Statistical Interactive Learning of Keyword Expansion

JinYeong Bak (jy.bak@kaist.ac.kr)
Imaduddin Amin (imaduddin.amin@un.or.id)
Jong Gun Lee (jonggun.lee@un.or.id)
Alice Oh (alice.oh@kaist.edu)

Motivation & Idea

- Understanding public opinions for crisis events helps to manage the disaster [1]
- Building lexicon of keywords of resource-limited languages rely on the huge feedback from human experts
- Including the explicit human feedback in probabilistic learning process on new data makes better results

Contributions

- Developed Statistical Interactive Learning of Keywords expansion
- Showed a significant performance gain with Indonesian tweets
  - 88M Indonesian tweets from March to June 2014
  - Compared to existing methods
  - Achieved high recall for identifying event related tweets
  - Achieved high tagging efficiency

Crisis Event in Twitter

Haze in Indonesia

- Air pollution caused mainly by forest fires.
- Affects tens of millions of people and happens every year
- People talk of the disaster in Twitter [1]

Identifying event related tweets

- Existing methods
  - SVM
  - Pre-defined keywords by crowd-sourcing
  - Pseudo-relevance feedback
- Limitations
  - Needs training data / large human efforts
  - No explanation for new keywords

Kabut Asap di Ku...ku semakin parah
Gubernur Riau belum ada tindakan
4:04 AM - 3 Mar 2014 from Senapelan, Indonesia

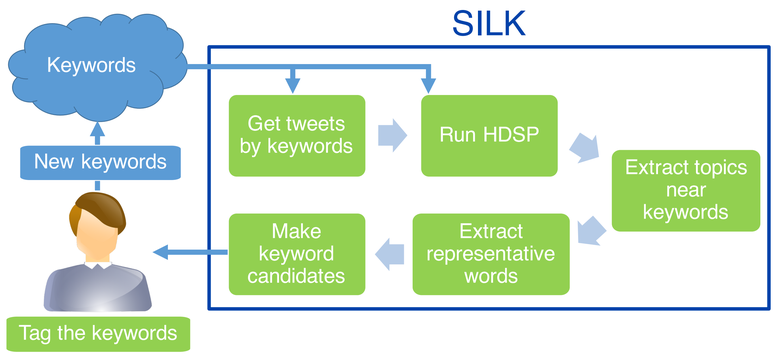
Pasang musik dengan volume max..
Pake masker...
Saatnya bongkar kamar...
… (at Rumah Keluarga Besar B. Siregar) —

SILK

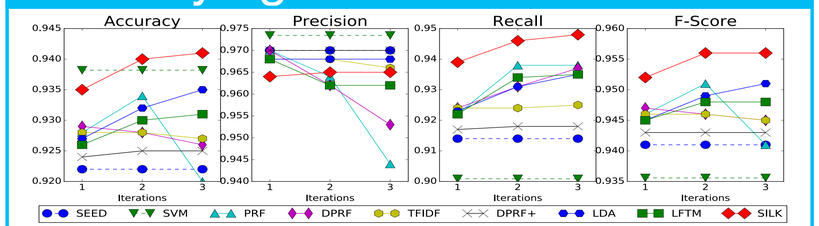
Statistical Interactive Learning of Keywords expansion

- Statistical topic model (HDSP) [2]
  - Make candidate keywords from topics near input keywords
- Iterative method
  - Reuse the output keywords of the previous iteration as input
- Interactive learning
  - Filter keywords by human experts
  - Include the filtered keywords in finding the topics in tweets

Algorithm Overview



Identifying Haze related Tweets



<Classification performance by initial keywords with Mar - June 2014 tweets>

Keywords	asap & #prayforriau (smoke & pray for riau)	hujan & semoga (rain & hopefully)
Tweet	RT" : Jarak pandang hanya 50 M akibat asap #PrayForRiau"	alhamdulillah.. kemaren banyak yg melaksanakan shalat istisqa' & sekarang hujan (Sumbar). semoga di Riau jg hujan.. aamiin
Translation	Visibility of only 50m due to smoke #PrayForRiau	Thanksgiving.Yesterday many prayer perform a ritual for rain & now it rains (West Sumatra). Hopefully in Riau.

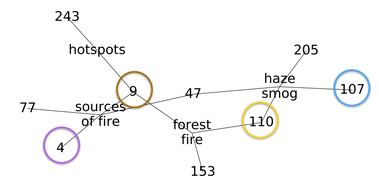
<New keywords and haze related tweets by SILK only >

Efficiency of Human Annotation

	Iter 1	Iter 2	Iter 3	All	Rate
TFIDF	24 / 55	15 / 90	17 / 82	56 / 227	0.25
DPRF+	14 / 43	13 / 24	9 / 22	36 / 89	0.40
LDA	46 / 61	52 / 98	63 / 67	161 / 226	0.71
LFTM	25 / 28	30 / 35	45 / 48	100 / 111	0.90
SILK	51 / 58	79 / 88	44 / 46	174 / 192	**0.91**

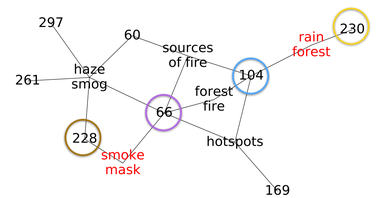
<Tagging efficiency over iterations, #acceptance / #candidates>

Visualising SILK Results



Topic 4	Topic 9
allah (Allah) semoga (hopefully) berkurang (diminish)	polusi (pollution) mengungsi (evacuate) bahaya (danger)
Topic 107	**Topic 110**
masker (mask) parah (severe) tebal (thick)	riau (a province) hujan (rain) semoga (hopefully)

<Seed keywords and its nearby topics in the first iteration of SILK>



Topic 66	Topic 228
parah (severe) sesak (gasping) nafas (breath)	masker (mask) pakai (wear) sabar (patient)
Topic 104	**Topic 230**
alhamdulillah (thanksgiving) tuhan (God)	lebat (heavy) angin (wind) semoga (hopefully)

<New keywords and its nearby topics in the second iteration of SILK>

Indonesian	kabut & asap	titik & panas	sumber & api	kebakaran & hutan	asap & makser	hujan & kebakaran
English	haze smog	hotspots	sources of fire	forest fire	smoke & mask	rain & forest

<Keywords in Indonesian and their English translations>

Future Work

- Make interpretable results (e.g. the reason of suggestions)
- Get more interactive feedback from human
- Apply various crisis such as earthquake and typhoon

[1] Kibanov, Mark, et al. "Mining social media to inform peatland fire and haze disaster management." Social Network Analysis and Mining 7.1 (2017): 30.
[2] Kim, Dongwoo, and Alice Oh. "Hierarchical Dirichlet scaling process." Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014.
Image sources: http://www.bbc.com/news/business-23026219, http://hazegazer.org/