A Framework and Tool for Collaborative Extraction of Reliable Information

Graham Neubig\textsuperscript{1}, Shinsuke Mori\textsuperscript{2},
Masahiro Mizukami\textsuperscript{1}

\textsuperscript{1}Nara Institute of Science and Technology
\textsuperscript{2}Kyoto University
Background
What is Information Extraction?

- Find useful information from large amounts of noise
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 $\rightarrow$ Wind, Earthquake $\rightarrow$ 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

| 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:** 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
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](image)
Constructing a Classifier Requires Lots of Data

Little Data

Lots of Data

Bold = Lots of Data
Active Learning

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

- Positive
- Negative
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
A Framework and Tool for Collaborative Extraction of Reliable Information

Web Interface

Find Information

Current Keywords

Keywords

Label Buttons

If the presented information was useful, press "+". 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.

| Label | Text                                                                 | Tweet ID          |
|-------|----------------------------------------------------------------------|-------------------|
| RT @  | 原発の近くにある菅野町などから276人避難。食事の不安はなく生活の燃料不足マツハ侯  | 46986480933019648 |
|       | 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 食事はおぎり |                   |
|       | ゴシラヤ ...                                                               |                   |
| RT @  | 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  | 4719570519322560  |
|       | みなさんが「お客さん」ではありません。幸いには皆同じです。避難所を運営している  |                   |
|       | 方も対応などなく、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 ... |                   |
| RT @  | 水戸市役所後ろ、水戸市水道部となり中央  | 4614866125320193  |
|       | 公園にて飲料水を確保しております。家庭用リットルお渡しています。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波小学校です。# ... |                   |
| RT @  | みなさんが、よく聞いて、このあと、日没が来て、日没がくると逃げられる。真っ暗になると津波が見えない。停電をしていて避難が難しい。夜の避難は犠牲を増やす。それはいろいろな災害での教訓だ。さあ、いますぐ避難だ。日没までがポイント ... | 4613368097980416  |

Submit Labels

Submit Button
1) **Simple keyword search filter**

| Type                      | Keywords                                |
|---------------------------|-----------------------------------------|
| Evacuation/Supplies       | evacuation area, water supplies, food supplies |
| Safety Info Request       | contact, cannot, waiting                |
| Safety Info Provision     | contact, safe                           |

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

![Graphs showing the effect of learning on information extraction and filtering accuracy for different tasks.](image-url)
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