Twitter Trending Topics Meaning Disambiguation

Soyeon Caren Han[1], Hyunsuk Chung[1], Do Hyeong Kim[2], Sungyoung Lee[2], and Byeong Ho Kang[1]

[1] School of Engineering and ICT
University of Tasmania, Sandy Bay, 7005, Tasmania, Australia
{Soyeon.Han,David.Chung,Byeong.Kang}@utas.edu.au
[2] Department of Computer Engineering
Kyung Hee University, Giheng-gu,Youngin, Korea
{dhkim,sylee}@oslab.khu.ac.kr

Abstract. Twitter is one of the most popular social media services that allow users to share and spread information. Twitter monitors their users' postings and detects the most discussed topics of the moment. Then, they publish these topics on the list, called 'Trending Topics'. Trending Topics on Twitter shows the list of top 10 trending topics but each topic consists of short phrase or keyword, which does not contain any explanation of those meanings. It is almost impossible to identify what a trending topic is about unless you read all related tweets. The goal of this paper is finding the most successful method that uses to retrieve the representative contents of trending topics in order to disambiguate the meaning of topics. We first collected the trending topics and tweets related to them. Then, we applied four types of information retrieval approaches (key factor extraction, named entity recognition, topic modelling, and automatic summarization) for extracting the representative contents of trending topics. We conducted human experiments with 20 postgraduate students.

Keywords: Twitter, Twitter Trending Topics, Social Media, Disambiguation.

1 Introduction

The rise of new types of online social media services (Twitter, Facebook, or YouTube) has caused a human communication paradigm shift and the exchange of unprecedented amounts of information on a wide variety of real-world events. The real-world events are not only the small-scaled local events, but also the large-scaled worldwide events. Twitter is one of the popular online social medias that enables post and share information of the local and worldwide events in real time. Twitter users post the message that is an answer of the question "What is happening now?" asked by Twitter. The answer should be limited to 140 characters. These short messages, called tweets, reflect the information of real-world event from the users' point of view. Several previous studies proved

Y.S. Kim et al (Eds.): PKAW 2014, LNCS 8863, pp. 126–137, 2014.

that these tweets are useful for notifying real-world events of different types and scale, regardless of the event types [1,7,12]. This unprecedented amount of events information from users also attracts Twitter's attention. Since then, Twitter has been monitoring and detecting the keywords and hash tags that are most often mentioned and discussed by their users. The detected top 10 popular topics of the moment are published in the list, called "Twitter Trending Topics". The 'trending topics' ist is located on the middle-right side of the Twitter interface, and displays trending topics from several small cities to worldwide. Twitter Trending Topics is now considered as the guide that displays what kind of event information is currently spread on Twitter. Several researchers also regard that the service is very useful for identifying both local and worldwide events [4].

However, there is one big problem in using the Twitter Trending Topics. As trending topics consist of short phrases, keywords, or hash tags, most of them are the term in an ambiguous sense. Lets assume that 'Galaxy' is one of top 10 trending topics. It is very difficult to define whether the topic is about the phone/tablet made by Samsung or a large group of the stars and planets. Without reading and analysing all related tweets of the topic 'Galaxy', it is almost impossible to fully-understand the exact meaning of the topic. In order to solve this issue, several researchers have investigated summarizing trending topics both manually and automatically[10,3,13]. The most popular trending topics summarisation site is 'What the Trend[1]' which provides the interface for users to manually type the explanation of what the trending topic is about. It is like a Wikipedia for Twitter Trending Topics. However, most of trending topics are not explained properly, but filled with spam and irrelevant content. Then, several researchers were applied automatic summarisation approaches to summarize the definition of trending topics. However, they focused on the readability of the summarized sentences, rather than the quality of content for disambiguating the exact meaning of specific topic.

In this paper, we focus on the finding successful method to retrieve the representative contents from related tweets of a specific trending topic for disambiguating the trending topics sense. We examine the ability of trending topics sense disambiguation, not the readability by human. For the evaluation, we selected four successful approaches in information retrieval area, including key factor extraction, named entity recognition, topic modelling, and automatic summarisation. The performance was evaluated by 20 postgraduate students in computing and information systems. Based on the evaluation result, we address the successful approach for twitter trending topics sense disambiguation.

The contributions of this paper are summarized as follows:

- This paper provides the first proper human evaluation results of the Twitter Trending Topic sense disambiguation
- This paper is the step forwards improving the performance of information retrieval research using trending topics as data

[1] HootSuite Media, Inc. 2011 `http://www.whatthetrend.com/`

2 Related Works

Twitter is one of the most popular social media services. It allows users to share and spread their interests in 140-characters short messages. It provides various ways for users to communicate with others by using their unique symbols, including @ or #. Twitter is aware of the value of these social data so they monitor and collect their data. Based on the data, Twitter currently extracts and displays real-time issues and events to the public by providing the service, called 'Twitter Trending Topics'. Twitter is now considered as very useful service for monitoring and detecting real-time events from local to worldwide level. According to this, several researchers are examined Twitter data for their researches as follows.

Event Detection and Extraction in Twitter. As real-time social data on Twitter is opened to the public, many researchers used the data to detect and extract real-time events. Benhardus and Kalita [1] used Twitter data to detect spikes in usage that related to particular topics: short-term, high intensity discussion in response to a recent event by applying statistical techniques, including TFIDF, normalised TF. Naaman et.al [7] developed taxonomy of trends present in the dataset, and then identified the dimensions that could be used to characterize the data. For characterizing tweets, they discovered various features, including content, interactivity, temporal features, participation level, and the level of reciprocity. Weng et al. [12] used Twitter data to detect real-time events by analyzing streams of tweets. They detected real-time events by clustering signals together using modularity-based graph partitioning. TEDAS (Twitter based Event Detection and Analysis System) is the system that analyses spatial and temporal patterns of events and identify their importance [6]. The system used several rules for classifying tweets and predicting the location of events. Although many researchers examined events mining in Twitter, Twitter detect and extract real-time event topics by applying their own algorithm, and provide the list of top 10 trending topics on their interface, and the list is called 'Twitter Trending Topics'.

Trending Topic Summarisation and Classification. It becomes very popular to detect real-time events in Twitter. 'Twitter Trending Topics', real-time event detection service provided by Twitter, attracts a lot of attention. Trending topics in Twitter are the most often mentioned or posted short phrases, words, and hash-tags but no detailed explanation. Because of this, it is almost impossible to fully understand the exact meaning and the content of the event topic. Hence, many researchers aimed to reveal the exact meaning of trending topics. Sharifi [10] applied phrase reinforcement algorithm to summaries related tweets of Twitter Trending Topics. Then, the author conducted evaluation for comparing hybrid TFIDF and phrase reinforcement in use of Trending topics summarising. Inouye [3] also conducted an experiment to compare twitter summarisation algorithms. They found that simple frequency-based techniques produce the best performance in tweets summarisation. Sport events are one of the popular types in twitter trending topics so Nichols summarises the sport events. All researches in trending topics summarisation applied ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, which is extremely popular evaluation method in automatic summarisation area. Those metrics are for evaluating the quality of a summary, such as the coherence, conciseness, grammaticality, or readability. However, they are not very evaluating whether the summary contains enough contents to fully understand what the trending topic is about. Some researchers examined classifying trending topics. Lee et al. [8] classifies trending topics into general 18 categories by labeling and applying machine-learning techniques. Zubiaga et al. [13] aimed to classify trending topics by applying several proposed features and used SVM to check the accuracy. However, those researches aimed to extract the abstract of twitter trending topics but not the exact meaning.

Approaches in Information Retrieval. In order to find successful approach to retrieve the representative contents for twitter trending topics sense disambiguation, we reviewed several well-known approaches in the general information retrieval field. Statistical key factor extraction with term weighting is widely applied to retrieve the important key terms in a document [9]. The representative keywords of a document can be the objects in it. Named Entity Recognition (NER) has been used for labeling the name of objects in a document, such as person, organisation, or location [8]. For this, topic model has been introduced for extracting the abstract topic that in multiple documents, Blei et al. [11] proposed Latent Dirichlet Allocation (LDA), which is a probabilistic model and present it as a graphical model for discovering the topics.

3 Data Collection

For our studies, it is necessary to collect trending topics on Twitter and tweets related to those topics. Twitter provides an API (Application Programming Interface) that allows developers or researchers to crawl and collect the data easily. Through this API service, we collected twitter trending topics in 3 years (until 30th June, 2014)

3.1 Trending Topics

Twitter monitors all users data and detects the popular trending topics that most people are currently discussing about. The detected popular trending topics are displayed on the service 'Twitter Trending Topics'. This trending topic service is located on the sidebar of Twitter interface by default so it is very easy for users to check the current trending topics and discuss about it. It provides top 10 trending topics in real time. Hence, we have collected those top 10 trending topics per hour using Twitter API. In total, we have collected 105354 unique trending topics in 3 years.

Trending topics in Twitter consist of short phrases, words, or hash-tags. Twitter never provides any detailed explanation of trending topics so it is very difficult to identify the meaning of trending topics until you have a look related tweets of those topics. For example, when a missile destroys Malaysian Airlines, the trending topics were 'Malaysia Airlines', 'Malaysian', etc. It is almost impossible to realise what happened to the Malaysia Airlines by only checking the trending topics. In order to reveal the exact meaning of each trending topic, we need to collect not only the trending topics, but also the related tweets of those topics.

3.2 Related Tweets

The goal of this paper is finding novel method to disambiguate the exact meaning and content of trending topics. To achieve this goal, it is necessary to collect the appropriate related tweets of a specific trending topic. The related tweets should not contain the contents that are irrelevant. If the trending topic is 'Malaysia Airline' which is about a missile attack happened on July 18th, we should not collect the related tweets about missing Malaysia Airline occurred on March 8th. It is extremely important to distinguish the tweets that are related to a specific trending topics. Twitter API provides the tweet/search crawling service that allows users to collect the tweets by using the search query. The concept of tweet/search service is same as the search engine. Users can search the tweets that contain the search keyword. The search results contain detailed information of each tweet, including content, username, location, created date-time, and etc. We used this created date-time to extract the appropriate tweets for the trending topics. As we collect the top 10 trending topics in an hourly basis, we search and collect the related tweets that users upload in last one hour. For example, when Malaysia Airline is on the trending topics list at 8pm, we search and collect the related tweets that users upload in last one hour, 7pm to 8pm. This collecting approach prevents irrelevant tweets.

4 Selected Approaches

As mentioned before, the aim of this paper is to find novel method for disambiguating the exact meaning of the trending topics in Twitter. We focused on examining whether the methods are sufficient to extract the appropriate contents that represent the specific trending topics. We experimented with four different methods that are applied in topic-sense disambiguation research field: Key Factor Extraction, Named Entity Recognition, Topic Modeling, and Automatic Summarisation. The philosophies behind these four methods are very different, but each has been shown to be very effective in the information retrieval area.

4.1 Key Factor Extraction: Term Frequency Weighting

The first selected method for twitter trending topic sense disambiguation is the key factor extraction by applying numerical statistic. There are several key factor extraction approaches that are aimed to find the most important keywords in the document by calculating the importance weights of each word.

TF (Term Frequency) weighting is a classic key factor extraction technique for automatic determination of term relevance [9]. The term frequency in the given tweets gives measure of importance of the term within the particular document. TF weighting is a classic approach but still widely used in Information retrieval area. TF can be determined the exact values in various ways, such as raw frequency, boolean frequency, logarithmically scaled frequency, and augmented frequency. We used raw frequency calculation, which is the most classical approach. The TF weighting tf(t,d) can be calculated by counting the number of times each term occurs in a document.

$$tf(t,d) = \frac{f(t,d)}{max\{f(w,d) : w \in d\}} \tag{1}$$

However, like most English sentences do, tweets include several common words, such as 'the' or 'a'. Assume we calculate TF weights for all terms in documents including those extremely common terms. Since the term 'the' is too common, the result will point the term 'the' as the most important word. In order to solve this issue, we eliminate all stop-words from tweets. The list of stop-words we used is based on the 'Full-Text Stopwords in MySQL'. After removing those stop-words, we applied TF weighting to identify the important terms in the related tweets of each specific trending topic.

4.2 Named Entity Recognition: CRF Sequence Model

Named entity recognition (NER) is widely used for labelling the name of objects in documents. It labels sequences of terms, which are about the name of objects, such as person, organisation, or location. By recognising named entities, it can be easy for people to identify what kind of subject/topic the document is discussing about. We applied one of the most popular Named Entity Recognition approach, Conditional Random Field (CRF) sequence model. CRF-based NER are investigated by Stanford NLP lab [8] and it is widely used as a standard NER technique. CRF is a type of probabilitstic sequence model, and it is applied for sequential data labelling. The basic idea of CRF sequence model is as follows. Assume X is a random variable over data sequences to be labelled, and Y is a random variable over corresponding label sequence. The nodes in the model are seperated into two different sets, X and Y. A conditional distribution p(Y—X) with an associated graphical structure will be modeled.

CRF-based NER models are trained by the official sources such as dictionary or WordNet. The applied CRF model for this study is trained on the CoNLL English training data [2]. For extracting the named entities in related tweets, we applied this trained 4 classes CRF model that contains the entity information of person, location, oragnisation and misc.

4.3 Topic Modelling: Latent Dirichlet Allocation

Topic Modelling is the approach that discovers the abstract topics in the multiple documents. The discovered topics consist of a cluster of words that frequently occur. LDA (Latent Dirichlet Allocation) is the most successful approaches in topic modelling area [11]. The concept of LDA can be explained with the following example. If multiple documents are randomly mixed over various types of topics, the topic can be characterized by a distribution over words. LDA is very different from the traditional Dirichlet-multinomial clustering model. Like many other clustering models, traditional clustering model does not allow a document to being clustered with a single topic. However, LDA has three levels, and notably the topic node is sampled repeatedly within the document. Under this model, documents can be associated with multiple topics.

We used LDA approach in Mallet Topic Modelling tool for training and testing the representative content extraction, with all parameter set to their default values.

4.4 Automatic Summarisation: SumBasics

Automatic Summarisation was introduced for people to save the document reading time by providing a summary that retains the most important points of the documents. There are two main approaches, extraction and abstraction, in automatics summarisation. According to the evaluation conducted by Inouye and Kalita [3], most extraction approaches produced better performance; especially SumBasic had the highest scores in ROUGE metrics. SumBasic is a frequency based summarisation system, which uses the following algorithm. First, it calculates the probability distribution over the words in the input data. For each sentence in the input, assign a weight equal to the average probability of the words in the sentence. Then, select the highest scored sentence that contains the best probability word. For each word in the chosen sentence, update the probability. If the desired summary length has not been reached, go back to the first step. In this paper, we applied SumBasics to extract the summary of related tweets of each specific trending topic.

5 Evaluation

5.1 Evaluation Set-up

After collecting data and selecting the approaches to apply, we conducted evaluation to find the most successful approach in twitter topic sense disambiguation. Unlike other studies, we do not focus on the readability or conciseness of extracted content, but examine which approach can extract the most relevant and representative content of a specific trending topic. As mentioned in section 3, we collected 105,354 twitter trending topics and tweets related to them in 3 years. With this in mind, we randomly selected 100 different trending topics and related tweets for each topic. Then, we selected four different types of information retrieval approaches, including Key Factor Extraction (TF), Named Entity Recognition (CRF sequence model), Topic Modelling (LDA), and Automatic Summarisation (SumBasics). Each selected approach disambiguates the sense of trending topics in their own way by using the related tweets.

Table 1 shows the example contents extracted from the related tweets of a trending topic ‘Susan Powell’. Those contents display the result of applying four different information retrieval approaches. The trending topic ‘Susan Powell’ was on the list in 7th February 2012. It was about the following news. Josh Powell, husband of missing Utah woman, killed himself and his two young sons in Washington house fire. He was a murder suspect of his wife. You can find specific information about the topic from the result of KFE with TF.

Table 1. The contexts extracted of a specific topic by applying four algorithms

Approaches	Extracted Contents
KFE with TF	Susan, Powell, Josh, powell, Utah, sons, woman, killed, Cox, boys, doubts, fate, missing, death, PollyDad, Charlie
NER with CRF	[Susan/P, Candlelight/P, Washington/P, Cox/P, Cheyenne/P, Utah/L, Miller/P, It/P, Powell/P, Charlie/P, Husband/P, City/L, Powell/P, Brandon/P, WEST/L, Wash/P, VALLEY/L, Josh/P, Mommy/P, Marc/P, Denise/P, Candlelight/L, Klaas/P, Kids/P, Dad/P, CITY/L, West/L, Valley/L, Tacoma/P]
TM with LDA	susan, family, lovely, watched, flips, middle, black, husband, children, afternoon
AS with SumBasic	josh powell Any doubts about Susan Powells fate should be dispelled in lieu of Josh Powells homicidal binge.

* The trending topics for this example is ‘Susan Powell’

For evaluating the performance, extracted contents of each 100 trending topic is assessed by 20 postgraduate students in Computing and Information Systems. All students are trained by attending 2 hours workshop for this evaluation. In the workshop, students are encouraged to understand that the evaluation is focusing on the quality of the extracted contents for trending topics, not the readability of the contents. For this evaluation, we developed the evaluation system as can be seen in the figure 1. Figure 1 displays the user interface after a student logged into the system.

In the workshop, participants are asked to use the evaluation system in the following order:

1. Choose one of the 100 topics on the top left section, Trending Topics
2. After selecting a specific topic, the related tweets will be shown on the top right section, Related tweets
3. Click any related tweet to read. The content of each tweet will be displayed on the middle section. By reading those related tweets, get the point what the trending topic is about.
4. After fully understanding the specific topic, grade based on the content extracted by four different approaches in the sense disambiguation area.

In the evaluation, the grades are given based on Likert scale (from highest to lowest) 1,2,3,4, and 5. The meanings of those five grades are as follows: 1=strong agree, 2=agree, 3=neutral, 4=disagree, 5=strong disagree. By using this grade, student give grades for each extracted contents. If a student agrees with only the output of the KFE with TF, but strong disagree with 3 other results, they can give grade '2' to KFE and '5' to the content of all 3 other approaches. The evaluation was successfully conducted in 10 days.

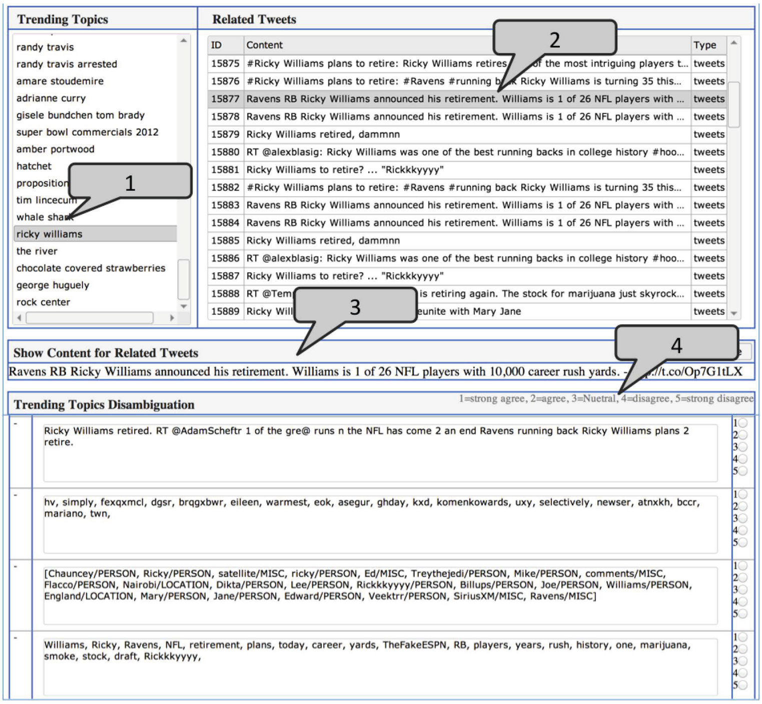


Fig. 1. The human evaluation system interface

5.2 Results

First, we begin the analysis of result with the following question: How much score each approach receives in average? As can be seen in the table 2, the average score for all 4 selected approaches are between 2(agree) and 3(neutral) level.

Table 2. Average Liert Score for each approaches

Approaches	KFE with TF	NER with CRF	TM with LDA	AS with SumBasic
Average score	2.12	2.66	2.90	2.49

It seems all 4 approaches are generally acceptable for twitter trending topics sense disambiguation, and the key factor extraction receives the highest grade among those approaches. However, the average score is not enough to define the successful approach. A graph in figure 2 shows that the distribution of the responses on each approach. The graph clearly indicates that only few participants (less than 10%) strong disagree with the output of sense disambiguation for all evaluated approaches, except topic modelling. The participants roughly understand the meaning of twitter trending topics with the extracted contents of all chosen approaches. As we have seen in the average scores, Keyword Factor Extraction (KFE) got the highest (almost 80%) positive responses (for both strongly agree or agree) and it can be clearly seen from the distribution as well. It is a quite interesting result that the participants provided positive responses with the output of KFE that is extracted based on the classical term frequency weighing technique. However, we found that the contents from Named Entity Recognition (CRF) and Automatic Summarisation (SumBasic) are not clear enough, since the neutral responses took the biggest percentage in those approaches.

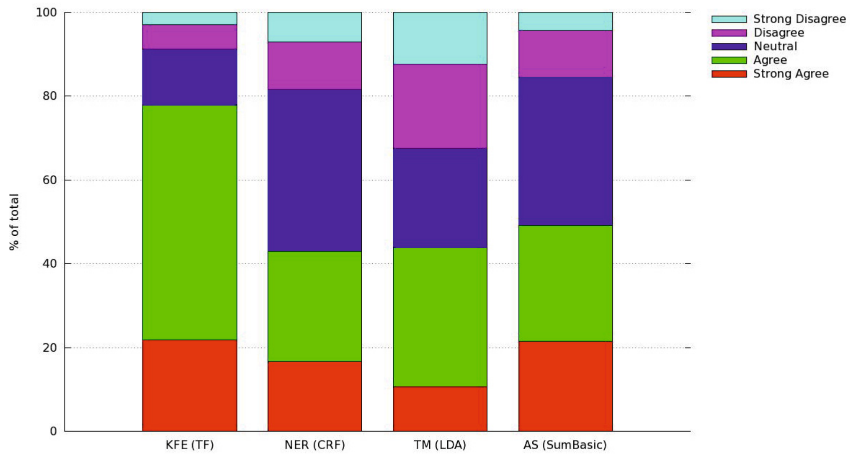
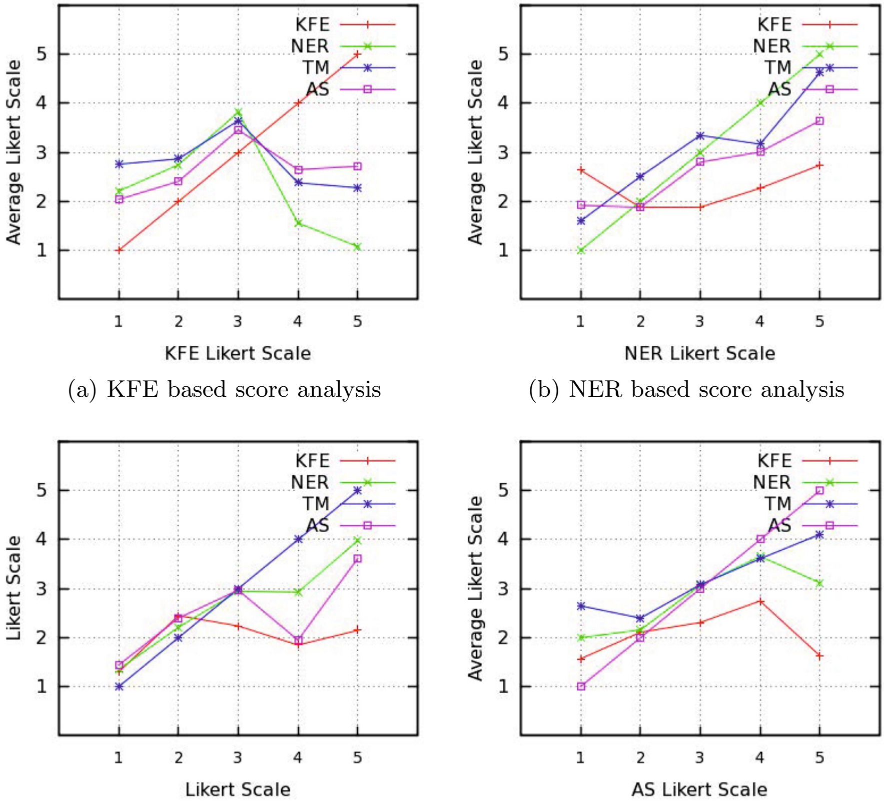


Fig. 2. The grade distribution for four different approaches

Based on the result shown in the above figure, it seems KFE is the good approach to extract the representative contents of the twitter trending topic. However, the KFE result is not fully covered, as there are few amount of neutral and negative response. About this issue, we analysed and found that there is a high correlationship between all four approaches. As you can see from the figure 3, KFE has inverse correlationship with all 3 different approaches in its negative responses (disagree-4 and strong disagree-5). This indicates that other approaches can be used as a substitute of KFE, when the extracted contents are not satisfied. However, this would require an analysis of twitter trending topics to find proper conditions for interchangeability and this is left for our future work.

Likert Scale	1=strong agree, 2=agree, 3=neutral, 4=disagree, 5=strong disagree



(a) KFE based score analysis (b) NER based score analysis

(c) TM based score analysis (d) AS based score analysis

Fig. 3. The grade correlation analysis among four different approaches

6 Conclusion

As mentioned before, the goal of this paper is finding the most successful method to retrieve the representative contents for twitter trending topics sense disambiguation. In order to achieve the goal, we first collected the trending topics and tweets related to them. Then, we applied four different information retrieval approaches, including key factor extraction, named entity recognition, topic modelling, and automatic summarisation. We conducted human experiments with 20 postgraduate students. Based on results reported in the paper, the statistical key factor extraction approach, a classical term weighting technique, provides the highest performance in retrieving the most representative contents for trending topics sense disambiguation. However, topic modelling does not work well on finding the topic words from those real-time events information. As mentioned before, we present the result of the first human evaluation in online trending topic sense disambiguation. We hope the paper is the step forwards improving the performance for any researches using trending topics as a data.

Acknowledgement. This paper was supported by Asian Office of Aerospace Research and Development (AOARD), Japan.

References

1. Benhardus, J., Kalita, J.: Streaming trend detection in twitter. International Journal of Web Based Communities 9(1), 122–139 (2013)
2. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics (2005)
3. Inouye, D., Kalita, J.K.: Comparing twitter summarization algorithms for multiple post summaries. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom). IEEE (2011)
4. Kwak, H., et al.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. ACM (2010)
5. Lee, K., et al.: Twitter trending topic classification. In: 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW). IEEE (2011)
6. Li, R., et al.: Tedas: A twitter-based event detection and analysis system. In: 2012 IEEE 28th International Conference on Data Engineering (ICDE). IEEE (2012)
7. Naaman, M., Becker, H., Gravano, L.: Hip and trendy: Characterizing emerging trends on Twitter. Journal of the American Society for Information Science and Technology 62(5), 902–918 (2011)
8. Ritter, A., Clark, S., Etzioni, O.: Named entity recognition in tweets: An experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2011)
9. Russell, M.A.: Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. O'Reilly Media, Inc. (2013)
10. Sharifi, B., Hutton, M.-A., Kalita, J.: Summarizing microblogs automatically. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics (2010)
11. Weinshall, D., Levi, G., Hanukaev, D.: Lda topic model with soft assignment of descriptors to words. In: Proceedings of the 30th International Conference on Machine Learning (2013)
12. Weng, J., Lee, B.-S.: Event Detection in Twitter. In: ICWSM 2011, pp. 401–408 (2011)
13. Zubiaga, A., et al.: Classifying trending topics: A typology of conversation triggers on twitter. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM (2011)