Cross-Lingual Classification Of Crisis Data

Prashant Khare\textsuperscript{1}, Gregoire Burel\textsuperscript{1}, Diana Maynard\textsuperscript{2}, Harith Alani\textsuperscript{1}

\textsuperscript{1}Knowledge Media Institute, The Open University, UK
\textsuperscript{2}Department of Computer Science, University of Sheffield, UK

\{prashant.khare, g.burel, h.alani\} @open.ac.uk, d.maynard@sheffield.ac.uk
People of NSW, be careful because there's fires spreading! Stay safe everyone!

Hundreds of volunteers in Mexico tried to unearth children they hoped were still alive beneath a school's ruins.

Two trucks and one car in the water after a road collapse at Hwy 287 and Dillon. #cowx #boulderflood
Challenges

- High volume of data.
  - Over 1m tweets were posted during the 2017 Hurricane Harvey.

- Overwhelming data for prolong periods.

- The irrelevant and repetitive content is in surplus.

- Variety in types of information
**Aspect- Diverse forms of Crises**

- Man-made disasters
- Quake
- Fire
- Floods
- Hurricane

- Different type of crises generates social media data of a diverse nature.
- Classifiers trained and tested on data from one type of crisis might be unfit to classify data from a new type of crisis.
- Semantic features from knowledge bases improve the performance of cross-crisis classifiers.

People be careful because there's fires spreading! Stay safe everyone!

Terremoto en México devasta gran área y; el número de personas muertas sigue subiendo. La más profunda simpatía para todos.

Sardegna: 17 Morti. perchè non donare incassi di tutte le partite agli aiuti e alla ricostruzione?

एक 35 फीट लंबा पुल बनाया गया था। सैनिकों ने बच्चों और वरिष्ठ नागरिकों सहित 100 से अधिक लोगों को बचाया।
Cross-Lingual Crisis Classification

- How do we identify *relevant* and *irrelevant* information in multi-lingual crisis content?

- Can we develop language agnostic classifiers?
  - Train on one language and detect crisis content relevancy from another language.
Language Agnostic Approaches

1. Translate content to a single language using automated translation approaches.

2. Enrich the content by adding semantics from knowledge graphs.

Annotated Tokens

-烧
-区域
-地图
-周一
-夜晚

Semantically enriched (augmented hypernyms)

- 烧 毁灭
- 周一 工作日
- 夜 晚间
Translation Approach

Create a classification model based on a particular language.

- Collected data from CrisisLex.org—collection of crises oriented tweets.
- Language detected via APIs—langdetect, detectlang, text blob. Selected tweets in English, Spanish, and Italian.
- **Curated training data only in a particular language.**
  - Extracted Statistical Features —via **sPacy** (NLP lib Python).
    - Text length, no. of tokens, POS (nouns, verbs, pronouns), hashtags
  - Trained a model with a machine learning algorithm.
    - SVM classifier with a linear kernel—SciKitLearn Python 2.7.
- **Translated (Google Translate) all new test data to the language of the training data.**
- Classified the test data.
Semantic Expansion Via KGs

Using NER tools, entity linking tools, and semantic databases we can extract semantic features:

- Entity synset sense in Knowledge base.
- Hierarchical Context: Hyponyms, Synonyms
- Dbpedia properties.

**BabelNet Semantics** (from BabelNet)

- **BabelNet Sense**: English labels of entities identified via Babelfy.
- **BabelNet Hyponym**: Direct English hyponyms of each entity (at a distance 1).

**Dbpedia Semantics** (from Dbpedia)

- List of properties associated with Dbpedia URI returned by Babelfy.
  - *subject, label, type, city, state, country* (via SPARQL endpoint).
Extracting Semantics

Available tools for entity extraction and knowledge expansion:

NER
- DBpedia Spotlight
- Alchemy (IBM)
- Babelfy (BabelNet)
- Text Razor NLP API
- Aylien Text Analysis API

Knowledge Bases
- Dbpedia
- YAGO
- BabelNet - a multilingual lexicalised semantic network formed by combining various knowledge resources that can enable multilingual NLP applications.
- WordNet
- Google Knowledge Graph
- Wikidata
Semantic Expansion Via KGs Approach

- Collected data from CrisisLex.org- collection of crises oriented tweets.
- Language detected via APIs— `langdetect`, `detectlang`, `text blob`. Selected tweets in English, Spanish, and Italian.

- Curated training data only in a particular language.

- Extracted Statistical Features - via `sPacy` (NLP lib Python).
  - Text length, no. of tokens, POS (nouns, verbs, pronouns), hashtags

- Semantic Features (via Babelfy, BabelNet, and Dbpedia):
  - BabelNet Semantics
  - Dbpedia Semantics

- Deploy SVM classifier with a linear kernel-
  - SciKit Learn Python 2.7. Chosen after determining statistical significance over other kernels.
### Post A (En)

Sad news to report from #Guatemala - at least 8 confirmed dead, possibly more, by this morning’s major earthquake.

### Post B (Es)

‘Terremoto 7,4 Ricther Guatemala deja 15 fallecidos, casas en el suelo, 100 desaparecidos, 100MIL personas sin luz FO’

### Translate

‘Earthquake 7,4 Ricther Guatemala leaves 15 dead, houses on the ground, 100 disappeared, 100MIL people without light FO’

### Semantic Expansion

<table>
<thead>
<tr>
<th><strong>Entities</strong></th>
<th>News, sadness, dead, describe, earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sense(EN)</strong></td>
<td>News, sadness, <strong>dead</strong>, describe, earthquake</td>
</tr>
<tr>
<td><strong>Hypernyms(EN)</strong></td>
<td>Broadcasting, communication, emotion, feeling, people, deceased, inform, <strong>natural disaster</strong>, geological phenomenon</td>
</tr>
<tr>
<td><strong>DBpedia</strong></td>
<td>dbc:Grief, dbc:Emotions, dbc:Demography, <strong>dbr:Death</strong>, dbc:Communication, dbc:News, dbc:Geological hazards, dbr:Earthquake</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Entities</strong></th>
<th>Terremoto, casas, suelo, luz, fallecidos</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sense(EN)</strong></td>
<td>Earthquake, house, soil, light, <strong>dead</strong></td>
</tr>
<tr>
<td><strong>Hypernyms(EN)</strong></td>
<td>Natural disaster, geological phenomenon, building, Structure, residential building granular material, people, deceased</td>
</tr>
<tr>
<td><strong>DBpedia</strong></td>
<td>dbc:Geological hazards, dbr:Earthquake, dbc:Home, dbc:Structural system, dbc:Light, <strong>dbr:Death</strong>, dbc:Demography</td>
</tr>
</tbody>
</table>
Data

From CrisisLex.org we selected 3 datasets:

- **CrisisLexT26** — 26 crisis events, 1000 labelled tweets in each event
  - 4 labels: Related & Informative, Related & Not Informative, Not Related, and Not Applicable.
  - For binary class dataset merged Related & Informative, Related & Not Informative — Related and other two as Not Related.

- **ChileEarthquakeT1** — 1000 tweets from Chile Earthquake in 2010
  - 2 labels: Relevant and Not Relevant.
  - For binary class dataset Relevant as Related and Not Relevant as Not Related.

- **SOSItalyT4** — Around 5600 tweets across 4 natural disaster events in Italy
  - 3 labels: Damage, Not About Damage, and Not Relevant.
  - Damage and Not About Damage reflected crisis relatedness.
  - For binary class dataset merged Damage/No Damage as Related and Not Relevant as Not Related.
Data

After removing duplicates

- CrisisLexT26: 21378 Related and 2965 Not Related.
- ChileEarthquakeT1: 924 Related and 1238 Not Related
- SOSItalyT4: 4372 Related and 878 Not Related

- Language detection phase.
- Removal of distributional bias in the Related and Not Related categories across each language.

<table>
<thead>
<tr>
<th>Language</th>
<th>Unbalanced</th>
<th>Balanced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Related</td>
<td>Related</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>English (en)</td>
<td>2060</td>
<td>2298</td>
</tr>
<tr>
<td>Italian (it)</td>
<td>813</td>
<td>812</td>
</tr>
<tr>
<td>Spanish (es)</td>
<td>1039</td>
<td>1124</td>
</tr>
<tr>
<td>Total</td>
<td>3912</td>
<td>4234</td>
</tr>
</tbody>
</table>
Language distribution (%) in crisis events data

% distribution

<table>
<thead>
<tr>
<th>Events</th>
<th>EN</th>
<th>IT</th>
<th>ES</th>
<th>OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorado Wildfire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Costa Rica Quake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guatemala Quake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italy Quake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippines Flood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typhoon Pablo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venezuela Refinery</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alberta Flood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia Bushfire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bohol E’quake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boston Bombing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil Club Fire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colorado Floods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colorado Helicopter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LA Train Shoot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LacMegantic Train</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>London Train Crash</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manila Flood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NY Train Crash</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Queensland Flood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russia Meteor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sardinia Flood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Savar Building</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singapore Haze</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain Train Crash</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typhoon Yolanda</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas Explosion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L’Aquila Quake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genova Flood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emilia Quake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chile Quake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiment Design

Feature Models:

- **Statistical Features (SF - baseline)**
- Statistical Features + *BabelNet Semantics* (SF + SemBN)
- Statistical Features + *Dbpedia Semantics* (SF + SemDB)
- Statistical Features + *BabelNet Semantics + Dbpedia Semantics* (SF + SemBNDB)

*Statistical Features (SF), BabelNet Semantics (SemBN), Dbpedia Semantics (SemDB), BabelNet and Dbpedia Semantics (SemBNDB)*
Experiment Design

Apply and validate models in three experimental scenarios:

Monolingual Classification with Monolingual Models:

- Train the model on one language and test it on data from the same language via 5-fold Cross Validation

Cross-lingual Classification with Monolingual Models:

- Train the model on one language and test on data from languages not observed in the training data.

Cross-lingual Classification with Machine Translation:

- Train the model on one language and test on data translated from other languages, to the language used in the training data.
Results – Monolingual Classification Model 5-fold CV

**Unbalanced Data**

- **SF+SemBN**: ΔF 0.54%
- **SF+SemDB**: ΔF -1.1%
- **SF+SemBNDB**: ΔF -0.71%

**Balanced Data**

- **SF+SemBN**: ΔF 0.04%
- **SF+SemDB**: ΔF -1.57%
- **SF+SemBNDB**: ΔF -0.59%

*Statistical Features (SF), BabelNet Semantics (SemBN), DBpedia Semantics (SemDB), BabelNet and Dbpedia Semantics (SemBNDB)*
Results - Cross-lingual Classification with Monolingual Model

Balanced Data

F-1 Score

Train-Test Language

SF  SF+SemBN  SF+SemDB  SF+SemBNDB

SF+SemBN: ΔF 8.26%
SF+SemDB: ΔF 8.71%
SF+SemBNDB: ΔF 9.07%

Statistical Features (SF), BabelNet Semantics (SemBN), DBpedia Semantics (SemDB), BabelNet and DBpedia Semantics (SemBNDB)
## Results - Cross-lingual Classification with Monolingual Model

<table>
<thead>
<tr>
<th>Balanced Data</th>
<th>SF</th>
<th>SF+SemBN</th>
<th>SF+SemDB</th>
<th>SF+SemBNDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>Test</td>
<td>Size</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>en</td>
<td>1224</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>it</td>
<td>401</td>
<td>0.429</td>
<td>0.549</td>
<td>28</td>
</tr>
<tr>
<td>es</td>
<td>401</td>
<td>0.688</td>
<td>0.668</td>
<td>-2.9</td>
</tr>
<tr>
<td>it</td>
<td>1224</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>401</td>
<td>0.521</td>
<td>0.580</td>
<td>11.3</td>
</tr>
<tr>
<td>es</td>
<td>401</td>
<td>0.640</td>
<td>0.647</td>
<td>1.1</td>
</tr>
<tr>
<td>es</td>
<td>1224</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>401</td>
<td>0.578</td>
<td>0.597</td>
<td>3.3</td>
</tr>
<tr>
<td>it</td>
<td>401</td>
<td>0.489</td>
<td>0.532</td>
<td>8.8</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.557</td>
<td>0.596</td>
<td><strong>8.26</strong></td>
<td>0.597</td>
</tr>
<tr>
<td>SDV</td>
<td>0.096</td>
<td>0.053</td>
<td><strong>10.94</strong></td>
<td>0.057</td>
</tr>
</tbody>
</table>

**Statistical Features (SF), BabelNet Semantics (SemBN), DBpedia Semantics (SemDB), BabelNet and DBpedia Semantics (SemBNDB)**
Results- Cross-lingual Classification with Machine Translation

Balanced Data

- **SF+SemBN**: $\Delta F 3.75\%$
- **SF+SemDB**: $\Delta F -1.59\%$
- **SF+SemBNDB**: $\Delta F -0.83\%$

**Statistical Features (SF)**, BabelNet Semantics (SemBN), DBpedia Semantics (SemDB), BabelNet and DBpedia Semantics (SemBNDB)
Results - Baseline Comparison

SF vs SF (translation)

Statistical Features (SF), BabelNet Semantics (SemBN), DBpedia Semantics (SemDB), BabelNet and DBpedia Semantics (SemBNDB)

SF(Translation): ΔF 11.25%
Observations

- Semantics do not aid much when the language is same (5-fold cross validation).

- In Cross-lingual Classification (with monolingual models),
  - Semantics improve the F1 score ($\Delta F1$) over the baseline by 8.26%-9.07% on an average, across all 3 semantic models.
  - SF+SemBN improves over the baseline in 5 out of 6 test cases.

- In Cross-lingual Classification (with Machine Translation)
  - SF (translation) improves over SF-baseline with an average gain of 11.25% in F-1 score.
  - SF+SemBN (translation) outperforms the SF (translation) on an average by 3.75% (with a Std Dev. Of 4.57%).
  - SF+SemDB and SF+SemBD do not improve over the SF (translation).
  - SF+SemBN (translation) improves over the SF-baseline with a gain of 15.23% (with Std Dev. 12.6%).
Feature Correlation Analysis

To understand the impact of *semantics* and the *translation* on the discriminatory nature of the cross-lingual data.

- Calculated **Information Gain** over each dataset (**en**, **es**, and **it** — and also across translated datasets), across all the 4 models, and create ranked list of features.

- For each pair of datasets, say **en-es**, consider common ranked features with **IG > 0**, and calculate **Spearman’s Rank Order Correlation** ([−1, 1]) across the two ranked lists.

![Spearman’s Rank Order Correlation](image)

- **Rank List A**
  - ITEM 1
  - ITEM 2
  - ITEM 3
  - ITEM 4

- **Rank List B**
  - ITEM 4
  - ITEM 1
  - ITEM 3
  - ITEM 2

**Legend**:
- SF
- SF+SemBN
- SF+SemDB
- SF+SemBNDB
- Translation A-B2A
- Translation B-A2B
Summary and Future Course

- **Semantics**, particularly SF + SemBN, improves the classifier for classifying cross-lingual data; the *translation* to the same language (as that of training data) also improves the performance.
  - SF (statistical) model might be sufficient if using *translation*.
  - Statistical + Semantics could be preferred if the translation is not viable (data arriving in unpredicted languages or translation is too inaccurate or expensive).

- The data, for each language, was not discrete in crisis type.
  - Supposedly resulted in vocabulary overlap (such as crisis name, places, people etc.).
  - What if the training and test data was discrete for its crisis type along with the languages? – *We are currently investigating this.*

- 3 languages were chosen (due to limited availability of annotated data).
  - Experimenting with more languages will help in generalising the findings. – *Ongoing exploration.*
Thank you!

prashant.khare@open.ac.uk