Natural Language Processing for Analyzing Disaster Recovery Trends Expressed in Large Text Corpora

**Lucy H. Lin**, Scott B. Miles, Noah A. Smith
University of Washington

19 October 2018
Problem: empirical data describing disaster recovery is scarce
⇒ constrains pre-/post-event recovery planning

Available data source: text corpora (e.g., news articles)

Manual analysis of large text corpora is slow...
⇒ + natural language processing
Introduction

**Proposition query:**
“Dealing with authorities is causing stress and anxiety.”

**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.”

**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>
Outline

1. Introduction
2. Case study: 2010-2011 Canterbury earthquake disaster
3. NLP method for semantic matching
4. User evaluation
5. Qualitative/quantitative output
6. Conclusions
1. Introduction

2. Case study: 2010-2011 Canterbury earthquake disaster

3. NLP method for semantic matching

4. User evaluation

5. Qualitative/quantitative output

6. Conclusions
2010–2011 Canterbury earthquake disaster: timeline

**September 2010:**

- Epicenter: 35km west of Christchurch
- Moderate damage

**February 2011:**

- Epicenter: 10km southeast of Christchurch
- Extremely high ground acceleration
- 185 deaths, thousands of felt aftershocks
2010–2011 Canterbury earthquake disaster: impacts

**Damages:**

- Estimated $40 billion
- Housing: 100k houses in need of repairs
- Water, utilities, road infrastructure: extensive damage

**Recovery groups:**

- Government: CERA, SCIRT (sunset after 5 years)
- Community: Regenerate Christchurch

Recovery still ongoing: public development projects, residential rezoning
2010–2011 Canterbury earthquake disaster: text data

- stuff.co.nz, nzherald.co.nz

Proposition queries: 20 queries, covering
- Community wellbeing
- Infrastructure
- Decision-making

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

2. Case study: 2010-2011 Canterbury earthquake disaster

3. NLP method for semantic matching

4. User evaluation

5. Qualitative/quantitative output

6. Conclusions
Semantic matching

**Goal:** find sentences with similar meaning to the query.

- Needs to be more powerful than word/phrase-level matching.
- Related to information retrieval, but want all matches.
Semantic matching: method overview

- corpus of sentences
  - fast filter
  - likely matches
    - syntax-based model
      - matched sentences
  - proposition query
Semantic matching: method overview

corpus of sentences

fast filter

likely matches

syntax-based model

matched sentences

proposition query
Semantic matching: fast filter

**Goal:** quickly filter out unlikely matches.

Word vector based comparison between two sentences:

- Average of word vectors: 
  \[
  \begin{pmatrix}
    \text{Unfamiliar} \\
    \text{bureaucratic} \\
    \vdots \\
    \text{stress}
  \end{pmatrix}
  \]

- Corpus sentence:
  \[
  \text{corpus sent.}
  \]

- Query sentence:
  \[
  \text{query sent.}
  \]

Cosine similarity:

\[
\text{cosine similarity}
\]
Semantic matching: method overview

corpus of sentences

fast filter

likely matches

syntax-based model

matched sentences

proposition query
**Semantic matching: syntax-based model**

**Finer-grained matching:** take word order/syntax into account.

**Intuition:** transformation between sentences is indicative of their relationship.
Semantic matching: syntax-based model

Candidate: unfamiliar bureaucratic systems are causing stress

Query: dealing with authorities is causing stress

Root: ?
Semantic matching: syntax-based model

candidate

- unfamiliar
- bureaucratic
- systems
- are
- causing
- stress

+DELETE(unfamiliar)
+DELETE(bureaucratic)

query

- dealing
- with
- authorities
- is
- causing
- stress

systems
are
causing
stress

root

?
Semantic matching: syntax-based model

candidate

unfamiliar bureaucratic systems are causing stress

+DELETE(unfamiliar)
+DELETE(bureaucratic)
+RELABEL(systems)

dealing are causing stress

query

dealing with authorities is causing stress
Semantic matching: syntax-based model

- candidate
  - +DELETE(unfamiliar)
  - +DELETE(bureaucratic)
  - +RELABEL(systems)
    - +RELABEL(are)

- query
  - dealing with authorities is causing stress

- root
  - dealing is causing stress

- root
Semantic matching: syntax-based model

candidate

- DELETE(unfamiliar)
- DELETE(bureaucratic)
- RELABEL(systems)
- RELABEL(are)
- INSERT(authorities)
- INSERT(with)

unfamiliar bureaucratic systems are causing stress

dealing with authorities is causing stress

dealing with authorities is causing stress

query

root

root

root
Semantic matching: method overview

corpus of sentences → fast filter → likely matches → syntax-based model → matched sentences

proposition query
Outline

1. Introduction
2. Case study: 2010-2011 Canterbury earthquake disaster
3. NLP method for semantic matching
4. User evaluation
5. Qualitative/quantitative output
6. Conclusions
User evaluation

Questions:

• How good are the sentences matched by our method?
• Do potential users think this kind of tool will be helpful?

User study: 20 emergency managers
User evaluation: output quality

Rated output from 20 proposition queries:

- Different method variants
- Different parts of method:
  - Not selected by filter
  - Selected by filter, but not part of final output
  - Top-scoring output from filter
  - Method output (from syntax-based model)

- 1-5 scale (Krippendorf’s $\alpha = 0.784$)
Query: **There is a shortage of construction workers.**

“The quarterly report for Canterbury included analysis on Greater Christchurch Value of **Work** projections.”

(1: completely unrelated to the query)
Query: There is a shortage of construction workers.

“The construction sector’s workload was expected to peak in December.”

(3: related to but does not adequately express the query)
Query: There is a shortage of construction workers.

“Greater Christchurch’s labour supply for the rebuild was tight and was likely to remain that way.”

(5: expresses the query in its entirety)
User evaluation: results

Best performing system

Average score

Not selected by filter
Selected by filter (unmatched)
User evaluation: results

Best performing system

Average score

- Not selected by filter
- Selected by filter (unmatched)
- Highest-scoring by filter

Scores:
- 1.06
- 2.03
- 3.1
User evaluation: results

Best performing system

Average score

- Not selected by filter: 1.06
- Selected by filter (unmatched): 2.03
- Highest-scoring by filter: 3.1
- Matched by method: 3.22
Other feedback:

• 17/20 respondents interested in measuring ideas in news/other text corpora
User evaluation: round two

Follow-up study:

- Participant-supplied queries (18)
- 7 return participants
- Replicated findings of first user study
1. Introduction

2. Case study: 2010-2011 Canterbury earthquake disaster

3. NLP method for semantic matching

4. User evaluation

5. Qualitative/quantitative output

6. Conclusions
The power system was fully restored quickly.
Recovery trends: example #1

The power system was fully restored quickly.

“Orion Energy CEO Roger Sutton says most of the west of Christchurch now has full restored power.”
The power system was fully restored quickly.

“He had no water but power had been restored in his area.”
Recovery trends: example #1

The power system was fully restored quickly.

"TV3 reports that power has now been restored to 60 per cent of Christchurch."
Recovery trends: example #1

The power system was fully restored quickly.

“It had been unable to access the electricity network to restore power and the situation could remain for the next few days.”
Dealing w/authorities is causing stress and anxiety.
Dealing w/authorities is causing stress and anxiety.

“The initial trauma may be over but [...] Christchurch residents will **endure** at least six years of ‘man-made’ stressors as the region **battles bureaucracy**.” (5)
Dealing w/authorities is causing stress and anxiety.

“Add to this the growing frustration among the new, youthful leaders of the community who emerged in the wake of the quakes.” (3)
Caveats

• Expected topics generally expressed in the output, but not necessarily relationships/quantities
  • Except some domain-specific entities, e.g., CERA & SCIRT

• Measurement plots best explored jointly w/text output (i.e., quantitative & qualitative)

• Small sample size (25 sentences per query)
• Reliance on sentences as unit of match
Outline

1. Introduction
2. Case study: 2010-2011 Canterbury earthquake disaster
3. NLP method for semantic matching
4. User evaluation
5. Qualitative/quantitative output
6. Conclusions
Conclusion

New NLP method to measure propositions in text corpora

- Potential applications: long-term recovery planning, exploratory research
- User study with participant interest in method
- Future work: richer models, further user engagement
Thanks!

Contact: lucylin@cs.washington.edu

Website: homes.cs.washington.edu/~lucylin/research/semantic_matching.html

Funding: National Science Foundation (grant #1541025, graduate fellowship)
(more slides)
In relation to other NLP problems...

- 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)
Pre-trained word vectors:

- word2vec (Mikolov et al., 2013), pre-trained on Google News
- paraphrastic word vectors (Wieting et al., 2015), based off the PPDB
Original model (Heilman and Smith, 2010):

- Extract 39 integer features from tree edit sequence: sequence length, counts of edit types
- Logistic regression (LR) $\rightarrow m(s_\rho, s)$
New variation: input tree edit sequence into a LSTM

Each operation in the sequence is vectorized as:

- One-hot encoding of the operation type
- Word vector $\Delta$ between the sentences pre- and post-operation
  - insert $\rightarrow$ word embedding of new word
  - relabel $\rightarrow$ difference between word embeddings
  - delete $\rightarrow$ negated word embedding of deleted word
Tree edit classifier details

**Training:** SNLI corpus (Bowman et al., 2015)

- 570k pairs of sentences
- labels: entailment, contradiction, neutral
- e.g.,: “A soccer game with multiple males playing.” *entails* “Some men are playing a sport.”

**Mapping to our problem:**

- $s \rightarrow$ premise, $s_p \rightarrow$ hypothesis
- match $\rightarrow$ entailment,
  - non-match $\rightarrow$ contradiction/neutral
Disaster recovery queries

Residents are frustrated by the slow pace of recovery.
The repair programme is on schedule to be completed.
Money for repairs is running out.
The council should have consulted residents before making decisions.
Mental health rates have been rising.
Dealing with authorities is causing stress and anxiety.
Most eligible property owners have accepted insurance offers.
Confidence in Cera has been trending downwards.
Water quality declined after the earthquakes.
The power system was fully restored quickly.
Cera missed several recovery milestones.
Prices levelled off as more homes were fixed or rebuilt.
People are suffering because they’ve lost the intimacy of their relationships.
Coordination between rebuild groups has been problematic.
Few people said insurance companies had done a good job.
Having the art gallery back makes the city feel more whole.
Scirt has spent less money than predicted.
Traffic congestion was severe due to road repairs.

Some of the businesses forced out by the earthquake are returning.

Some of the burden on mental health services is caused by lack of housing.