Assessing Semantic Similarity of Short Texts

Mihai Lintean

Carney Labs

Excerpts from a Dissertation Defense
University of Memphis
Advisor: Dr. Vasile Rus

DC-NLP Meetup - Oct 09, 2013
The Goal

Addressing the challenging task of automatically assessing the semantic similarity of texts
The Problem

Text A: York had no problem with MTA’s insisting the decision to shift funds had been within its legal rights.

Text B: York had no problem with MTA’s saying the decision to shift funds was within its powers.

Paraphrasing - a clear case of semantic similarity
The Problem

**Text A:** York had no problem with MTA’s insisting the decision to shift funds had been within its legal rights.

**Text B:** York had no problem with MTA’s saying the decision to shift funds was within its powers.

*Paraphrasing* - a clear case of semantic similarity

**Text A:** About 1,417 schools statewide receive Title I money.

**Text B:** That applies only to schools that get federal Title I money.

A clear case when two texts are NOT semantically similar
Applications

- Question Answering Systems
  - compare the input question to a list of known questions
Applications

- **Question Answering Systems**
  - compare the input question to a list of known questions

- **Dialogue-Based Tutoring Systems**
  - compare student's answer to a list of known answers
Applications

- **Question Answering Systems**
  - compare the input question to a list of known questions

- **Dialogue-Based Tutoring Systems**
  - compare student's answer to a list of known answers

- **Text-based Clustering and Classification**
  - gather news articles about same story, event or person
  - cluster and classify retrieved documents by their topics
Contributions

▶ Investigate the role of linguistic information on semantic similarity

▶ Propose a Semantic Representation for the input texts

▶ Design, implement and evaluate a variety of Methods

▶ Develop a general Framework for assessing the semantic similarity of texts
Outline

Introduction

Semantic Similarity in Short Texts
A Framework to Measure Semantic Similarity
A Shallow Representation of Meaning

Lexical Methods

A Simple Example
Methodology
Results

Word-Semantics

Word Semantics (WordNet, LSA)
Methodology
Results
Outline

Introduction

Semantic Similarity in Short Texts
A Framework to Measure Semantic Similarity
A Shallow Representation of Meaning

Lexical Methods

A Simple Example
Methodology
Results

Word-Semantics

Word Semantics (WordNet, LSA)
Methodology
Results
The Problem Reviewed

**Text A:** York had no problem with MTA’s insisting the decision to shift funds had been within its legal rights.

**Text B:** York had no problem with MTA’s saying the decision to shift funds was within its powers.

**Qualitative Judgement - Paraphrase**

**Text A:** About 1,417 schools statewide receive Title I money.

**Text B:** That applies only to schools that get federal Title I money.

**Qualitative Judgement - NOT Paraphrase**
The Problem Reviewed

Text A: York had no problem with MTA's insisting the decision to shift funds had been within its legal rights.

Text B: York had no problem with MTA's saying the decision to shift funds was within its powers.

Quantitative Judgement - are similar to a degree of 0.9 (on a normalized scale)

Text A: About 1,417 schools statewide receive Title I money.

Text B: That applies only to schools that get federal Title I money.

Quantitative Judgement - are similar to a degree of 0.4 (on a normalized scale)

Quantitative Judgement ⟷ Qualitative Judgement
Why Quantitative Analysis

Text A: Ricky Clemons’s brief, troubled Missouri basketball career is over.

Text B: Missouri kicked Ricky Clemons off its team, ending his troubled career there.
Why Quantitative Analysis

**Text A:** Ricky Clemons’s brief, troubled Missouri basketball career is over.

**Text B:** Missouri kicked Ricky Clemons off its team, ending his troubled career there.
**Why Quantitative Analysis**

**Text A:** Ricky Clemons’ brief, troubled Missouri basketball career is over.

**Text B:** Missouri kicked Ricky Clemons off its team, ending his troubled career there.

A *paraphrase* example from the **Microsoft Research Paraphrase (MSR) Corpus**

Symmetric Relation
Why Quantitative Analysis

Text A: Ricky Clemons’ brief, troubled Missouri basketball career is over.

Text B: Missouri kicked Ricky Clemons off its team, ending his troubled career there.

A paraphrase example from the Microsoft Research Paraphrase (MSR) Corpus
Symmetric Relation

Text A: There are also tanneries, sawmills, textile mills, food-processing plants, breweries, and a film industry in the city.

Text B: Movies are also made in the city
Why Quantitative Analysis

**Text A:** Ricky Clemons’s brief, troubled Missouri basketball career is over.

**Text B:** Missouri kicked Ricky Clemons off its team, ending his troubled career there.

*A paraphrase example from the Microsoft Research Paraphrase (MSR) Corpus
Symmetric Relation*

**Text A:** There are also tanneries, sawmills, textile mills, food-processing plants, breweries, and a film industry in the city.

**Text B:** Movies are also made in the city.
Why Quantitative Analysis

Text A: Ricky Clemons’s brief, troubled Missouri basketball career is over.

Text B: Missouri kicked Ricky Clemons off its team, ending his troubled career there.

A paraphrase example from the Microsoft Research Paraphrase (MSR) Corpus
Symmetric Relation

Text A: There are also tanneries, sawmills, textile mills, food-processing plants, breweries, and a film industry in the city.

Text B: Movies are also made in the city

An entailment example from the Recognizing Textual Entailment (RTE) Corpus
Asymmetric Relation
Example #1

**Text A:** York had no problem with MTA’s **insisting** the decision to shift funds had been within its **legal rights**.

**Text B:** York had no problem with MTA’s **saying** the decision to shift funds was within its **powers**.

*A paraphrase example from the Microsoft Research (MSR) Paraphrase Corpus*
Example #1

Text A: York had no problem with MTA’s insisting the decision to shift funds had been within its legal rights.

Text B: York had no problem with MTA’s saying the decision to shift funds was within its powers.

A paraphrase example from the Microsoft Research (MSR) Paraphrase Corpus

Word-to-word Semantics
Challenges (which we address)

Example #1

Text A: York had no problem with MTA’s **insisting** the decision to shift funds had been within its **legal rights**.

Text B: York had no problem with MTA’s **saying** the decision to shift funds was within its **powers**.

A *paraphrase* example from the **Microsoft Research (MSR) Paraphrase** Corpus

**Word-to-word Semantics**

Example #2

Text A: Besançon is the **capital** of France’s watch and clock-making **industry** and of high precision **engineering**.

Text B: Beançon is the **capital** of France.

An *non-entailment* example from the **Recognizing Textual Entailment (RTE-1)** Corpus
Challenges (which we address)

Example #1

**Text A:** York had no problem with MTA’s *insisting* the decision to shift funds had been within its *legal rights.*

**Text B:** York had no problem with MTA’s *saying* the decision to shift funds was within its *powers.*

A *paraphrase* example from the Microsoft Research (MSR) Paraphrase Corpus

Word-to-word Semantics

Example #2

**Text A:** Besançon is the *capital* of France’s watch and clock-making *industry* and of high precision *engineering.*

**Text B:** Beançon is the *capital* of *France.*

An *non-entailment* example from the Recognizing Textual Entailment (RTE-1) Corpus

Syntactic Relations between words in a sentence
Example #3

Text A: That information was first reported in today’s edition of the New York Times.

Text B: The information was first printed yesterday in the New York Times.
Example #3

**Text A:** That information was first reported in today’s edition of the New York Times.

**Text B:** The information was first printed yesterday in the New York Times.
Challenges (which we do not address)

Example #3

**Text A:** That information was first *reported in today’s edition of* the New York Times.

**Text B:** The information was first *printed yesterday in* the New York Times.

Need knowledge on: 1) Time 2) Printing business
Example #3

Text A: That information was first reported in today’s edition of the New York Times.

Text B: The information was first printed yesterday in the New York Times.

Need knowledge on: 1) Time 2) Printing business

Example #4

Text A: John bought 3 apples and 2 pears.

Text B: John bought 5 fruits.

Need to know how to Add two integers (Mathematics)
Assessing Semantic Similarity between Texts

The approach is based on the Principle of Compositionality
- the meaning of a text is determined by the meaning of its constituents and the rules used to combine them

\[
\begin{align*}
\text{words} & \quad \text{numbers} \\
\text{punctuation} & \quad \text{lexical tokens}
\end{align*}
\]

How do we compare two words?
Semantic Similarity between Words

- $\text{dog} \leftrightarrow \text{mutt}$
Semantic Similarity between Words

- dog $\iff$ mutt $\iff$ animal
Semantic Similarity between Words

- $\text{dog} \leftrightarrow \text{mutt} \leftrightarrow \text{animal}$

- $\text{dog} \leftrightarrow \text{bark} \leftrightarrow \text{apple} \leftrightarrow \text{pie}$
Semantic Similarity between Words

- $\text{dog} \leftrightarrow \text{mutt} \leftrightarrow \text{animal}$
- $\text{dog} \leftrightarrow \text{bark} \leftrightarrow \text{apple} \leftrightarrow \text{pie}$
- $\text{hot} \leftrightarrow \text{cold} \leftrightarrow \text{agree} \leftrightarrow \text{disagree}$
Semantic Similarity between Words

- $\text{dog} \iff \text{mutt} \iff \text{animal}$

- $\text{dog} \iff \text{bark} \iff \text{apple} \iff \text{pie}$

- $\text{hot} \iff \text{cold} \iff \text{agree} \iff \text{disagree}$

The diagram illustrates the semantic similarity and agreement between various word pairs.
Semantic Similarity between Words

- $\text{dog} \leftrightarrow \text{mutt} \leftrightarrow \text{animal}$
- $\text{dog} \leftrightarrow \text{bark} \leftrightarrow \text{apple} \leftrightarrow \text{pie}$
- $\text{hot} \leftrightarrow \text{cold} \leftrightarrow \text{agree} \leftrightarrow \text{disagree}$
Semantic Similarity between Words

- **dog** $\leftrightarrow$ **mutt** $\leftrightarrow$ **animal**
- **dog** $\leftrightarrow$ **bark**  $\leftrightarrow$ **apple** $\leftrightarrow$ **pie**
- **hot** $\leftrightarrow$ **cold**  $\leftrightarrow$ **agree** $\leftrightarrow$ **disagree**

Semantic Similarity versus Semantic Agreement
Semantic Similarity between Sentences

- Different, unrelated topics
- Topic or Event Related
- Similar opinions on same topic
- Elaboration, Entailment
- Paraphrasing
- Semantic Equivalence

Text A: The Dow finished the volatile day with a modest gain.

Text B: US stocks rose in volatile trading, thanks only to technical factors.

Same topic...

Text A: It is now time to bring our combat troops home from Afghanistan.

Text B: NATO's secretary general argued against a retreat from Afghanistan.

...but different opinions
**Semantic Similarity between Sentences**

- **Text A:** The Dow finished the volatile day with a modest gain.
- **Text B:** US stocks rose in volatile trading, thanks only to technical factors.

*Same topic...*
Semantic Similarity between Sentences

Text A: The Dow finished the volatile day with a modest gain.
Text B: US stocks rose in volatile trading, thanks only to technical factors.

Same topic...

Text A: It is now time to bring our combat troops home from Afghanistan.
Text B: NATO’s secretary general argued against a retreat from Afghanistan.

...but different opinions
Semantic Similarity between Sentences

Text A: The Dow finished the volatile day with a modest gain.
Text B: US stocks rose in volatile trading, thanks only to technical factors.

Same topic...

Text A: It is now time to bring our combat troops home from Afghanistan.
Text B: NATO’s secretary general argued against a retreat from Afghanistan.

...but different opinions

Words → Sentences → Paragraphs → Documents
Our Dataset - The MSR Paraphrase Corpus

- identify sentential paraphrases
- 5801 instance pairs
  - 70% training (.67 T)
  - 30% testing (.66 T)
- average sentence length:
  - 17 words
- a challenging dataset
  - inconsistent labeling
  - 83% inter-rater agreement

The ULPC Corpus (2000 #instances)
The RTE Corpus (4657 #instances)
Our Dataset - The MSR Paraphrase Corpus

- identify sentential paraphrases
- 5801 instance pairs
  - 70% training (.67 T)
  - 30% testing (.66 T)
- average sentence length:
  - 17 words
- a challenging dataset
  - inconsistent labeling
  - 83% inter-rater agreement

The ULPC Corpus (2000 #instances)

The RTE Corpus (4657 #instances)
A Framework to Measure Semantic Similarity

Our Goal: to offer a fully automated and robust process

- Step 1: Semantic Mapping
  - covert the input into semantic representations
  - retain the Lexical, Syntax and Semantics
A Framework to Measure Semantic Similarity

Our Goal: to offer a fully automated and robust process

▶ Step 1: Semantic Mapping
  - covert the input into semantic representations
  - retain the Lexical, Syntax and Semantics

▶ Step 2: Compare
  - compare the representation
  - extract features that quantify the semantic similarity
A Framework to Measure Semantic Similarity

Our Goal: to offer a fully automated and robust process

▶ Step 1: Semantic Mapping
- covert the input into semantic representations
- retain the Lexical, Syntax and Semantics

▶ Step 2: Compare
- compare the representation
- extract features that quantify the semantic similarity

▶ Step 3: Learn and classify
- learn from the features
- assess qualitatively the semantic similarity
A Shallow Representation of Meaning

SR: (Word, Lemma, POS, Specificity, WN-SENSE|LSA-Vector,
(< − : dep_type : dep_mod > | < dep_head : dep_type : − >)+)+

Peter went to Seattle last Thursday.

[ (Word=Peter, lemma=peter, POS=NNP, WNSENSE=1,Deps=(went:nsubj:-)),
(went, go, VBP, 1, (−:nsubj:peter; −:prep_to:seattle; −:tmod:thursday)),
(to, to, N, 1, ()),
(Seattle, seattle, NNP, 1, (went:prep_to:-)),
(last, last, JJ, 1, (thursday:amod:-)),
(Thursday, thursday, NNP, 1, (went:tmod:-; −:amod:last)) ]
(., ., PERIOD, 1, ()) ]

▶ Easy extraction of data
▶ Human friendly
▶ Encode all lexical, syntactic and semantic facts of the input
A Shallow Representation of Meaning

SR: (Word, Lemma, POS, Specificity, WN-SENSE|LSA-Vector,
(\langle \neg: dep_{type}: dep_{mod} \rangle \mid \langle dep_{head}: dep_{type}: \neg \rangle \rangle)^+)^+

Preprocessing the Input

- Tokenize text ↦ lexical tokens (words)
A Shallow Representation of Meaning

SR: (Word, Lemma, POS, Specificity, WN-SENSE|LSA-Vector,
(< dep < dep mod | < dep head : dep type : - >) + ) +

Preprocessing the Input

- Tokenize text $\leadsto$ lexical tokens (words)
- Lematize tokens $\leadsto$ lemmas
A Shallow Representation of Meaning

SR: (Word, Lemma, POS, Specificity, WN-SENSE|LSA-Vector, 
(<−:dep_type:dep_mod>|<dep_head:dep_type:−>)+)+

Preprocessing the Input

- Tokenize text $\leadsto$ lexical tokens (words)
- Lemmatize tokens $\leadsto$ lemmas
- Part-of-Speech Tagging $\leadsto$ POSs
A Shallow Representation of Meaning

SR: (Word, Lemma, POS, Specificity, WN-SENSE|LSA-Vector,
     (< − : dep_type : dep_mod > | < dep_head : dep_type : − >)+)+

Preprocessing the Input

- Tokenize text → lexical tokens (words)
- Lematize tokens → lemmas
- Part-of-Speech Tagging → POSs
- Extract Word Specificity from precalculated Indices (IDF, Entropy)
A Shallow Representation of Meaning

SR: (Word, Lemma, POS, Specificity, WN-SENSE|LSA-Vector,
    (< − : dep_{type} : dep_{mod} > | < dep_{head} : dep_{type} : − >)+)+

Preprocessing the Input

- Tokenize text $\leadsto$ lexical tokens (words)
- Lemmatize tokens $\leadsto$ lemmas
- Part-of-Speech Tagging $\leadsto$ POSs
- Extract Word Specificity from precalculated Indices (IDF, Entropy)
- Compute the Meaning of Words $\leadsto$ WordNet Sense $\leadsto$ LSA Vector
A Shallow Representation of Meaning

SR: (Word, Lemma, POS, Specificity, WN-SENSE|LSA-Vector,
    (< −: dep\textsubscript{type}: dep\textsubscript{mod} > | < dep\textsubscript{head}: dep\textsubscript{type}: − >)+)+

Preprocessing the Input

- Tokenize text \(\leadsto\) lexical tokens (words)
- Lematize tokens \(\leadsto\) lemmas
- Part-of-Speech Tagging \(\leadsto\) POSs
- Extract Word Specificity from precalculated Indices (IDF, Entropy)
- Compute the Meaning of Words \(\leadsto\) WordNet Sense — LSA Vector
- Syntactic Parsing \(\leadsto\) dependency relations between words
A Shallow Representation of Meaning

SR: (Word, Lemma, POS, Specificity, WN-SENSE|LSA-Vector,
( < - : dep\text{type} : dep\text{mod} > | < dep\text{head} : dep\text{type} : - > ) + ) +

Extracting Dependencies

The decision had been within its legal rights.
Outline

Introduction
- Semantic Similarity in Short Texts
- A Framework to Measure Semantic Similarity
- A Shallow Representation of Meaning

Lexical Methods
- A Simple Example
- Methodology
- Results

Word-Semantics
- Word Semantics (WordNet, LSA)
- Methodology
- Results
A Simple Method to Measure Similarity of Texts

Compute the degree of token overlap between the texts

| Text A: Peter went to Seattle last Thursday. |
| Text B: Last Thursday, my friend Peter flew to Seattle for a business meeting. |

- Number of common tokens = 6 (including punctuation)
- Average number of tokens = \( \frac{7(\text{TextA}) + 14(\text{TextB})}{2} = 10.5 \)
- Similarity Score: \( \text{Sim} = \frac{6}{10.5} = 0.57 \)
- Paraphrasing: Is \( \text{Sim} \geq \) Threshold?
- Learn optimum threshold  \( \leftrightarrow \) Maximum accuracy on training
A Simple Method to Measure Similarity of Texts

Compute the degree of token overlap between the texts

| Text A: | Peter went to Seattle last Thursday. |
| Text B: | Last Thursday, my friend Peter flew to Seattle for a business meeting. |

- Number of common tokens = 6 (including punctuation)
- Average number of tokens = \( \frac{7(\text{TextA}) + 14(\text{TextB})}{2} = 10.5 \)
- Similarity Score: \( Sim = \frac{6}{10.5} = 0.57 \)
- Paraphrasing: Is \( Sim \geq \) Threshold?
- Learn optimum threshold \( \leftrightarrow \) Maximum accuracy on training

Our process:

- **Step 1)** Find all distinct pairs
- **Step 2)** Count the pairs (or do a weighted sum)
- **Step 3)** Normalize (use average or maximum length)
Decisions to Consider (when counting common tokens)

▶ Ignore Punctuation

”I am not a business man. I am a business, man.”  
Jay-Z
Decisions to Consider (when counting common tokens)

▶ Ignore Punctuation

”I am not a business man. I am a business, man.” Jay-Z

▶ Consider only Content Words or Ignore Stop-Words

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>John is flying</td>
<td>John is flying</td>
</tr>
<tr>
<td>from Seattle.</td>
<td>to Seattle.</td>
</tr>
</tbody>
</table>
Decisions to Consider (when counting common tokens)

▶ Ignore Punctuation  
“ I am not a business man. I am a business, man.”  
Jay-Z

▶ Consider only Content Words or Ignore Stop-Words

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>John is flying from Seattle.</td>
<td>John is flying to Seattle.</td>
</tr>
</tbody>
</table>

▶ Use lemmas vs. words

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>The children are playing in the courtyard.</td>
<td>The child was playing in the courtyard.</td>
</tr>
</tbody>
</table>
Decisions to Consider (when counting common tokens)

- Ignore Punctuation
  
  ”I am not a business man. I am a business, man.”  
  Jay-Z

- Consider only Content Words or Ignore Stop-Words

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>John is flying from Seattle.</td>
<td>John is flying to Seattle.</td>
</tr>
</tbody>
</table>

- Use lemmas vs. words

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>The children are playing in the courtyard.</td>
<td>The child was playing in the courtyard.</td>
</tr>
</tbody>
</table>

- Ignore Case

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>People were having a good time.</td>
<td>Most people were having a good time.</td>
<td>They made US proud.</td>
<td>They made us proud.</td>
</tr>
</tbody>
</table>
Decisions to Consider (when counting common tokens)

- Ignore Punctuation
  
  "I am not a business man. I am a business, man." Jay-Z

- Consider only Content Words or Ignore Stop-Words

  Text A: John is flying from Seattle.
  Text B: John is flying to Seattle.

- Use lemmas vs. words

  Text A: The children are playing in the courtyard.
  Text B: The child was playing in the courtyard.

- Ignore Case

  Text A: People were having a good time.
  Text B: Most people were having a good time.

  Text A: They made US proud.
  Text B: They made us proud.

- Compare with POS

  Text A: Trees line the riverbank.
  Text B: The riverbank ends the line of trees.

  Text A: They had a pleasant walk in the park.
  Text B: They pleasantly walked in the park.
Decisions to Consider (when counting common tokens)

- Ignore Punctuation
  "I am not a business man. I am a business, man."  Jay-Z

- Consider only Content Words or Ignore Stop-Words

  Text A: John is flying from Seattle.
  Text B: John is flying to Seattle.

- Use lemmas vs. words

  Text A: The children are playing in the courtyard.
  Text B: The child was playing in the courtyard.

- Ignore Case

  Text A: People were having a good time.
  Text B: Most people were having a good time.
  Text A: They made US proud.
  Text B: They made us proud.

- Compare with POS

  Text A: Trees line the riverbank.
  Text B: The riverbank ends the line of trees.
  Text A: They had a pleasant walk in the park.
  Text B: They pleasantly walked in the park.

- Unigram versus Bigram overlap
  - pair bigrams of tokens
Decisions to Consider (when counting common tokens)

- Ignore Punctuation
  
  "I am not a business man. I am a business, man."  
  Jay-Z

- Consider only Content Words or Ignore Stop-Words

<table>
<thead>
<tr>
<th>Text A:</th>
<th>Text B:</th>
</tr>
</thead>
<tbody>
<tr>
<td>John is flying from Seattle.</td>
<td>John is flying to Seattle.</td>
</tr>
</tbody>
</table>

- Use lemmas vs. words

<table>
<thead>
<tr>
<th>Text A:</th>
<th>Text B:</th>
</tr>
</thead>
<tbody>
<tr>
<td>The children are playing in the courtyard.</td>
<td>The child was playing in the courtyard.</td>
</tr>
</tbody>
</table>

- Ignore Case

<table>
<thead>
<tr>
<th>Text A:</th>
<th>Text B:</th>
</tr>
</thead>
<tbody>
<tr>
<td>People were having a good time.</td>
<td>Most people were having a good time.</td>
</tr>
<tr>
<td>They made US proud.</td>
<td>They made us proud.</td>
</tr>
</tbody>
</table>

- Compare with POS

<table>
<thead>
<tr>
<th>Text A:</th>
<th>Text B:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees line the riverbank.</td>
<td>The riverbank ends the line of trees.</td>
</tr>
<tr>
<td>They had a pleasant walk in the park.</td>
<td>They pleasantly walked in the park.</td>
</tr>
</tbody>
</table>

- Unigram versus Bigram overlap - pair bigrams of tokens

- Weighting, Normalization
Local and Global Weighting Schemas

Local Weighting

\[ w_{binary}(i, j) = \begin{cases} 
1 & \text{if } i \in j \\
0 & \text{if } i \notin j
\end{cases} \]

\[ l_{weight}(i, j) = \begin{cases} 
tf_{ij} & \text{if } i \in j \\
0 & \text{if } i \notin j
\end{cases} \]

\[ l_{weight}(i, j) = \log[l_{weight}(i, j) + 1] \]

Global Weighting

\[ g_{weight}(i) = 1 + \sum_j \frac{p_{ij} \log_2(p_{ij})}{\log_2(n)} \]

, where \( p_{ij} = \frac{tf_{ij}}{\sum_{k \in D} tf_{ik}} \)

\[ g_{weight}(i) = \log \frac{|D|}{\sum_{j \in D} w_{binary}(i, j)} \]

i = type of a lexical token
j = a text instance or a document
D = a collection of documents
\( tf_{ij} = \) frequency of i in j
Understanding the Results

- We show accuracy and precision ...
  
  ... on both training and testing data

- We compare several methods in a graph (around 8 methods/graph)


  (ST|OP) = Stanford Processing | OpenNLP Processing

  (P|W|C|S) = Punctuation | Words Only | Content Words | No Stop-Words

  (W|B|P) = Compare Words | Lemmas | Lemmas with POS

  (C|I) = Case Sensitive | Case Insensitive

  (U|B) = Unigrams | Bigrams

  (I|E|N) = Global Weighting (IDF | Entropy | NoWeight)

  (F|N) = Local Weighting (Frequency | NoWeight)

**Example:** ST-W.B.I.U.N.F
Lexical Methods with Average-Norm

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc. on Train</th>
<th>Prec. on Train</th>
<th>Acc on Test.</th>
<th>Prec on Test.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST-PBIUNN</td>
<td>0.755</td>
<td>0.732</td>
<td>0.761</td>
<td>0.755</td>
</tr>
<tr>
<td>ST-WWIUNN</td>
<td>0.740</td>
<td>0.727</td>
<td>0.757</td>
<td>0.760</td>
</tr>
<tr>
<td>ST-SBIUNN</td>
<td>0.743</td>
<td>0.709</td>
<td>0.735</td>
<td>0.729</td>
</tr>
<tr>
<td>ST-PBIUNF</td>
<td>0.732</td>
<td>0.730</td>
<td>0.732</td>
<td>0.737</td>
</tr>
<tr>
<td>ST-PBIUEN</td>
<td>0.737</td>
<td>0.705</td>
<td>0.711</td>
<td>0.725</td>
</tr>
<tr>
<td>ST-PBIUN</td>
<td>0.714</td>
<td>0.697</td>
<td>0.710</td>
<td>0.692</td>
</tr>
<tr>
<td>ST-PBIUNF</td>
<td>0.715</td>
<td>0.693</td>
<td>0.710</td>
<td>0.692</td>
</tr>
<tr>
<td>prev-#4</td>
<td>0.743</td>
<td>0.709</td>
<td>0.729</td>
<td>0.740</td>
</tr>
</tbody>
</table>
Outline

Introduction
Semantic Similarity in Short Texts
A Framework to Measure Semantic Similarity
A Shallow Representation of Meaning

Lexical Methods
A Simple Example
Methodology
Results

Word-Semantics
Word Semantics (WordNet, LSA)
Methodology
Results
Motivation for Word-to-Word Similarity Metrics

Text A: York had no problem with MTA’s *insisting* the decision to shift funds had been within its *legal rights*.

Text B: York had no problem with MTA’s *saying* the decision to shift funds was within its *powers*.
Motivation for Word-to-Word Similarity Metrics

Text A: York had no problem with MTA’s *insisting* the decision to shift funds had been within its *legal rights*.

Text B: York had no problem with MTA’s *saying* the decision to shift funds was within its *powers*.

We use **WordNet Similarity** and **LSA**-based metrics

**insisting** versus **saying**

<table>
<thead>
<tr>
<th>W2W Metric</th>
<th>insist ⇔ say</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNS Path</td>
<td>0.333</td>
</tr>
<tr>
<td>WNS Lin</td>
<td>0.594</td>
</tr>
<tr>
<td>WNS Lch</td>
<td>0.670</td>
</tr>
<tr>
<td>WNS HSO</td>
<td>0.375</td>
</tr>
<tr>
<td>LSA</td>
<td>0.126</td>
</tr>
</tbody>
</table>
Motivation for Word-to-Word Similarity Metrics

Text A: York had no problem with MTA's insisting the decision to shift funds had been within its legal rights.

Text B: York had no problem with MTA's saying the decision to shift funds was within its powers.

We use WordNet Similarity and LSA-based metrics

insisting versus saying

<table>
<thead>
<tr>
<th>W2W Metric</th>
<th>insist ⇔ say</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNS Path</td>
<td>0.333</td>
</tr>
<tr>
<td>WNS Lin</td>
<td>0.594</td>
</tr>
<tr>
<td>WNS Lch</td>
<td>0.670</td>
</tr>
<tr>
<td>WNS HSO</td>
<td>0.375</td>
</tr>
<tr>
<td>LSA</td>
<td>0.126</td>
</tr>
</tbody>
</table>

- first, pair identical tokens
- then, pair words on $W2W \geq Th_{sim}$
- Weight-sum the pairs on their W2W metric
Computing Similarity for W2W Methods

Symmetric Similarity \((A \leftrightarrow B)\)

\[
Sim_{W2W}(A, B) = \frac{2 \cdot \sum_{w_A \in A, w_B \in B \text{(paired)}} \frac{\text{weight}(w_A) + \text{weight}(w_B)}{2} W2W(w_A, w_B)}{\sum_{w \in A} \text{weight}(w) + \sum_{w \in B} \text{weight}(w)}
\]  

(1)
Computing Similarity for W2W Methods

Symmetric Similarity ($A \Leftrightarrow B$)

$$\text{Sim}_{W2W}(A, B) = \frac{2 \times \sum_{w_A \in A, w_B \in B(\text{paired})} \frac{\text{weight}(w_A) + \text{weight}(w_B)}{2} W2W(w_A, w_B)}{\sum_{w \in A} \text{weight}(w) + \sum_{w \in B} \text{weight}(w)}$$ (1)

Normalization on Maximum Length (Max-Norm) ($A \Leftrightarrow B$)

$$\text{Sim}_{W2W}(A, B) = \frac{\sum_{w_A \in A, w_B \in B(\text{paired})} \frac{\text{weight}(w_A) + \text{weight}(w_B)}{2} W2W(w_A, w_B)}{\text{Max}(\sum_{w \in A} \text{weight}(w), \sum_{w \in B} \text{weight}(w))}$$ (2)
Computing Similarity for W2W Methods

**Symmetric Similarity (A ⇔ B)**

\[
Sim_{W2W}(A, B) = \frac{2 \sum_{w_A \in A, w_B \in B} \frac{weight(w_A) + weight(w_B)}{2} W2W(w_A, w_B)}{\sum_{w \in A} weight(w) + \sum_{w \in B} weight(w)}
\]

(1)

**Normalization on Maximum Length (Max-Norm) (A ⇔ B)**

\[
Sim_{W2W}(A, B) = \frac{\sum_{w_A \in A, w_B \in B} \frac{weight(w_A) + weight(w_B)}{2} W2W(w_A, w_B)}{\text{Max}(\sum_{w \in A} weight(w), \sum_{w \in B} weight(w))}
\]

(2)

**Asymmetric Similarity (A ⇒ B)**

\[
Sim_{W2W}(A, B) = \frac{\sum_{w_A \in A, w_B \in B} \frac{weight(w_A) + weight(w_B)}{2} W2W(w_A, w_B)}{\sum_{w \in B} weight(w)}
\]

(3)
Results on W2W Methods

W2W Methods with Max-Norm (WSD-all)

- Acc. on Train
- Prec. on Train
- Acc on Test.
- Prec on Test.