Microsoft Research

Each year Microsoft Research hosts hundreds of influential speakers from around the world including leading scientists, renowned experts in technology, book authors, and leading academics, and makes videos of these lectures freely available.

2016 © Microsoft Corporation. All rights reserved.
Semantic matching against a corpus: new applications and methods

Lucy H. Lin, Scott Miles, Noah A. Smith
27 April 2018
How can we apply recent advances in broad-coverage modeling of sentential semantics?

⇒ Match natural language propositions against a corpus.

<table>
<thead>
<tr>
<th>Possible end user</th>
<th>tracking occurrences of...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historian of science</td>
<td>“vaccines cause autism”</td>
</tr>
<tr>
<td>Political scientist</td>
<td>“immigrants are used as scapegoats for problems in society”</td>
</tr>
<tr>
<td>Public servant</td>
<td>“dealing with authorities is causing stress and anxiety”</td>
</tr>
</tbody>
</table>
Introduction

Proposition query:
"Dealing with authorities is causing stress and anxiety."

query corpus

Matched sentences:
"Unfamiliar bureaucratic systems are causing the majority of the stress."
"Those in charge of recovery are making moves to appease the growing anger among homeowners."

aggregate

Frequency across time:

<table>
<thead>
<tr>
<th>Year</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
</tr>
</tbody>
</table>
Related to...

- Paraphrase (Dolan et al., 2004)
- Entailment (Dagan et al., 2006)
- Semantic similarity (Agirre et al., 2012)

Information retrieval, passage retrieval for QA (Tellex et al., 2003)

Dynamics of language across a corpus (e.g., Blei & Lafferty, 2006)
1. Introduction
2. Problem formulation
3. Matching exemplars from a codebook
   Domain: media framing of policy issues
4. Matching specific expert queries
   Domain: analysis of disaster recovery
5. Using semantic matching output for measurement
6. Discussion/conclusions
Problem formulation

- corpus of sentences, $C$
- proposition query, $s_p$
- matching: $\text{score}(s_p, s)$
- top $n$
- matched sentences, $C_m$

where $\text{score}(s_p, s)$ should be high iff $s$ expresses $s_p$
1. Introduction

2. Problem formulation

3. Matching exemplars from a codebook
   Domain: media framing of policy issues

4. Matching specific expert queries
   Domain: analysis of disaster recovery

5. Using semantic matching output for measurement

6. Discussion/conclusions
Matching exemplars from a codebook

Inputs:

- **C**: Media Frames Corpus (Card et al., 2015)
  - thousands of news articles on immigration (and other policy issues)
  - spans of text annotated with *framing dimensions*

- **s_p**: 30 annotation codebook examples
  e.g.: “immigration rules have changed unfairly over time” evokes the *fairness and equality* frame

Note: many ways to evoke a frame outside of the codebook.
Matching exemplars from a codebook: scoring

Scoring function $f(s_p, s)$:

1. each sentence is the average of its word vectors
2. cosine similarity between two sentence vectors $\rightarrow$ score

We use two word vector variants (300D):

- paraphrastic word vectors (Wieting et al., 2016)
- word2vec (Mikolov et al., 2013) pretrained on Google News
How well does output align with corpus annotations?

**Finding:**
- paraphrastic > word2vec > tf-idf
1. Introduction

2. Problem formulation

3. Matching exemplars from a codebook
   Domain: media framing of policy issues

4. Matching specific expert queries
   Domain: analysis of disaster recovery

5. Using semantic matching output for measurement

6. Discussion/conclusions
Matching specific expert queries

Motivation: Researchers, public servants want to understand community challenges post-disaster.

Inputs:

- C: 982 NZ news articles after the 2010/2011 earthquakes
- \( s_p \): 20 queries provided by domain expert, covering community wellbeing, infrastructure, and decision-making

  e.g.: “The council should have consulted residents before making decisions.”

  \( \Rightarrow \) more fine-grained matching.
Matching specific expert queries: scoring

Scoring function $m(s, s_p)$:

- dependency parses $T, T_p$
- sequence of tree edit operations transforming $T$ into $T_p$
  (Heilman and Smith, 2010)
- classify as match/non-match using tree edit sequence
Matching specific expert queries: scoring

Scoring function $m(s, s_p)$:

⇒ dependency parses $T, T_p$

⇒ sequence of tree edit operations transforming $T$ into $T_p$
  (Heilman and Smith, 2010)

⇒ classify as match/non-match using tree edit sequence
Matching specific expert queries: scoring

\[ T \]
unfamiliar bureaucratic systems are causing stress

? 

\[ T_p \]
dealing with authorities is causing stress
Matching specific expert queries: scoring

$T$

unfamiliar bureaucratic systems are causing stress

$T$

+DELETE(unfamiliar)
+DELETE(bureaucratic)

systems are causing stress

$T_p$

dealing with authorities is causing stress
Matching specific expert queries: scoring

\[ T \]

unfamiliar bureaucratic systems are causing stress

\[ +\text{DELETE}(\text{unfamiliar}) \]
\[ +\text{DELETE}(\text{bureaucratic}) \]
\[ +\text{RELABEL}(\text{systems}) \]

\[ T \]

dealing are causing stress

? 

\[ T_p \]
dealing with authorities is causing stress
Matching specific expert queries: scoring

1. \( T \{\)
   - unfamiliar bureaucratic systems are causing stress

2. \( T \{\)
   - +DELETE(unfamiliar)
   - +DELETE(bureaucratic)
   - +RELABEL(systems)
   - +RELABEL(are)
   - +INSERT(authorities)
   - +INSERT(with)
   - dealing with authorities is causing stress

3. \( T_p \{\)
   - dealing with authorities is causing stress
Matching specific expert queries: scoring

Scoring function $m(s, s_p)$:

- dependency parses $T, T_p$
- sequence of tree edit operations transforming $T$ into $T_p$
  (Heilman and Smith, 2010)
- classify as match/non-match using tree edit sequence
  - logistic regression (39 features), LSTM
  - trained on SNLI (entail vs. neutral/contradiction)
Matching specific expert queries: scoring

For reasonable runtime:

1. $f$: word-vector-based matching on $C$
   $\Rightarrow$ obtain top $k$ matches ($C_f$)

2. $m$: entailment-based model on just $C_f$
   $\Rightarrow$ obtain top $n$ matches ($C_m$)

Four combinations of $f, m$:

- w2v, LR
- w2v, LSTM(w2v)
- para, LR
- para, LSTM(para)
Matching specific expert queries: evaluation

User study:

- surveyed 20 emergency managers
- output of the four $f, m$ combinations obtained per query
- candidate sentences rated 1-5
  (Krippendorf’s $\alpha = 0.784$)
Example $s_p$: There is a shortage of construction workers.

<table>
<thead>
<tr>
<th>Score</th>
<th>Example candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The quarterly report for Canterbury included analysis on Greater Christchurch Value of Work projections.</td>
</tr>
<tr>
<td>3</td>
<td>The construction sectors workload was expected to peak in December.</td>
</tr>
<tr>
<td>5</td>
<td>Greater Christchurchs labour supply for the rebuild was tight and was likely to remain that way.</td>
</tr>
</tbody>
</table>
Matching specific expert queries: study results

Scores by system and category

<table>
<thead>
<tr>
<th>Systems</th>
<th>Scores</th>
<th>top(m)</th>
<th>top(f)</th>
<th>rand(f)</th>
<th>rand(¬f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>para,LR</td>
<td>3.22</td>
<td>2.03</td>
<td>1.06</td>
<td>1.04</td>
<td>1.03</td>
</tr>
<tr>
<td>para,LSTM</td>
<td>3.31</td>
<td>2.08</td>
<td>1.04</td>
<td>1.08</td>
<td>1.03</td>
</tr>
<tr>
<td>w2v,LR</td>
<td>2.79</td>
<td>1.95</td>
<td>1.08</td>
<td>1.03</td>
<td>1.03</td>
</tr>
<tr>
<td>w2v,LSTM</td>
<td>2.81</td>
<td>1.88</td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Findings: paraphrastic > word2vec; m is useful and LSTM > LR if using paraphrastic vectors.
Other feedback:

- 17/20 respondents interested in a way to match ideas in news or other text corpora
- Half of respondents interested in follow-up study, with their own idea queries (in progress!)
1. Introduction
2. Problem formulation
3. Matching exemplars from a codebook
   Domain: media framing of policy issues
4. Matching specific expert queries
   Domain: analysis of disaster recovery
5. Using semantic matching output for measurement
6. Discussion/conclusions
Example measurement:

$s_p$: “Dealing with authorities is causing stress and anxiety.”
Outline

1. Introduction
2. Problem formulation
3. Matching exemplars from a codebook
   Domain: media framing of policy issues
4. Matching specific expert queries
   Domain: analysis of disaster recovery
5. Using semantic matching output for measurement
6. Discussion/conclusions
Discussion

Idea complexity:
- queries that are too general/too specific
- user guidance for writing good queries?

Entities/coreference: e.g., “Cera”

Sentence-related issues:
- surrounding context invalidates a match
- potential match spread across a sentence boundary
In conclusion...

In this talk, we:

- Demonstrated viability of semantic matching methods in two different domains
- Performed a user study to establish end user interest
- Motivated future work on semantic matching/measurement applications
Thanks!

(contact: lucylin@cs.washington.edu)
Syntetic ant Natural Noize doth Breack Neural Machine Translation

Yonatan [Belinkov & Bisk]

*Mistakes generated from Wiki-edits corpus*
Synthetic ant Natural
Noise doth Break Neural
Machine Translation

Yucatan [Blink & Busk]

*Mistakes “fixed” by a spell-checker*
Synthetic and Natural Noise Both Break Neural Machine Translation

Yonatan [Belinkov & Bisk]
What is noise?

**Vision**

```
“panda” 57.7% confidence

+ ε =

“gibbon” 99.3% confidence
```
What is noise?

Vision

\[ \text{“panda”} + \epsilon = \text{“gibbon”} \]

57.7% confidence

99.3% confidence

Natural Language

\[ ? \]

Noise distorts the perceptual process without hindering semantic understanding.

Perception is noisy (e.g. saccades)
What do you actually see?

lawsuit wass filed mach 27th and de preliminary

inestigation wass opend by ttt proseccutor’s
What do you actually see?

- Lawsuit was filed on March 27th and the preliminary investigation was opened by the prosecutor's office.
What do you actually see?

lawsuit was filed on the 27th and the preliminary investigation was opened by the prosecutor's.

Words: 14
Fixations: 9
Perception is noisy and sparse

What about handwritten text?
The Meaning of Ugly Handwriting
Ugly handwriting can indicate a certain amount of emotional baggage.
Geniuses with Ugly Handwriting. The Secrets finally Exposed.
Setup

Models

**Nematus**  
Sennrich et. al *Nematus: a Toolkit for Neural Machine Translation* EACL-Demos 2017  
Used in top performing contributions to WMT shared task  

**char2char**  
Lee, Cho, & Hofmann *Fully Character-Level Neural Machine Translation without Explicit Segmentation* TACL 2017  
Complex encoder with convolutional, highway, and recurrent layers, and a standard recurrent decoder  

**charCNN**  
Basic seq2seq with attention and convolution embeddings  

Data

TED talks parallel corpus prepared for IWSLT 2016  (Cettolo et al. 2012)
Synthetic Errors

Swap
- Switch two adjacent letters from the middle of the word
  - noise \rightarrow nosie

Middle
- Randomly permute all but the first and last letters
  - noise \rightarrow nisoe

Random
- Randomly permute all letters
  - noise \rightarrow iones

Keyboard
- Replace a random letter with one adjacent on the keyboard
  - noise \rightarrow noide
Effects of Synthetic Noise on Char-NMT

- Nematus
- Random
- Swap

- Char2Char
- Random
- Swap

BLEU vs % tokens changed
Natural Noise

Disclaimer: I don’t speak German

Phonetic

Tut → Tud  sieht → zieht  Natürlich → Naturlich

Omission

efahren → erfaren  Babysitter → Babysiter

Morphological

wohnt → wonnen  fortsetzt → fortzusetzen  wünsche → wünchen

Keyboard, etc ...

Agglomerationen → Agromelationen  (omission + letter swap)
Natural Decay
Looks like Synthetic

We max out at 40% token coverage
But, like, spell checkers, amiright?
But, like, spell checkers, amiright?
But, like, spell checkers, amiright?
Training with Noise
Can Training on Noise Generalize?

Noisy Training: Noise → Robust on Clean?

- French
- Czech

<table>
<thead>
<tr>
<th>Method</th>
<th>French</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>Swap</td>
<td>37.5</td>
<td>22.5</td>
</tr>
<tr>
<td>Middle</td>
<td>35</td>
<td>20</td>
</tr>
<tr>
<td>Random</td>
<td>32.5</td>
<td>17.5</td>
</tr>
<tr>
<td>Keyboard</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>Natural</td>
<td>27.5</td>
<td>12.5</td>
</tr>
<tr>
<td>Rand + Key</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>Rand + Nat</td>
<td>22.5</td>
<td>8</td>
</tr>
<tr>
<td>Key + Nat</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>Rand + Key + Nat</td>
<td>17.5</td>
<td>5.5</td>
</tr>
</tbody>
</table>
Can Training on Noise Generalize?

Noisy Training: Noise $\rightarrow$ Robust on Clean?

- French
- Czech

Bar chart showing different training methods and their performance on different data sets.
Modeling Multiple Types of Noise

Average BLEU (tested on all 5 types of noise)

Value Axis

Swap  Middle  Random  Keyboard  Natural  Rand+Key  Rand + Nat  Key + Nat  Rand + Key + Nat
Modeling Multiple Types of Noise

Average BLEU (tested on all 5 types of noise)

Value Axis

Swap | Middle | Random | Keyboard | Natural | Rand + Key | Rand + Nat | Key + Nat | Rand + Key + Nat

French | Czech
Mean Char

Can a model be robust by definition?

Average vectors are structure invariance

\[ \bar{w} = \sum \frac{c_i}{N} \]

Train on Vanilla, Test on Scrambled

Good news

French: 34.26 (20% performance hit)
German: 27.53 (21% performance hit)

Bad news

Czech: 3.73 (86% performance hit)
Model Weights

What happens when you train on noise?

- Keyboard
- Natural
- Random
- Ensemble

Variance

CNNs collapse to meanChar
Thank’s!

Questiones?
Syntactic Scaffolds for Semantic Structures

Swabha Swayamdipta
Sam Thomson
Kenton Lee
Luke Zettlemoyer
Chris Dyer
Noah A. Smith
Semantic Structured Prediction

Ivanka *told* Fei-Fei that she is *inspiring*

- **speaker**: TELLING
- **addressee**: message
- **entity**: SUBJECTIVE INFLUENCE
Semantic Structured Prediction

Ivanka told Fei-Fei that she is inspiring

- speaker: TELLING
- addressee: Fei-Fei
- message: that she is inspiring

Coreference Resolution
Frame Semantics
Role of Syntax

Ivanka told Fei-Fei that she is inspiring

speaker  TELLING  addressee  message

entity  SUBJECTIVE INFLUENCE
Incorporating Syntax

- Syntax in a Pipeline
  - Expensive
  - Cascaded Errors
Incorporating Syntax

- Syntax in a Pipeline
  - Expensive
  - Cascaded Errors

- End-to-end syntax free
  - Syntax still helps... (He et. al., 2017)
Incorporating Syntax

- Syntax in a Pipeline
  - Expensive
  - Cascaded Errors
- End-to-end syntax free
  - Syntax still helps... (He et. al., 2017)

Syntactic Scaffolds
Talk Outline

1. Introduction

2. Syntactic Scaffolds

3. Semantic Structured Prediction
   i. Frame-Semantics
   ii. Coreference Resolution

4. What’s next?
Syntactic Scaffolds

- Multitask Learning with primary and auxiliary tasks
  - Better contextualized representations of spans
- No data overlap required
- Relaxed syntactic prediction
- Scaffold not required at test time
Learning with Syntactic Scaffolds

The joint scaffold objective is:

$$\sum_{(x,y) \in D_{pr}} \mathcal{L}_{pr}(x,y) + \delta \sum_{(x,z) \in D_{sc}} \mathcal{L}_{sc}(x,z)$$

(1)

Local classification decisions:

$$\mathcal{L}_{sc}(x,z) = - \sum_{1 \leq i,j \leq n, |j-i| \leq D} \log p(z_{i,j} | x_{i,j}).$$

(2)
Types of Syntactic Scaffolds

- Phrase Identity
- Phrase Type
  - includes NULL
- Phrase and Parent Phrase Type
- Special Phrase Types (e.g. Noun Phrase and Prepositional Phrase)
Talk Outline

1. Introduction

2. Syntactic Scaffolds

3. Semantic Structured Prediction
   i. Frame-Semantics
   ii. Coreference Resolution

4. What’s next?
Frame-Semantic Semantic Role Labeling

- FrameNet semantic graphs (Baker et. al, 1998)
- **Focus**: Identifying and labeling spans

Ivanka *told* Fei-Fei that she is *inspiring*

- **speaker**
- **TELLING**
- **addressee**
- **message**
- **SUBJECTIVE INFLUENCE**
Semi-Markov CRFs
(Sarawgi et. al., 2004)

Given an input $x$ and a segmentation $s = \langle s_1, s_2, \ldots, s_m \rangle$, model $p(s \mid x)$. Each segment is assigned a factor, and is scored by

$$\phi(s, x) = \sum_{k=1}^{m} \phi(s_k, x).$$

The inference problem is given by

$$\hat{s} = \arg \max_s \phi(s, x).$$

Exact inference takes $O(n d \ell)$, $n$ being the length of sentence, $d$ maximum length of spans, and $\ell$ the number of labels.
Modeling Spans

General Electric  Electric said the  the Postal Service  Service contacted the  the company

Lee et. al., 2017
Scaffolded Frame SRL

\[
L_{pr} = - \sum_{(x,s^*) \in D_{pr}} \log \frac{\exp \Psi(s^*, x)}{Z(x)},
\]

\[
Z(x, s^*) = \sum_s \exp \{ \Psi(s, x) + \text{cost}(s, s^*) \}.
\]

Scaffolded frame-SRL involves plugging in the objective:

\[
\sum_{(x,y) \in D_{pr}} L_{pr}(x, y) + \delta \sum_{(x,z) \in D_{sc}} L_{sc}(x, z)
\]

Note that the corpora to learn syntax and semantics need not be identical.
Frame SRL Results

- Yang & Mitchell
- Baseline
- Phrase Type Scaf.
- Identity Scaf.
- NP-PP Scaf.

Test F1:
- Models
  - 65.5
  - 67
  - 67.7
  - 68.5
  - 68.7
  - 69.1
Talk Outline

1. Introduction

2. Syntactic Scaffolds

3. Semantic Structured Prediction
   i. Frame-Semantics
   ii. Coreference Resolution

4. What’s next?
Coreference Resolution

Ivanka told Fei-Fei Li from Stanford University that she is inspiring

- Identify clusters of mentions referring to the same entity
- Span-based task
- Series of span-classification decisions (Lee et. al., 2017)
Model

\[
\sum_{(x, y) \in D_{pr}} \mathcal{L}_{pr}(x, y) + \delta \sum_{(x, z) \in D_{sc}} \mathcal{L}_{sc}(x, z)
\]

Here, we use the same corpus for both kinds of supervision.
Coreference Resolution Results

- Lee et. al.
- Phrase Type Scaf.
- NP Scaf.

<table>
<thead>
<tr>
<th>Avg F1</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>67.2</td>
<td></td>
</tr>
<tr>
<td>67.7</td>
<td></td>
</tr>
<tr>
<td>67.8</td>
<td></td>
</tr>
</tbody>
</table>
In Summary

- Syntax is extremely useful for semantics 🧡
- Syntax is expensive 😞
- Scaffolds are an inexpensive alternative to incorporate syntactic information
What’s next?

- Tasks which could benefit from syntactic span information, such as NER and extractive question answering
- Syntactic dependency scaffolds
- Semantic scaffolds
Thank You!