Extracting Meta Statements from the Blogosphere

Filipe Mesquita
University of Alberta

Denilson Barbosa
University of Alberta
Information overload in the blogosphere

- Spinn3r.com indexes over 1 million posts per hour
  - 1 billion posts every 45 days!

- Users have to cope with information overload.

- There is a need for summarizing the conversation in the blogosphere.

What are people talking about?

mesquita@cs.ualberta.ca
Information networks [Mesquita et al. 2010]

• Nodes are entities.
• Edges are relations.
• Extracted from the ICWSM 2009 Spinn3r Dataset
Information networks

• Allow direct statements only
  ▪ (entity1, relation, entity2)

• Unable to represent more subtle statements:
  ▪ “The AP reported Russia’s conflict with Georgia.”
Meta statement [Yang & Kiffer et al 2003]

• Also known as reification.
• Statements about statements.
  - “The AP reported Russia’s conflict with Georgia.”
Reified information networks

• Nodes or arguments can be either entities or statements

• Provide even richer, more useful networks
  - Ex.: Source, repercussion, consequences and context of a statement
Our problem

- Extract reified networks from natural language text found in blog posts.
**Related work: relation extraction (RE)**

- **State-of-the-art for direct statements:**
  - TextRunner’s O-CRF [Banko & Etzioni 2008]
- **Our work is the first to address meta statement extraction from natural language text.**

<table>
<thead>
<tr>
<th></th>
<th>Traditional RE</th>
<th>Open RE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of target relations</strong></td>
<td>One</td>
<td>All</td>
</tr>
<tr>
<td><strong>Relation-specific training</strong></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>Linear on the number of relations</td>
<td>Constant</td>
</tr>
</tbody>
</table>

Our method
South Ossetia was protected by Russia. North Ossetia is an autonomous republic within the Russian Federation.

---

Blog Posts

Split Sentences → Recognize Entities → Resolve Coreferences → Annotated Blog Posts

Gate

[http://alias-i.com/lingpipe]

LBJ Tagger [Ratinov & Roth, CoNLL'09]

Gate

[http://gate.ac.uk/]

mesquita@cs.ualberta.ca
Baseline approach (O-CRF)

- Extracting statements is a sequence labeling problem:
  - Given a sequence of tokens, produce a sequence of labels:
    - relational token
    - non relational token

Russia is definitely in conflict with Georgia

Relation

Russia conflict with Georgia
Baseline approach (O-CRF)

• Conditional Random Fields (CRF)
  ▪ Sequence of tokens $X$
  ▪ Output sequence of labels $Y$
  ▪ Maximize the conditional probability $p(Y|X)$

• $p(Y|X)$ is learnt from training examples with hand-tagged labels.

• Features:
  ▪ **Token**: “in”, “with”
    • closed classes only (e.g., prepositions and determiners)
  ▪ **Part of speech**: AP/NNP reported/VBD Russia/NNP ’s/POS ...
Our approach (meta-CRF)

• Extends the CRF model to extract relations between statements as well.
The need for additional features

• Part of speech is insufficient for extracting meta statements.

• Unable to differentiate:
  ▪ Meta statement:
    AP reported Russia’s conflict with Georgia
  ▪ No statement:
    AP reported Russia’s conflict with Georgia

• Learning meta statements confuses the model!
  ▪ Tokens with the same features but different labels

• Solution: full parsing.
Our additional features

- Parse Tree & Dependency Tree
  - Path length between a token and arguments
  - Relational tokens are expected to have shorter paths than non-relational tokens

- Argument type
  - “Entity” or “Statement” (instead of “Argument” only)

mesquita@cs.ualberta.ca
Experiment setup

- 100 sentences from ICWSM blog dataset
- 496 pairs of arguments were extracted as examples
  - Tokens manually labeled as relational or non-relational tokens.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Sentences</td>
<td>100</td>
</tr>
<tr>
<td>Examples</td>
<td>496</td>
</tr>
<tr>
<td>Meta statements</td>
<td>107</td>
</tr>
<tr>
<td>Direct statements</td>
<td>111</td>
</tr>
<tr>
<td>No statement</td>
<td>278</td>
</tr>
<tr>
<td>Tokens</td>
<td>1245</td>
</tr>
<tr>
<td>Relational tokens</td>
<td>364</td>
</tr>
<tr>
<td>Non relational tokens</td>
<td>881</td>
</tr>
</tbody>
</table>
Experiment setup (cont’d)

• Baseline: O-CRF
  ▪ Using token and part of speech features only

• Our system: meta-CRF
  ▪ Using all features

• Metric:
  ▪ Accuracy = Correct Labels / Number of Tokens

• Tenfold cross validation
  ▪ 10 non-overlapping partitions
  ▪ Training on 9 partitions and testing in 1 partition
Improvement of our additional features

Combination of features outperformed individual ones.

- Arg Type
- Dependency
- Parse
- All

Improvement
Baseline
Accuracy for different types of statement

Meta-CRF outperformed the baseline by 190%.

86% improvement in direct statements.

Meta-CRF

O-CRF

No statement

Meta statement

Direct statement

All

mesquita@cs.ualberta.ca
Conclusion

• Meta statements allow for richer, more useful information networks.

• Part of speech is insufficient to extract meta statements.
  ▪ Full parsing helps!

• Our method has shown great improvement over O-CRF.
  ▪ Double the accuracy for direct statements
  ▪ Triple the accuracy for meta statements

• Our results indicate that a method aware of meta statements may extract direct statements more accurately.
Future work

• Collect statistics about meta statements
  ▪ Ex.: how often can we find meta statements in blog posts?
• Reduce the effort to produce training examples.
  ▪ Can TextRunner’s self supervision be adapted to meta statements?
• Identify other features from full parsing to improve accuracy.
• Replace heavyweight full parsing by shallow parsing.
Thank you

Questions?

mesquita@cs.ualberta.ca
References

• [Yang & Kiffer’03]
• [Banko’08]
• [Mesquita’10]
Best feature set with average link

Unigrams is simpler, faster and as effective as uni+bigrams

Average link and unigrams are effective
Extracting Information Networks from the Blogosphere:
State-of-the-art and Challenges

BACKUP SLIDES
<table>
<thead>
<tr>
<th></th>
<th>O-CRF</th>
<th>meta-CRF</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta statement</td>
<td>0.271</td>
<td>0.785</td>
<td>189.7%</td>
</tr>
<tr>
<td>Direct statement</td>
<td>0.4392</td>
<td>0.801</td>
<td>82.4%</td>
</tr>
<tr>
<td>No statement</td>
<td>0.9259</td>
<td>0.8965</td>
<td>-3.2%</td>
</tr>
<tr>
<td>All examples</td>
<td>0.71</td>
<td>0.86</td>
<td>20.1%</td>
</tr>
</tbody>
</table>
Current relation extraction systems

- Extract **information networks** from blog posts.

  - Nodes are **entities**.
  - Edges are **relations**.
Terminology

• Argument
  ▪ An entity or statement.

• Statement
  ▪ Triple: (argument, relation, argument)
  ▪ In text: “Russia started the conflict with Georgia.”

• Direct statement
  ▪ A statement about entities only.

• Meta statement
  ▪ A statement about entities or statements.

• Relational tokens
  ▪ Tokens that describe a relation between two arguments.
Our approach

• Extract relational tokens for every pair of arguments.
  ▪ Relational tokens are expected to be found in between the arguments.
    • “The AP reported that Russia started the conflict with Georgia.”
    • This is the case for more than 80% of the sentences [Banko’08]

• We provide an algorithm for enumerating all pairs of arguments.
  ▪ Phase 1: Enumerate pairs of entities only:
    • “The AP reported that Russia started the conflict with Georgia.”
  ▪ Phase 2: Enumerate pairs containing extracted statements:
    • “The AP reported that Russia started the conflict with Georgia.”
Our approach (cont’d)

• Sequence labeling problem:
  - Given sequence of tokens $X$, produce a sequence of labels $Y$.

• Labels $Y$ (BIO encoding)
  - B-REL: beginning of the relation
  - I-REL: continuation of the relation
  - O-REL: not a relation

• Our Conditional Random Fields (CRF) model estimates conditional probability distribution $p(Y|X)$
Features

• Tokens
  ▪ Closed classes only (e.g., prepositions and determiners).
  ▪ Function words are not useful for open relation extraction [Banko’08].

• Part of speech
  ▪ A recent study shows that 95% of the relations follow 8 simple part-of-speech patterns [Banko’08].
Results: rounds

<table>
<thead>
<tr>
<th>Round</th>
<th>O-CRF</th>
<th>meta-CRF</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.78</td>
<td>0.89</td>
<td>14.4%</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>0.89</td>
<td>17.7%</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.89</td>
<td>14.6%</td>
</tr>
<tr>
<td>4</td>
<td>0.72</td>
<td>0.91</td>
<td>25.6%</td>
</tr>
<tr>
<td>5</td>
<td>0.69</td>
<td>0.85</td>
<td>22.7%</td>
</tr>
<tr>
<td>6</td>
<td>0.71</td>
<td>0.83</td>
<td>16.3%</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>0.79</td>
<td>11.6%</td>
</tr>
<tr>
<td>8</td>
<td>0.67</td>
<td>0.89</td>
<td>34.1%</td>
</tr>
<tr>
<td>9</td>
<td>0.63</td>
<td>0.77</td>
<td>21.5%</td>
</tr>
<tr>
<td>10</td>
<td>0.68</td>
<td>0.85</td>
<td>25.0%</td>
</tr>
<tr>
<td>Average</td>
<td>0.71</td>
<td>0.86</td>
<td>20.1%</td>
</tr>
</tbody>
</table>

Over 20% improvement over the state-of-the-art

meta-CRF consistently outperform O-CRF
The need for richer features

• O-CRF relies almost exclusively on part of speech.
• Part of speech is insufficient for extracting meta statements.
• Training examples:

  AP reported Russia’s conflict with Georgia

• Learning meta statements confuses the model!
  ▪ Same token, different labels
• Solution: full parsing.
Our approach (meta-CRF)

- Extends the CRF model to extract relations between statements as well.

AP reported Russia’s conflict with Georgia