Reddit Temporal N-gram Corpus and its Applications on Paraphrase and Semantic Similarity in Social Media using a Topic-based Latent Semantic Analysis

DALHOUSIE UNIVERSITY

1. Introduction

- This paper introduces a new large-scale n-gram corpus that is created specifically from social media text.
- Two distinguishing characteristics of this corpus are:
  - monthly temporal attribute.
  - is created from 1.65 billion comments of user-generated text in Reddit.
- The usefulness of this corpus is exemplified and evaluated by a novel Topic-based Latent Semantic Analysis (TLSA) algorithm.
- Unsupervised TLSA outperforms all the state-of-the-art unsupervised and semi-supervised methods in SEMEVAL 2015: paraphrase and semantic similarity in Twitter tasks.

2. The New Reddit Temporal N-gram Corpus

Table 1: Examples of n-grams (1-5) of the newly created corpus about the topic "Donald Trump". Each entry includes the word (n-gram), its frequency, its month, and its year from October 2007 to August 2016.

Reddit temporal n-gram corpus	word	frequency	year	month
1-gram	trump	981	2015	01
2-gram	trump apprentice	31	2015	01
3-gram	donald trump battle	16	2015	01
4-gram	donald trump ignorant tweet	8	2015	01
5-gram	take donald trump advice in	2	2015	01

Table 2: Statistical comparison between Reddit temporal n-gram corpus and its counterparts.

Corpus	1-gram	2-gram	3-gram	4-gram	5-gram
Google web 1T n-gram corpus	13.5M	314M	977M	1.3B	1.12B
Microsoft web n-gram corpus	1.2B	11.7B	60.1B	148.5B	237B
Reddit temporal n-gram corpus	170.2M	1.2B	6.7B	18.4B	30.1B

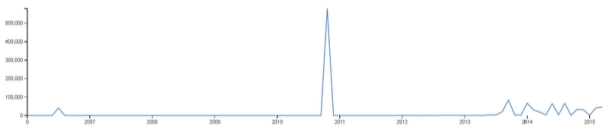

Figure 1: The frequency count of unigram "ISIS" in Reddit from 2007 to 2016. The x-axis represents the year while the y-axis shows the frequency count per month. The highest peak in the graph represents the rise of "ISIS" in October, 2010 following the outbreak of the Syrian Civil War in August, 2010.

Figure 2: An example of word cloud showing the context words of "ISIS" in the 5-grams of the corpus before and after August, 2010. a) "ISIS" is mainly discussed as an Egyptian god before August, 2010, b) "ISIS" means the Islamic State in Iraq and Syria after August, 2010.

3. Topic-based Latent Semantic Analysis

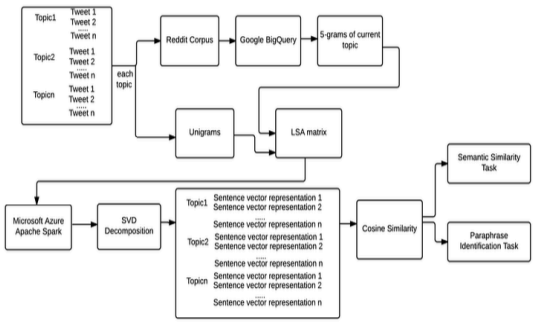

Figure 3: The proposed Topic-based Latent Semantic Analysis using distributed parallel computing, Google BigQuery, and Microsoft Azure Apache Spark. The semantic similarity between two sentences is computed with regard to a specific topic being discussed in two sentences.

4. Evaluations of Paraphrase Identification and Semantic Similarity for Social Media Text

- We use the PIT-2015 Twitter dataset [1].
- TLSA can be extended to any general topic-based datasets.

Table 3: Examples of Paraphrase Identification and Semantic Similarity sentence pairs. All three sentence pairs are about the movie "8 Mile" which is a topic for TLSA. A sentence pair is a paraphrase if its Pearson Correlation score is above 0.6. A sentence pair is a non-paraphrase if its Pearson Correlation score is below 0.6. A sentence pair is debatable if its Pearson Correlation score is equal to 0.6.

Topic	Paraphase	Sentence 1	Sentence 2
8 mile	True	The Ending to 8 Mile is my fav part of the whole movie	Those last 3 battles in 8 Mile are THE shit
8 mile	False	All the home alones watching 8 mile	The last rap battle in 8 Mile nevr gets old ahah
8 mile	Debatable	8 mile is just a classic	After watching 8 mile I feel like such a thug

Table 4: PIT-2015 Twitter dataset. The test data is more balanced than MRPC as it has a higher percentage of non-paraphrase sentence pairs. The unsupervised TLSA only uses the test data for evaluation.

	Sent Pairs	Paraphase	Non-paraphrase	Debatable
Train	13063	3996 (30.6%)	7534 (57.7%)	1533 (11.7%)
Test	972	175 (18.0%)	663 (68.2%)	134 (13.8%)