SLSTC at the NTCIR-12 STC task

Japanese subtask

Waseda University
Hiroto Denawa / Tomoaki Sano / Yuta Kadotami / Sosuke Kato / Tetsuya Sakai

Approaches

SLSTC-J-R1

- Create word vectors with Word2vec (using Japanese Wikipedia and Nicopedia as corpora)
- Weight each word vector using tf-idf
- Represent each tweet as a sum of word vectors
- Generate replies using three-layered neural networks.
- Select top five tweets from the repository that are similar to the generated reply.

SLSTC-J-R2

Based on word co-occurrence networks

Network Generation

- Perform morphological analysis for all tweets in the repository
- *V*: set of nodes, each representing a noun from the dataset
- *E*: set of edges defined as follows:
  - For each word bigram in each tweet, an edge goes out from the left word into the right word
  - For each post-reply pair, an edge goes out from a word in the post into a word in the reply

Subnetwork Extraction

- Given a test post tweet, obtain a subnetwork from the above as follows:
  - $V' \subseteq V$: words in the test tweet, plus words connected by an edge $\in E$ outgoing from a word in the new tweet
  - $E' \subseteq E$: edges involving words from V'

PageRank Calculation

- Calculate the PageRank $PR(w)$ of each node $w \in V'$. $PR(w)$ is initially set to $1-d/|V'|$ and updated through 100 iterations as follows:

$$PR(w) = \frac{1-d}{|V'|} + \sum_{w' \in V'(w)} \frac{d * PR(w')}{|E(w')|}$$

$E'(w)$: set of edges from w
d: parameter (set to 0.9 based on a preliminary experiment)

Scoring Tweets

$$idf(w) = \log_{10} \frac{|M|}{|\{m|w \in m\}|} + 1$$

$$tfidf(w,m) = \frac{|m| - tf(w,m) + 1}{|m|} idf(w)$$

$$Score(t) = \sum_{w \in W} tfidf(w,m) * PR(w)$$

M: set of word sequences $\{m\}$ obtained from all tweets in the repository
$tf(w,m)$: number of occurrences of w in m

Select top ten tweets as a result of test tweet

SLSTC-J-R3

Same as R2, except that we removed all continuous occurrences of "w" in the test post tweets (since "ww" "www" in Japanese tweets are similar in meaning to the English "lol" and can be considered as noise)

Results and failure analysis

Results

Not very successful...

Table 1. Official Result (mean accuracy)

	Acc_{L2}@1	Acc_{L2}@5	$Acc_{L1,L2}$@1	$Acc_{L1,L2}$@5
SLSTC-J-R1	0.0381	0.0364	0.1644	0.1650
SLSTC-J-R2	0.0782	0.0332	0.3416	0.1795
SLSTC-J-R3	0.0054	0.0032	0.0391	0.0196
MAX	0.4574	0.3583	0.7817	0.7050
MIN	0.0054	0.0032	0.0391	0.0196

Failure analysis

- Our systems returned the same nonrelevant tweets for many of test post tweets
- Each subnetwork includes too many words
- Morphological analysis often did not perform well due to spelling variations and symbols in the tweets

Table 2. Examples of nouns extracted

test post tweet	nouns extracted	words in subgraph with highest scores	PageRank	IDF	PageRank*IDF
ゆうくりっどさんが言いたいことにプラスして言及してくれてた	さんが, プラス, して	PC ゲーム	0.0000296	4.75	0.000141
		トレシュー	0.0000240	5.23	0.000125
		女川	0.0000237	5.23	0.000124
【自動】 お待たせしました。 7 号線、各駅停車、神戸三宮行き ただいま発車します。	自動, 7 号, 各 駅 停 車, 神戸三宮	やん いってらっしゃいだよっ	0.0000208	5.23	0.000109
		バスターミナル	0.0000187	5.23	0.0000975
		PC ゲーム	0.0000187	4.75	0.0000887

Future Work

- Improve the scoring scheme and setting a threshold for constructing subnetworks
- Improve the morphological analysis dictionary
- Remove symbols (e.g. ゛, ゜,)
- Normalize expressions that are characteristic in microblogs