(\mathcal{A}, \mathcal{B})).$$

Temporal distances (Method 2): Temporal-weighted cross correlation



Time-weighted distance between $\mathcal{A}$ and $\mathcal{B}$:

$$D(\mathcal{A}, \mathcal{B}) = \sum_{t=0}^{n} w(t)x(t)y(t-s)$$

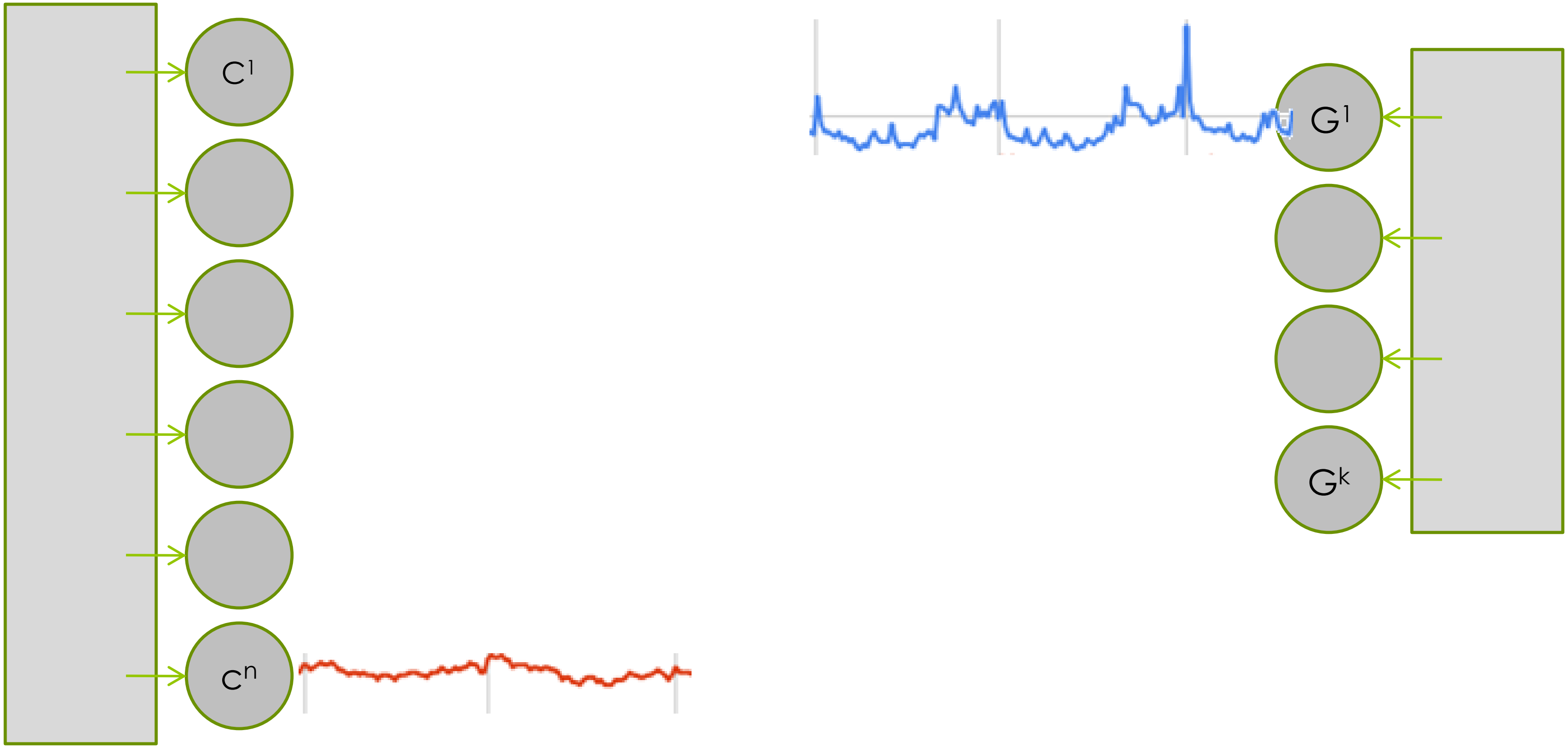
$s = 0, \pm 1, \pm 2, \ldots$

$w(t) > 0$: weighting coefficient (with decay over time)

Best alignment path between $\mathcal{A}$ and $\mathcal{B}$:

$$P_0 = \arg\min_{s}(D(\mathcal{A}, \mathcal{B})).$$

Reminder:
Temporal distance between different concepts



The sets of support concepts on both sides are DIFFERENT

Greedy temporal distance function

Procedure GREEDY DISTANCE FUNCTION(o_1, o_2)

$F(o_1) = \{ts^1_1, \ldots, ts^1_n\}$

$F(o_2) = \{ts^2_1, \ldots, ts^2_m\}$

$R(o_1, o_2) \leftarrow 0$

While $F(o_1) \neq \emptyset$ **AND** $F(o_2) \neq \emptyset$

$\quad \langle \hat{ts}_1, \hat{ts}_2 \rangle = arg\max_{\langle ts_1, ts_2 \rangle \in F(o_1) \times F(o_2)} Q(ts_1, ts_2)$

$\quad R(o_1, o_2) \leftarrow R(o_1, o_2) + Q(\hat{ts}_1, \hat{ts}_2)$

$\quad F(o_1) \leftarrow F(o_1) \backslash \{\hat{ts}_1\}$

$\quad F(o_2) \leftarrow F(o_2) \backslash \{\hat{ts}_2\}$

Return $R(o_1, o_2)$

Word-similarity benchmarks

In our experiments we have used two datasets:

1. **WS-353 dataset**: standard in the field.
   - 353 pairs of words (manually selected)
   - Each pair judged by 13 or16 human annotators
2. **MTurk dataset**: a new dataset, in which pairs of words are selected automatically
   - 287 pairs of words
   - Each pair judged by 23 human annotators

Evaluation metric: correlation with human judgments is the most commonly used metric

Main result: TSA outperform ESA

TSA algorithm vs. state-of-the-art (WS-353 dataset)

Algorithm	Correlation with humans
ESA-Wikipedia	0.75
ESA-ODP	0.65
TSA	**0.80**

TSA algorithm vs. state-of-the-art (MTurk dataset)

Algorithm	Correlation with humans
ESA-Wikipedia	0.59
TSA	**0.63**

On both datasets our algorithm outperform the state of the art.

TSA outperforms ESA mainly on low word frequency

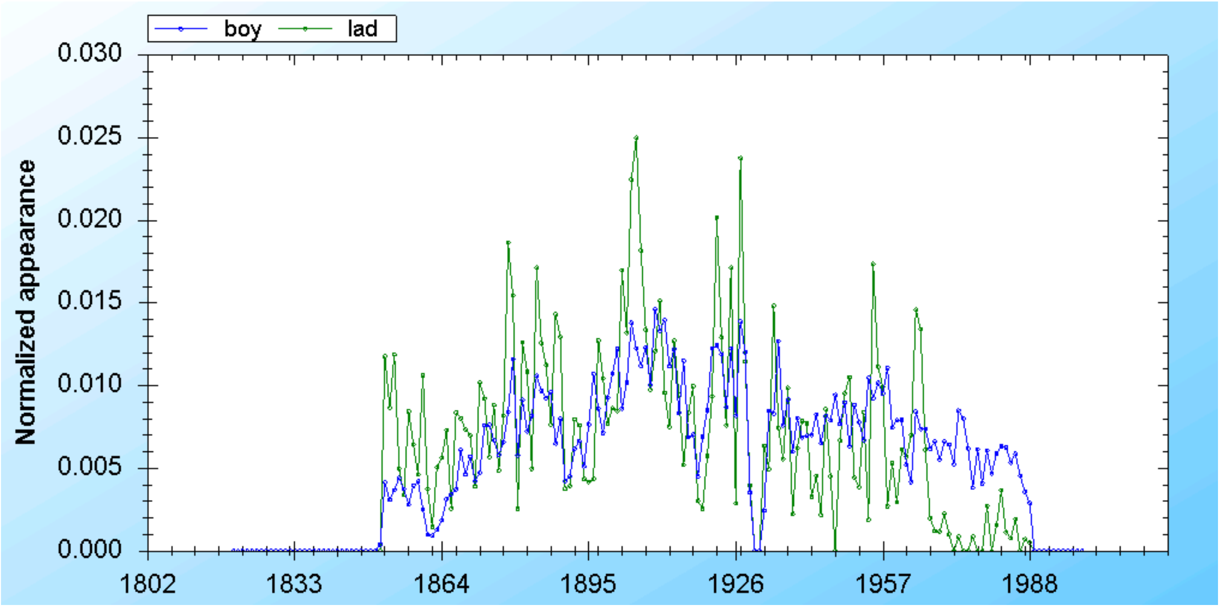
Grouping word pairs by frequency (WS-353 dataset)

Type of Bucket	ESA correlation with humans	TSA correlation with humans
Rare	0.73	**0.82**
Medium	0.74	0.76
High	0.76	0.79

We see that we best perform on low frequency:

- ESA is based on statistical information about words → Requires many examples.
- Low-frequency word-pairs have little statistical data → Any additional temporal signal can improve the performance.

Strength of TSA: Synonyms ("boy" & "lad")

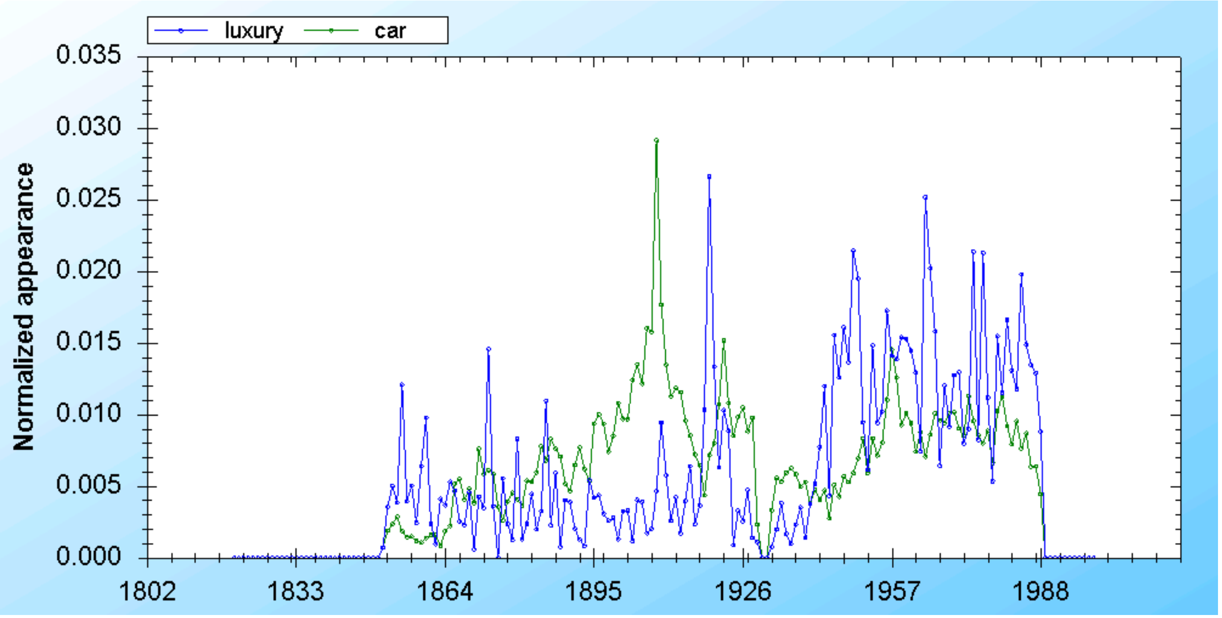


word pairs are ordered by similarity (WS353 dataset)

Human ranking: 16 TSA Ranking: 62 ESA ranking: 155

synonyms have similar patterns of occurrence over time, as writers in the news corpus tend to use them interchangeably

Strength of TSA: Compound terms ("luxury" & "car")

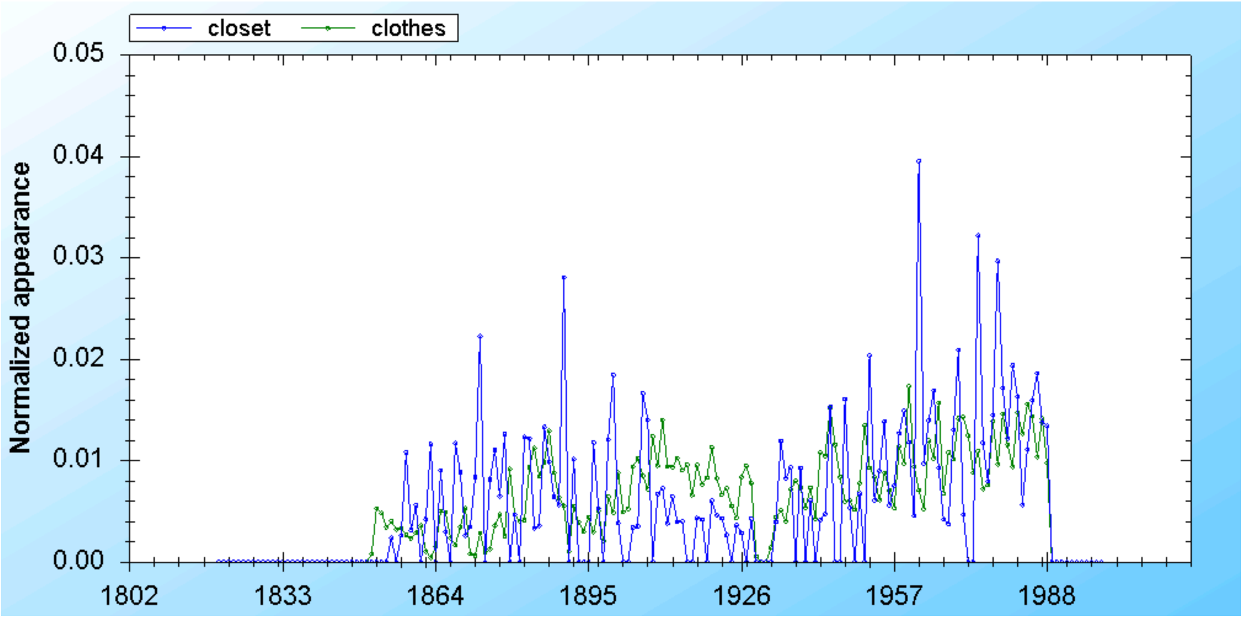


word pairs are ordered by similarity (WS353 dataset)

Human ranking: 164 TSA Ranking: 118 ESA ranking: 12

TSA captures co-occurrences of words in a single article, as we construct time-series aggregated over all articles on a certain date.

Strength of TSA: Implicit relations ("closet" & "clothes")

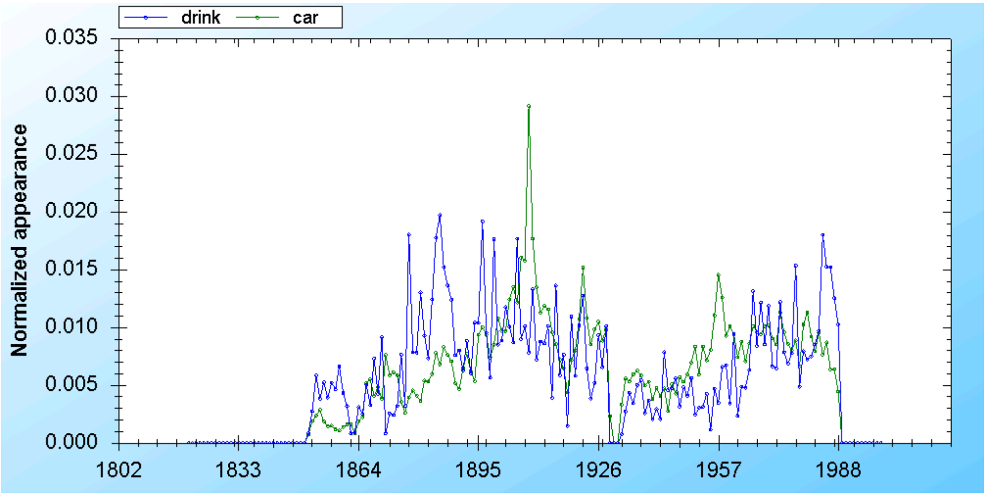


word pairs are ordered by similarity (WS353 dataset)

Human ranking: 57 TSA Ranking: 56 ESA ranking:173

Additional Implicit relations: summer-draught , canyon-landscape etc.

Limitations of TSA

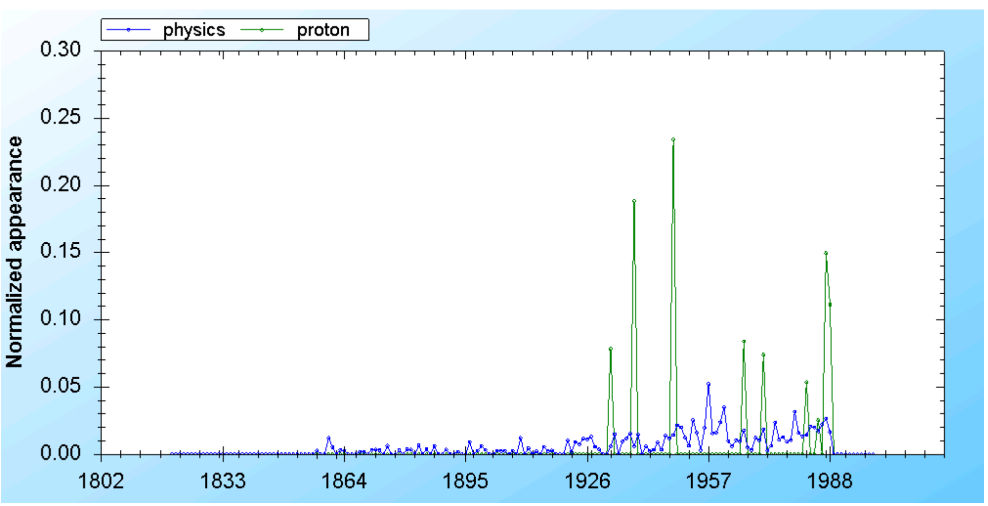


Complex implicit relations ("drink" & "car")

Human ranking: 303

TSA ranking: 150

ESA ranking: 313



News corpus bias - coverage problem ("physics" & "proton")

Human ranking: 56

TSA ranking: 322

ESA ranking: 55

Summary

1. Temporal Semantic Analysis main contributions:
   - Semantic **representation** of natural language terms using **temporal corpus** (NYT 1850-2009).
   - Semantic relatedness distance **algorithms** using temporal data.
2. Automatic algorithm for semantic relatedness datasets construction.
3. Empirical evaluation confirms using TSA outperforms current state of the art.
4. Many other temporal datasets: Nature and Science archives, Google Books, and more.

Temporal information holds a lot of promise for NLP tasks

References + supplemental materials

- Word relatedness datasets

1. WS353: www.cs.technion.ac.il/~gabr/resources/data/wordsim353
2. MTurk: www.technion.ac.il/~kirar/Datasets.html

- For any questions please email to Kira Radinsky

kirar@cs.technion.ac.il

Thank you!