Semantic Textual Similarity

Mona Diab
mtdiab@gwu.edu
The George Washington University

(joint effort with Eneko Agirre, Daniel Cer, Aitor Gonzalez-Agirre, Weiwei Guo)

JSSP 2013
Intuition

John loves the taste of apples.

John thinks apples are delicious.
Intuition

John loves the taste of apples.

John thinks apples are delicious.
Intuition

John finds apples to be quite tasty.

John thinks apples are delicious.
Intuition

John likes the taste of apples.

John thinks apples are delicious.
Intuition

John loves the smell of apples.

John thinks apples are delicious.
Intuition

John loves the smell of apples.

John thinks apples are delicious.

Kinda Similar
Intuition

John loves Apple tea.

John thinks apples are delicious.
Intuition

Making delicious apple tarts is John’s hobby.

John thinks apples are delicious.
Intuition

Making delicious apple tarts is John’s hobby.

John thinks apples are delicious.
BUT hold on!
Intuition

Apple trees ripen in the winter time.

John thinks apples are delicious.
Intuition

Bananas grow in tropical regions.

John thinks apples are delicious.

Not similar
Intuition

Making delicious apple tarts is John’s hobby. Apple trees ripen in the winter time. Bananas grow in tropical regions.

John thinks apples are delicious.
Intuition

Making delicious apple tarts is John’s hobby.

Apple trees ripen in the winter time.

Bananas grow in tropical regions.

John thinks apples are delicious.
People’s judgment and intuition seem to be good at discerning meaning equivalence.

Actually, they are good at discerning various degrees of meaning equivalence.

Can we quantifiably characterize that intuition and use it effectively?
RoadMap

• What is the task of Semantic Textual Similarity (STS)?
• Why STS?
• Task so far
• 2014
• And beyond....
What is STS?

A metric that reflects the degree of similarity between two snippets of text $t_1$ and $t_2$

Degree of similarity quantifies semantic equivalence, i.e. $t_1$ and $t_2$, if extremely similar, then they bear the same meaning

A forest is a large area where trees grow close together.

The coast is an area of land that is next to the sea.

Woodland is land with a lot of trees.
What is STS?

A metric that reflects the degree of similarity between two snippets of text $t_1$ and $t_2$

Degree of similarity quantifying semantic equivalence, i.e. $t_1$ and $t_2$, if extremely similar, then they bear the same meaning

A forest is a large area where trees grow close together.
The coast is an area of land that is next to the sea.
Woodland is land with a lot of trees.
Graded Semantic Textual Similarity

A forest is a large area where trees grow close together.
Woodland is land with a lot of trees.

[Semantic Similarity Score: 2.51]

VS

Once there was a Czar who had three lovely daughters.
There were three beautiful girls, whose father was a ruler of Russia.

[Semantic Similarity Score: 4.3]
Desiderata for STS: Nuanced Semantics with utility

• STS should inform us on
  • *how* quantifiably similar \( t_1 \) and \( t_2 \) are \( \Rightarrow \) a graded similarity score
  • *Why* \( t_1 \) and \( t_2 \) are similar giving a nuanced interpretation of similarity based on semantic components’ contributions

• \( t_1 \) and \( t_2 \) could be of variable size and of different languages

• Plug & play environment for semantic components

• A hub for semantic processing as a *gray box* for NLP applications & beyond

• Lends itself to an *extrinsic* evaluation of scattered semantic components
Semantic Textual Similarity

Welcome to my world, trust me you will never be disappointed😊

Yo! Come over here, you will be pleasantly surprised😊

بس تعالى ومالكش دعوه، هتبنبسط اخر انبساط😊

• Quantitative Graded Similarity Score
• Principled Interpretability, i.e. which semantic components led to results
• Confidence Score
Semantic Textual Similarity

Welcome to my world, trust me you will never be disappointed 😊

Yo! Come over here, you will be pleasantly surprised 😊

بس تعالى ومالكش دعوه، هتينبسط اخر انسساط😊

- **Semantic Similarity score:** 4
- **Interpretation:** Lexical X Y, Syntactic AB, CD, Scoping xyz, etc
- **Confidence:** 0.8
Welcome to my world, trust me you will never be disappointed😊

Yo! Come over here, you will be pleasantly surprised😊

بس تعالى ومالكش دعوه هتبنبسط اخر انسساط😊

**Semantic Similarity**

- **Semantic Similarity score:** 4.3
- **Interpretation:** Lexical Z P, Informal Language, Modality xyz, etc
- **Confidence:** 0.95
Why STS?

- Most NLP applications have some core STS components
- STS is relevant in the monolingual space
  - MT evaluation, Summarization, Paraphrase Generation, *Populating Entailment Graphs Nodes*
Why STS?

• Most NLP applications have some core STS components

• STS is relevant in the monolingual space
  • MT evaluation, Summarization, Paraphrase Generation, Populating Entailment Graphs Nodes

• STS is relevant in the cross lingual space
  • Direct MT evaluation, X-lingual Summarization, X-lingual Generation

• Overall better understanding of semantic spaces
  • How do different languages carve up the space
  • What impact does it have on our thinking
But why *graded* STS?

- **Jives with human intuition!**
- **More relevant:** various types of NLP applications probably need different degrees of STS *(empirically decided on)*, e.g.:
  - Redundancy Detection for summarization/smart filtering in massive data conditions (medium to high similarity)
  - Machine Translation evaluation
    - Different domains might require different levels of similarity (homeland security vs. literature)
  - Novelty Detection in events (medium to high similarity)
Hence

A need for large scale graded similarity metric

...A nice unifying umbrella task for semantic components
To date...

Well till 2010 or so....
Established Word Level STS

Word similarity has been relatively well studied.

Word similarity and word relatedness metrics highly correlate with human judgments.

- cord smile 0.02
- fruit furnace 0.05
  ...
- hill woodland 1.48
- car journey 1.55
- cemetery mound 1.69
  ...
- cemetery graveyard 3.88
- automobile car 3.92

More similar
Gaps for Bigger Chunks of Text

- **Tiny** amounts of sentence level Graded similarity existed
  - Li et al. (2006) 65 pairs of glosses
  - Lee et al. (2005) 50 documents on news
- **Large** Paraphrase datasets judge semantic equivalence between text fragments produce **binary** similarity score
- Textual entailment is **directional**

*STS:* Provides a graded notion of semantic equivalence on a relatively large scale
Desiderata for STS: Nuanced Semantics with utility

- STS should inform us on
  - *how* quantifiably similar $t_1$ and $t_2$ are $\Rightarrow$ a graded similarity score
  - *Why* $t_1$ and $t_2$ are similar giving a nuanced interpretation of similarity based on semantic components’ contributions
- $t_1$ and $t_2$ could be of variable size and of different languages
- Plug & play environment for semantic components
- A hub for semantic processing as a *gray box* for NLP applications & beyond
- Lends itself to an *extrinsic* evaluation of scattered semantic components
Possible Contributors

- STS as a unified framework that combines and evaluates semantic (and pragmatic components) word sense disambiguation and induction, lexical substitution, semantic role labeling, multiword expression detection and handling, anaphora and coreference resolution, time and date resolution, named-entity handling, underspecification, hedging, semantic scoping, discourse analysis, etc.
We want a shared *commons* of good STS model components.

But good components are very hard to distribute/share

- Work involved in packaging and supporting research software
- Complex dependencies on other site specific software packages
- Data redistribution restrictions
- Incompatible software licenses

**Distributed feature extraction**

- Allow sites to provide model Features as a Service (FaaS)
- REST like API over HTTP - avoids firewall issues
- Language Independent
  - APIs for Python & Java
  - Priorities for adding more?

---

**Plug & Play: STS Common**

**Problem**

- Allow sites to easily build on the best components from other teams

---

**Work in Progress**
Operationalizing STS

- A system with various (explicit) semantic features
  - Ideally an interoperable pipeline of semantic components
- Input
  - Two text snippets
- Output
  - Numerical score of similarity
  - Confidence level in response
  - Why deemed similar
  - What semantic components led to score (interpretability)
- Evaluation
  - STS Intrinsic evaluation of sentence similarity
  - STS Extrinsic evaluation in MT evaluation
  - Intrinsic semantic component evaluations \(\rightarrow\) feature confidence scores
  - Extrinsic semantic component evaluation in STS evaluation
Operationalizing STS

• A system with various (explicit) semantic features
  • Ideally an interoperable pipeline of semantic components

• Input
  • Two text snippets

• Output
  • Numerical score of similarity
  • Confidence level in response
  • Why deemed similar
  • What semantic components led to score (interpretability)

• Evaluation
  • STS Intrinsic evaluation of sentence similarity
  • STS Extrinsic evaluation in MT evaluation
  • Intrinsic semantic component evaluations ➔ feature confidence scores
  • Extrinsic semantic component evaluation in STS evaluation

So far
A Pilot Task in SemEval 2012

- To set a definition of STS as a graded notion which can be easily communicated to non-expert annotators beyond the likert-scale
- To gather a substantial amount of sentence pairs from diverse datasets, and to annotate them with high quality
- To explore evaluation measures for STS
- To explore the relation of STS to paraphrase and Machine Translation Evaluation exercises
Description of the task

- Given two sentences, $s_1$ and $s_2$:
  - Return a **similarity** score
  - ... and an **optional confidence** score

- **Evaluation**
  - Correlation (**Pearson**) with average of **human scores**
Data Sets (sentence pairs)

- MSR paraphrase: train (750), test (750)
- MSR video: train (750), test (750)
- WMT 07–08 (EuroParl): train (734), test (499)
- Surprise test datasets
  - WMT 2007 news: test (399)
  - Ontonotes - WordNet glosses: test (750)
Instructions for annotation

Compare the Meaning of Two Statements (v.2.5)

Instructions

Two statements can mean the same thing even if they use very different words and phrases. Conversely, two statements that are superficially very similar in their word choice, phrasing and overall composition can have very different meanings.

Your job is to compare two statements and decide the type of relationship that holds between their underlying meanings or messages (i.e., what they say about or refer to in the world).

To do this task successfully, **picture** what is being described and contrast **exactly** what is conveyed by one statement versus what is being conveyed by the other.

Do the statements refer to the exact same person, action, event, idea or thing? Or, are they similar but differ according to either large or small details?

Tips:

- Be **precise** in your assignments and **try to avoid overusing any one of the category labels** (e.g., don’t just label most of the pairs as “mostly equivalent” or “roughly equivalent”).
- Be careful of **subtle differences** between the pairs that have an important impact on what is being said or described.
- Ignore grammatical errors and awkward wordings within the statements as long as they do not obscure what a statement is suppose to convey.
Definition of similarity

Likert scale with definitions

Compare Two Similar Sentences

Score how similar two sentences are to each other according to the following scale.

The sentences are:

(5) Completely equivalent, as they mean the same thing.
(4) Mostly equivalent, but some unimportant details differ.
(3) Roughly equivalent, but some important information differs/missing.
(2) Not equivalent, but share some details.
(1) Not equivalent, but are on the same topic.
(0) On different topics.
Instructions for annotation

**Likert scale** with definitions and examples.

- (5) The two sentences are completely equivalent, as they mean the same thing.
  
  *The bird is bathing in the sink.*
  
  *Birdie is washing itself in the water basin.*

- (4) The two sentences are mostly equivalent, but some unimportant details differ.
  
  *In May 2010, the troops attempted to invade Kabul.*
  
  *The US army invaded Kabul on May 7th last year, 2010.*

- (3) The two sentences are roughly equivalent, but some important information differs/missing.
  
  *John said he is considered a witness but not a suspect.*
  
  *"He is not a suspect anymore." John said.*

- (2) The two sentences are not equivalent, but share some details.
  
  *They flew out of the nest in groups.*
  
  *They flew into the nest together.*

- (1) The two sentences are not equivalent, but are on the same topic.
  
  *The woman is playing the violin.*
  
  *The young lady enjoys listening to the guitar.*

- (0) The two sentences are on different topics.
  
  *John went horse back riding at dawn with a whole group of friends.*
  
  *Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.*
Annotation Process

- Pilot gold 3 in-lab annotators (200 pairs)
  - Pairwise correlation (0.84\(r\) to 0.87\(r\)), with average (0.87\(r\) to 0.89\(r\))
- Amazon Mechanical Turk
  - 5 annotations per pair, averaged
  - Remove low performing turkers relative to gold
  - Correlation with in-lab 0.90\(r\) to 0.94\(r\)
  - MSR: 2.76 mean similarity, 0.66 sdv.
  - WMT-EuroParl: 4.05 mean similarity, 0.66 sdv.
2012 Results

- Baselines: random, cosine of tokens
- Participation: 120 hours to submit three runs.
  - 35 teams, 88 runs

Evaluation

- Pearson for each dataset
- Three measures:
  - \textit{All}: Concatenate all 5 datasets
  - \textit{Mean}: Weighted mean (micro-average)
    - Statistical significance
  - \textit{AllNorm}: Normalize each dataset and concatenate
2012 Results

- Majority of runs \textit{better than both baselines}
- Best three runs
  - All: 0.82 \( r \) UKP, TAKELAB, TAKELAB
  - Mean: 0.67 \( r \) TAKELAB, UKP, TAKELAB
  - AllNorm: 0.86 \( r \) UKP, TAKELAB, SOFT-CARDINALITY
- Per Dataset
  - MSRpar 0.73 \( r \) TAKELAB, ITA 0.71
  - MSRvid 0.88 \( r \) TAKELAB, ITA 0.874
  - WMT-eur 0.57 \( r \) SRANJANS, ITA 0.53
  - Surprise: On-WN 0.73 \( r \) WEIWEI, ITA 0.63
  - Surprise: WMT-news 0.61 \( r \) FBK, ITA 0.564
- Very few systems reported confidence scores
  - Weighted Pearson correlation
    \( \Rightarrow \) scores improved for two systems
2012 Pilot worked!

- Defined STS as likert scale with definitions

- Produced several publically available STS systems

- Produced a wealth of publically available data of high quality (~ 5250 sentence pairs)

- Very successful participation (35 teams, corresponding to 88 runs, unprecedented in SemEval)

- DARPA endorsement as one of the official evals for a major program DEFT
STS 2013: *SEM Shared Task

- **Core Task** (similar to 2012)
  - New related within genre data
  - Variable genre data
  - Check if new developments in systems

- **New Pilot Task**
  - **Typed-similarity Task**
    - scoring similarity between semi structured data records pertaining to cultural heritage items, while providing reason (similar author, time period, location)

- Similar data annotation process using crowd sourcing for both subtasks
Core Data sources

- 2012 Data (could be used to train in 2013)
  - MSR paraphrase (1500), MSR video (1500), WMT-EuroParl & News (1500), Ontonotes - WN glosses: test (750)
- 2013 new data (test only), crowd sourced annotation
  - News Headlines: (750), ITA 80.7%
  - SMT (HTER and HYTER): (750), ITA 37.7%
- Glosses
  - OntoNotes-WN (561), ITA 83.2%
  - FrameNet-WN: (189), ITA 42.7%
Data Examples

- **SMT**
  - Best seasoning - edible vinegar: Helps you relieve tiredness and is very good in inhibiting bacteria.
  - the top condiment -- vinegar: it will assist you to diminish fatigue, also, it has particularly strong power of controlling the bacterium

- **HDL**
  - Bahrain Grand Prix Underway Amid Protests
  - Bahrain race goes ahead amid unrest
Data Examples

• **ON-WN**
  - a boundary area or edge of something
  - a boundary marking the extremities of something.

• **FN-WN**
  - a phenomenon is portrayed with respect to the degree of likelihood that it will be perceived and known, given the (usually implicit) evidence, perceiver, and the circumstances in which it is considered.
  - be or become visible or noticeable;
Item 1

**Title**
Sculptured slabs of Aditya and Buddha, photographed at the Bihar Museum.

**Creator**
Photographer: Beglar, Joseph David

**Subject**
Bihar Bihar Sharif India Archaeological Survey of India Collections Archaeological Survey of India Collections (Indian Museum Series) Indian sculpture Indian sculpture (Buddhist) South Asia -- History 954

**Description**
This photograph showing sculpture fragments was taken by Joseph David Beglar in the 1870s. The sculptures were located in the Bihar museum and the photograph is part of the Archaeological Survey of India Collections. A note written by Bloch reads: "The sculptures photographed while exhibited in the Bihar Museum were collected from various places in Bihar, and are now in the Indian Museum."

**Date**
[1870]

**Source**

---

Item 2

**Title**
Buddhist sculpture pieces from Jamal-Garhi. 1003995

**Creator**
Photographer: Craddock, James

**Subject**
North-West Frontier Province Pakistan Buddha images Gandharan art Indian sculpture Indian sculpture (Buddhist) museum objects South Asia -- History 954

**Description**
Photograph of Buddhist sculpture pieces from Jamal-Garhi. This print shows boxed sculpture fragments. A note with Jamal-Garhi prints reads: "The plates entered here also include photographs taken from sculptures coming from Takht-i-Bahi and Shahr-i-Buhul. No separate arrangement was possible. Nearly all the sculptures coming from these places are now in the Indian Museum, Calcutta."

**Date**
[1880]

**Source**

---

**General Similarity** *(required)*

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completely Unrelated</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Identical</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

**Author Similarity** *(required)*

<table>
<thead>
<tr>
<th></th>
<th>Not Applicable</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completely Unrelated</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>
Typed Similarity Task

Instructions

The aim of this survey is to collect information about how people judge the relatedness of cultural heritage items in an online collection. You will be presented with pairs of cultural heritage items, including an image and additional textual information, and asked to judge how similar you think they are on the following scale:

5 - Identical
4 - Strongly Related
3 - Related
2 - Somewhat Related
1 - Unrelated
0 - Completely Unrelated

For each pair you will be asked to provide a general similarity score, plus an additional score for each of the types of similarity considered, as follows:

- similar author
  (e.g. two items with the same creator should be rated 5 while two items with similar creators should be rated 4-3, etc)
- similar people involved
  (e.g. two items showing the same people should be rated 5, two items showing children should be rated 4, showing similar people 4-3, etc.)
- similar time period
  (e.g. two items from 1914 should be rated 5, from the World War II should be rated 4, etc.)
- similar location
  (e.g. two items that showing scenes of the same street should be rated 5, of London should be rated 4, etc.)
- similar event or action involved
  (e.g. two items showing weddings or people eating an ice-cream should be rated 5, etc.)
- similar subject
  (e.g. two items about cars or cats should be rated 5, etc.)
- similar description (e.g. two items with identical description should be rated 5, etc.)
Typed Task Data

750 pairs for training
750 pairs for test
20 Gold for eval

Annotations obtained via crowd sourcing

ITA >77% overall
Participation

• 34 teams participated amounting to 104 system runs
  • 89 core
  • 14 typed
## Results

<table>
<thead>
<tr>
<th>Task</th>
<th>Baseline</th>
<th>Baseline Rank</th>
<th>Best Result (ITA)</th>
<th>Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core All Data sets</td>
<td>36.39</td>
<td>73/89</td>
<td>61.8</td>
<td>UMBC</td>
</tr>
<tr>
<td>Core-HDL</td>
<td>54</td>
<td></td>
<td>78.4 (80.7)</td>
<td>UMBC</td>
</tr>
<tr>
<td>Core-SMT</td>
<td>28.6</td>
<td></td>
<td>40.3 (37.7)</td>
<td>NTNU</td>
</tr>
<tr>
<td>Core-ON_WN</td>
<td>28.3</td>
<td></td>
<td>84.3 (83.2)</td>
<td>DEFT</td>
</tr>
<tr>
<td>Core-FN-WN</td>
<td>21.5</td>
<td></td>
<td>58.2 (42.7)</td>
<td>UMBC</td>
</tr>
<tr>
<td>Typed Mean</td>
<td>48.9</td>
<td>8/14</td>
<td>76.2 (77)</td>
<td>Unitor-SVR</td>
</tr>
<tr>
<td>Typed General</td>
<td>66.9</td>
<td></td>
<td>79.81</td>
<td>Unitor-SVR</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Task</th>
<th>Baseline</th>
<th>Baseline Rank</th>
<th>Best Result (ITA)</th>
<th>Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core All Data sets</td>
<td>36.39</td>
<td>73/89</td>
<td>61.8</td>
<td>UMBC</td>
</tr>
<tr>
<td>Core-HDL</td>
<td>54</td>
<td></td>
<td>78.4 (80.7)</td>
<td>UMBC</td>
</tr>
<tr>
<td>Core-SMT</td>
<td>28.6</td>
<td></td>
<td>40.3 (37.7)</td>
<td>NTNU</td>
</tr>
<tr>
<td>Core-ON_WN</td>
<td>28.3</td>
<td></td>
<td>84.3 (83.2)</td>
<td>DEFT</td>
</tr>
<tr>
<td>Core-FN-WN</td>
<td>21.5</td>
<td></td>
<td>58.2 (42.7)</td>
<td>UMBC</td>
</tr>
<tr>
<td>Typed Mean</td>
<td>48.9</td>
<td>8/14</td>
<td>76.2 (77)</td>
<td>Unitor-SVR</td>
</tr>
<tr>
<td>Typed General</td>
<td>66.9</td>
<td></td>
<td>79.81</td>
<td>Unitor-SVR</td>
</tr>
</tbody>
</table>

**Observation**

Improvement over previous ON-WN results: 73% ↑84.3%
## Results

<table>
<thead>
<tr>
<th>Task</th>
<th>Baseline</th>
<th>Baseline Rank</th>
<th>Best Result (ITA)</th>
<th>Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core All Data sets</td>
<td>36.39</td>
<td>73/89</td>
<td>61.8</td>
<td>UMBC</td>
</tr>
<tr>
<td>Core-HDL</td>
<td>54</td>
<td></td>
<td>78.4 (80.7)</td>
<td>UMBC</td>
</tr>
<tr>
<td>Core-SMT</td>
<td>28.6</td>
<td></td>
<td>40.3 (37.7)</td>
<td>NTNU</td>
</tr>
<tr>
<td>Core-ON_WN</td>
<td>28.3</td>
<td></td>
<td>84.3 (83.2)</td>
<td>DEFT</td>
</tr>
<tr>
<td>Core-FN-WN</td>
<td>21.5</td>
<td></td>
<td>58.2 (42.7)</td>
<td>UMBC</td>
</tr>
<tr>
<td>Typed Mean</td>
<td>48.9</td>
<td>8/14</td>
<td>76.2 (77)</td>
<td>Unitor-SVR</td>
</tr>
<tr>
<td>Typed General</td>
<td>66.9</td>
<td></td>
<td>79.81</td>
<td>Unitor-SVR</td>
</tr>
</tbody>
</table>

**Observation**

Decrease for SMT sets due to HTER (2013) vs. Ref (2012)  
57-61% ↓ 40.3%
Comparing ON-WN rating distributions 2012-2013

Values distribution for 2012 ON-WN dataset pairs

Values distribution for 2013 ON-WN dataset pairs
Comparing ON-WN rating distributions 2012-2013

Distributions look very different (ideally uniform)
However, no significant confusability across ratings
ITA improved from 0.63 (2012) to 0.83 (2013) - probably due to more bimodal rating distribution
Rating Distributions per data set

Values distribution for 2013 HDL dataset pairs

Values distribution for 2013 SMT dataset pairs

Values distribution for 2013 OnWN dataset pairs

Values distribution for 2013 FNWN dataset pairs
Data distribution especially in 2-4 region leads to more rating confusion.
Data set Observations

• **FN-WN (Significant confusability ratings 1-4)**
  - a certain idiosyncrasy belongs to an entity distinguishing it from other entities
  - unique or specific to a person or thing or category

• **HDL (Significant confusability ratings 3-1)**
  - 13 dead in Iraq bomb attacks
  - 16 killed in north Iraq attacks

  - Chicago teachers strike in bitter contract dispute
  - Chicago teachers union delegates could end strike

  - China's Cabinet promises to boost economy
  - China's Wen promises job creation
**Data set Observations**

- **FN-WN (Significant confusability ratings 1-4)**
  - a certain idiosyncrasy belongs to an entity distinguishing it from other entities
  - unique or specific to a person or thing or category

- **HDL (Significant confusability ratings 3-1)**
  - 13 dead in Iraq bomb attacks
  - 16 killed in north Iraq attacks
  - **Chicago teachers strike in bitter contract dispute**
  - **Chicago teachers union delegates could end strike**
  - **China's Cabinet promises to boost economy**
  - **China's Wen promises job creation**
Resources and tools used

- WordNet, corpora and Wikipedia most used
- Distributional and Knowledge-based tools
- Beyond lemmatization and Pos tagging
  - Syntax
- Machine learning widely used for combination
- Less used
  - LDA, word embeddings
- Best systems not always used most resources
STS 2013

- Another successful task with extremely high participation (104 system runs)
- Typed Similarity Pilot Task was a success
- Currently undergoing extensive analysis
  - Revising Guidelines and annotation presentation
    - Without examples
    - Reversing pair orderings
  - Excluding cases of high confusability and measuring impact on results
- Refining pipeline
2014 STS

- Part of SemEval 2014, Co-located with *SEM co-located with COLING, Dublin August 2014

- More data sets
  - Discussion Forum Data, Twitter, Image caption, More Reference Translation, Twitter-Newswire, viable RTE data (sentence pairs)
  - Coordination with Tasks 1 (Compositional Distributional Semantics), Task 3 (Cross size STS), Task 10 (Spanish STS)

- Improve the Sampling of random sentence pair selection (employing meteor, bleu, word overlap)

- Improve Pipeline

- Refine Guidelines

- Take better advantage of confidence scores

- Attempt viable extrinsic evaluation within MT
A note about data choices

- Targeting already existing annotated data for
  - Mining correlations with STS on syntactic-semantic levels
    - Abstract meaning representations (AMRs)
    - Treebanks/Propbanks from NW, BN, WB, BC
  - Mining correlation with STS on semantic-pragmatic levels
    - Modality: Belief, negation, permission
    - Contradiction: including negation, sarcasm
    - Sentiment & Affect
Beyond 2014

- Toward interpretability (gray boxes) (leveraging 2014 Task 1)
  - Layers of annotation on the same data
- Multilingual data for Cross lingual STS (leveraging 2014 Task 10)
- Different data sizes pairing (leveraging 2014 Task 3)
- Beyond sentence similarity to taking context seriously
  - Dialogue chains and discussion threads
- Robust pipeline: STS Common
- Stable set of extrinsic evaluation tasks
Resources and References

STS Webpages

http://www-nlp.stanford.edu/wiki/STS
http://ixa2.si.ehu.es/sts/

A couple of STS system already packaged ready to use:

check with Weiwei Guo  weiwei@columbia.edu

DKPro system: https://code.google.com/p/dkpro-similarity-asl/
Thank you for your attention!

Questions??