A Framework and Tool for Collaborative Extraction of Reliable Information

Graham Neubig\(^1\), Shinsuke Mori\(^2\), Masahiro Mizukami\(^1\)

\(^1\)Nara Institute of Science and Technology
\(^2\)Kyoto University
Background
What is Information Extraction?

- Find useful information from large amounts of noise

Information Source (e.g. Internet)

- Info about Hobbies
- Word-of-mouth Information
- Info. about Events
Information Extraction in Times of Crisis

• Noise is particularly prevalent in times of crisis

Information Source (e.g. Internet)

- Provision of Safety Info. ANPI_NLP Project [Neubig+ 11]
- Requests for Safety Info #99japan Project [Aida+ 13]
- Evacuation Shelters/Rescue Supplies
Necessities for Crisis-time Information Extraction

- **Speed**
  - Necessary to provide information ASAP to those in need

- **Absolute Reliability**
  - Provision of mistaken information could be deadly
  - In general, info will likely require confirmation before consumption

- **Difficult to Predict Needs**
  - Wildfire → Wind, Earthquake → Diapers, Radiation

- **Many volunteers!** [Starbird+10, Neubig+11]

- **Challenge:** How do we let volunteers work efficiently as possible to provide reliable information quickly?
This Work

● We propose a method for efficient extraction of reliable information:
  ● Use machine learning (relevance feedback) to decide which examples to show to annotators
  ● Web-based collaborative interface to allow multiple annotators to work on a single task

● Evaluation on data from Twitter

● Toolkit freely available open source

webigator:
http://www.phontron.com/webigator
Information Extraction Framework
Information Extraction Task

- **Information filtering**: Remove documents with no actionable information
- **Information extraction**: Identify which terms fill slots (e.g. status, location)
- For Twitter, documents are small but numerous, so filtering is a challenge

They really need to open more evacuation areas in Sendai!

They are distributing water at Ishinomaki High School today.

I was able to fill up my car at the gas station at XXX.

Got to the evacuation center, but I'm almost out of battery!
Information Filtering as Classification

- **Binary classification** of “useful or not?”
  - Define features, use machine learning to learn weights
  - Notable for large proportion of negative examples

![Diagram showing normal classification vs filtering with positive and negative examples](image-url)
Constructing a Classifier Requires Lots of Data

<table>
<thead>
<tr>
<th>Little Data</th>
<th>Lots of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Graph showing little data and lots of data" /></td>
<td></td>
</tr>
</tbody>
</table>

Bold = Lots of Data
Active Learning

- Way to create a good classifier efficiently
- Choose examples to annotate based on predictions
Active Learning

- Way to create a good classifier efficiently
- Choose examples to annotate based on predictions
Active Learning

- Way to create a good classifier efficiently
- Choose examples to annotate based on predictions
Active Learning

- Way to create a good classifier efficiently
- Choose examples to annotate based on predictions
Active Learning

- Way to create a good classifier efficiently
- Choose examples to annotate based on predictions
Problems with Unbalanced Data

• In information extraction, almost everything is negative
Problems with Unbalanced Data

- In information extraction, almost everything is negative.
Problems with Unbalanced Data

- In information extraction, almost everything is negative
Problems with Unbalanced Data

- In information extraction, almost everything is negative
Problems with Unbalanced Data

- In information extraction, almost everything is negative
Our Simple Fix

- Small change to example selection criterion
  - **Standard**: Select *low confidence* examples
  - **Proposed**: Select examples with *high probability of being positive*

- Effective when *final human check is necessary*
  - Labeling a positive example = finding a highly reliable piece of information
Our Simple Fix

- Finds many positive examples quickly

- Using these positive examples, learn characteristics that help pick out more
Scaling Up
Too Much Data!

- e.g. Twitter after the Great East Japan Earthquake = peak of 1237 tweets/second
- Problems with:
  - Viewing even the high scoring tweets with one person
  - Rescoring every tweet after each round of learning
Collaborative Web-based Interface

- **Allow multiple annotators to cooperate**

![Diagram of Collaborative Web-based Interface]

- Workers can use Web UI interfaces to display text, submit labels, and contribute to the Information List.
- Text Retrieval/Scoring occurs on the Server, which then communicates with the Internet.

---

**Collaborative Web-based Interface**

- **Allow multiple annotators to cooperate**

![Diagram of Collaborative Web-based Interface]

- Workers can use Web UI interfaces to display text, submit labels, and contribute to the Information List.
- Text Retrieval/Scoring occurs on the Server, which then communicates with the Internet.
### Web Interface

- **Find Information**
- **Keywords**
- **Label Buttons**

If the presented information was useful, press "+", and if not press "-".
If you cannot decide, you do not need to press either. After the current batch is submitted new ones will be displayed.

<table>
<thead>
<tr>
<th>Label</th>
<th>Text</th>
<th>Tweet ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT @</td>
<td>被災した方へ【給水／茨城県神栖市では次の3カ所で給水しています。若松公民館、平泉コミュニティセンター、波崎総合支所。このほかに井戸水の給水を市内19カ所で行っています。尚、土合1号公園前空き地での給水は取り止めとなりました。http: ...</td>
<td>4680161723630720</td>
</tr>
<tr>
<td>RT @</td>
<td>【避難の細かい情報(福島県①)] ▼三春町「三春交流館まほろ」: 原発の近くにある富岡町などから278人避難。食事の不安はないが暖房の燃料不足マ川俣町飯坂小学校: 避難指示が出ている双葉町から650人避難。電気使えず。食事はおどり、歯ブラシや ...</td>
<td>46986480933091648</td>
</tr>
<tr>
<td>RT @</td>
<td>すみません、ちょっとキツイことを言います。避難所に避難されているみなさんが、みなさんは「お客さん」ではありません。辛いのは皆さんです。避難所を運営しているものも多いことなど、すべてボランティアです。食べ物や充電、対応が不備の時にでも、暴 ...</td>
<td>47195705193922560</td>
</tr>
<tr>
<td>RT @</td>
<td>水戸市役所後、水戸市水道部となり中央公園にて飲料水を確保しております。一時家庭6リットル出荷しています。避難場所ではありませんのでご注意ください。水戸市役所付近の避難場所は千葉小学校です。# ...</td>
<td>46148666125320193</td>
</tr>
<tr>
<td>RT @</td>
<td>みなさんが、よく聞いて、このあと、日没が来る。日没がくると逃げられない、真っ暗になると津波が見えない。停電をしていて避難も難しい。夜の避難は掟を増やす。それはいろいろな災害での教訓だ、あはいすぐ避難だ。日没までがポイント ...</td>
<td>4613368809780416</td>
</tr>
</tbody>
</table>
Efficiency Improvements

1) **Simple keyword search filter**

<table>
<thead>
<tr>
<th>Type</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evacuation/Supplies</td>
<td>evacuation area, water supplies, food supplies</td>
</tr>
<tr>
<td>Safety Info Request</td>
<td>contact, cannot, waiting</td>
</tr>
<tr>
<td>Safety Info Provision</td>
<td>contact, safe</td>
</tr>
</tbody>
</table>

2) **Rescoring policy**

- Maintain a sorted list of highly scored examples
- When retrieving next example:
  - Choose the example highest in the cache, rescore
  - After rescoring, still better than second best, return
  - Otherwise, return to beginning
Experiments
Evaluation

- **Compared Methods:**
  - Keyword search
  - Proposed learning-based method

- **Target:**
  - 179M tweets week after Great East Japan Earthquake
  - Three types of info: evacuation/rescue supplies, safety info request, safety info provision

- **Evaluation measure:**
  - Amount of reliable information extracted in 30 mins.
  - Use shared Google Doc as repository for information
Effect of Learning

- Experiments with one annotator for three tasks
- Observable increase in amount of information extracted and accuracy
- Some tasks easier than others
Effect of Collaboration

• Experiments with 1-3 users using same interface

• As expected, increasing users = increasing efficiency
Conclusion

- A method for information filtering that focuses on positive examples
- More effective than simple keyword search
- Remaining challenges:
  - Identification/clustering of duplicates
  - Application to identification of slots as well

webigator: http://www.phontron.com/webigator