Social media content: Challenges

Saptarshi Ghosh
Department of CSE, IIT Kharagpur
Social Computing course, CS60017

Challenges in processing OSM content

- Crowdsourced
  - Large variation in spelling
  - Various ways of expressing same meaning
- Limitations on length
  - Arbitrary abbreviations
  - Words merged to form one word
- Multilingual and transliterated content

- Conventional IR / NLP algorithms tend to perform poorly

OUT OF VOCABULARY WORDS

Study on OOV words in social media

- WASSUP? LOL: Characterizing Out-of-Vocabulary Words in Twitter, Maity et al., CSCW 2016

- Filtered out English tweets from the Twitter 1% random sample over 6 months
- Used an English dictionary to identify OOV words that are stable
  - Occur across all months – so no event-specific terms selected

OOV Categories	Examples
Emoticons	:), :(, :D, :P, :/
Word Lengthenings	noooo, pleaseeee, okk, damnnn
Expressions	haha, uhh, ughh, ahah, grr
Word Shortenings + Abbreviations	lol, omg, yolo, rofl, oomf
Proper Nouns	instagram, miley, bieber, mcdonalds, tumblr
Word Mergings	wassup, iknow, followback

How to deal with OOV words

- Dealing with some categories easier
  - Emoticons
  - Word lengthenings

- Dealing with other categories is more difficult

- Developed a classifier for the other four categories

OOV classifier

- Lexical features
  - Distribution of POS tags of words appearing with OOV word
  - Distribution of named entities (NE) appearing with OOV word
- Content features
  - Length of the OOV word
  - Topic distribution (using LDA)
- Context features
  - Distribution of other OOV words around the OOV word

OOV classifier

- Used SVM and Logistic Regression classifiers
- Achieved 81.26% accuracy

UNIFYING VARIATIONS IN SPELLING

Challenges in processing OSM content

- Crowdsourced
  - Large variation in spelling
  - Various ways of expressing same meaning
- Limitations on length
  - Arbitrary abbreviations
  - Words merged to form one word
- Multilingual and transliterated content

Spelling variations

#Sindhupalchowk 1100+deaths and 99% Houses are Down

Indian national Azhar 23, missing. Last location Sindhupalchok. Plz help.

Food Distribution in sindupalchowk, sufficient for 7 days for 500 victims

Arbitrary shortenings of words

Foreign Secy & Defence Secy giving latest updates on earthquake relief [url]

4 planes to leave for #Nepal tmrw carry meds, med team, 30-Bed Hospital

Nepal quake stresses importance of earthquake resistant bldg designs in entire NCR.

How to handle such variations?

- Traditional technique: Stemming
  - E.g., Porter stemmer
  - Relies on rules of English language
- Will not perform well on arbitrary shortened words on OSM
- Need better methods to understand similarity of two words: (1) string similarity, (2) contextual similarity

Measuring similarity of two words

- String similarity
  - Length of common prefix (has to be at least *p ~ 2*)
  - Longest Common Subsequence of the words

- Contextual similarity
  - Applied Word2vec to get word vectors, where vector of a word is expected to capture the context
  - Cosine similarity of the word vectors

$$Stem_{score}(w, w^*) = \beta * cos\_sim(\overrightarrow{w}, \overrightarrow{w^*}) + (1 - \beta) * LCS_{length}(w, w^*)$$

Contextual stemming algorithm

- For a word *w*
  - Identify a group of words G_w having sufficient string similarity and contextual similarity – candidate stems
  - The word in G_w having minimum length is chosen as the stem of the set $\{w \cup G_w\}$
- Identifies groups of similar words which can be replaced by a common stem

Contextual stemming algorithm

Group of words stemmed to a common stem	Stem
Contribute, contributed, contribution, contributions	Contribute
Donating, donate, donated, donates, donation, donations	Donate
Collapse, collapsing, collapses, collapsed	Collapse
Gurudwaras, gurudwara, gurdwaras, gurdwara	Gurdwara
Organisations, organizations, organisation, organization, orgs, org	Org
Medical, medicine, medicines, medics, meds, med	Med

Contextual stemming algorithm

- Experiments over English tweets posted during 2015 Nepal earthquake

- A set of queries formulated based on discussion with NGOs

- Retrieval from two versions of the data using same retrieval algorithm
  - Stemmed by Porter stemmer
  - Stemmed by contextual stemmer

Contextual stemming algorithm

Query	Average Precision			Recall@1000		
	Unstemmed	Porter	Proposed	Unstemmed	Porter	Proposed
food send	0.1251	0.2356	**0.2542**	0.6214	**0.9660**	0.9563
food packet distributed	0.1930	0.2283	**0.2645**	0.9515	**0.8835**	0.8350
house damage collapse	0.0065	0.0254	**0.0296**	0.2264	0.5283	**0.6226**
medicine need	0.2029	**0.3528**	0.1390	0.4561	0.6140	**0.9298**
tent need	0.1110	**0.5962**	0.5718	0.5195	0.9870	**1.0000**
medicine medical send	0.1806	0.2851	**0.3775**	0.8333	**0.9808**	0.9744
Sindhupalchok	0.4457	0.4457	**0.9493**	0.4457	0.4457	**0.9620**
medical treatment	**0.8003**	0.7998	0.7417	0.8471	0.8471	**1.0000**
medical team send	0.5506	0.7358	**0.7548**	0.9290	0.9484	**0.9935**
NDRF operation	0.7337	0.9006	**0.9065**	0.9653	0.9653	**0.9722**
rescue relief operation	0.5342	0.7205	**0.7440**	0.5846	0.8338	**0.9154**
relief organization	0.2405	0.3015	**0.3293**	0.3448	**0.5460**	0.4598
Dharahara collapse	0.2659	0.6424	**0.9599**	0.7692	0.7692	**0.9780**
epicentre	0.3613	0.3612	**0.9847**	0.3621	0.3642	**0.9853**
gurudwara meal	0.2067	0.6116	**0.8429**	0.2671	0.7671	**0.9795**
All	0.3305	0.4828	**0.5900**	0.6082	0.7631	**0.9042**

A Novel Word Embedding based Stemming Approach for Microblog Retrieval during Disasters, ECIR 2017