Identification of Implicit Topics in Twitter Data Not Containing Explicit Search Queries

Suzi Park **Hyopil Shin**
Department of Linguistics, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul, 151-745, Republic of Korea
{mam3b,hpshin}@snu.ac.kr

Abstract

This study aims at retrieving tweets with an implicit topic, which cannot be identified by the current query-matching system employed by Twitter. Such tweets are relevant to a given query but do not explicitly contain the term. When these tweets are combined with a relevant tweet containing the overt keyword, the "serialized" tweets can be integrated into the same discourse context. To this end, features like reply relation, authorship, temporal proximity, continuation markers, and discourse markers were used to build models for detecting serialization. According to our experiments, each one of the suggested serializing methods achieves higher means of average precision rates than baselines such as the query matching model and the tf-idf weighting model, which indicates that considering an individual tweet within a discourse context is helpful in judging its relevance to a given topic.

1 Introduction

1.1 Limits of the Twitter Query-Matching Search

Twitter search was not a very crucial thing in the past (Stone, 2009a), at least for users in its early stages who read and wrote tweets only within their curated timelines real-time (Dorsey, 2007; Stone, 2009b; Stone, 2009c). Users' personal interests became one of the motivations to explore a large body of tweets only after commercial, political and academic demands, but it triggered the current extension of the Twitter search service. The domain of Twitter search was widened, for example, from tweets in the recent week to older ones (Burstein, 2013), and from accounts that have a specific term in their name or username to those that are relevant to that particular subject (Stone, 2007; Stone, 2008; Twitter, 2011; Kozak, Novermber 19, 2013). However, the standard Twitter search mechanism is based only on the presence of query terms.

Even though the Twitter Search API provides many operators, the current query matching search does not guarantee retrieving a complete list of all relevant tweets.[1][2] The 140-character limit sometimes forces a tweet not to contain a term, not because of its lack of relevance to the topic represented by the term, but due to one of the following:

Reduction the query term is written in an abbreviated form or in form of Internet slang,

Expansion the query term is in external text that can be expanded through other services such as TwitLonger (http://twitlonger.com) and twtkr (http://twtkr.olleh.com), while the part exceeding 140 characters is shown only as a link on twitter.com, or

Serialization the query term is contained as an antecedent in some previous tweet.

[1] "[T]he Search API is focused on relevance and not completeness." 2 October 2013. Using the Twitter Search API. https://dev.twitter.com/docs/using-search

[2] "[T]he Search API is not meant to be an exhaustive source of Tweets." 7 March 2013. GET search/tweets https://dev.twitter.com/docs/api/1.1/get/search/tweets

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 58–68, Dublin, Ireland, August 23-29 2014.

If these cases are frequent enough, the current query matching search in Twitter will get a low recall rate. Considering that tweets are usually used to obtain as various views on a topic as possible, in addition to accurate and reliable information about it, this setback would block attempts to collect diverse opinions in Twitter.

These three different cases require different approaches. First, *reduction*, one of the most significant characteristics of Twitter data in natural language processing, can be solved by building a dictionary of Internet slang terms or learning them. Second, in case of *expansion* tweets are always accompanied with short URLs (`http://tl.gd` for TwitLonger and `http://dw.am` for twtkr) and the full text is reachable through them. In these two cases, tweets correspond one-on-one with documents, whether reduced internally or expanded externally. This study will focus on the third case, *serialization*, where several tweets may be interpreted as a single document.

1.2 Serialization of Tweets: An Overlooked Aspect of Twitter

Though little reported before, serialization of tweets is frequently observed in Korean data.[3] Influential users like famous journalists, columnists and critics as well as ordinary users often publish multiple tweets over a short period of time instead of using other media such as blogs or web magazines. Types of tweets published in this way by Korean users include reports, reviews, and analysis on political or social affairs, news articles, books, films and dramas. The content users intend to express is longer than a tweet but shorter than a typical blog post. Examples from our dataset will be introduced in Section 3.

This study aims at retrieving tweets on a topic, which cannot be found by the current query-matching system. Such tweets are relevant to a given query but do not contain the necessary words. Under the hypothesis that a considerable number of these tweets not containing the query term are serialized with one containing it overtly, and that serialized segments are integrated into the same discourse context, we built a model that allows us, when given a tweet that includes a query or a mentioned topic, to find the other tweets serialized with it and count them as relevant to the topic. We primarily focused on Korean Twitter data, but we believe that the methods developed here are also applicable to other languages with similar phenomena.

2 Previous Studies

Our study is based on the observation that a tweet in a "serialization" does not necessarily correspond to a full document. In fact, it has already been reported (Hong and Davison, 2010; Weng et al., 2010; Mehrotra et al., 2013) that a single tweet is too short to be treated as an individual document, especially considering that word co-occurrence in a tweet is hardly found. Studies proved that performance of Latent Dirichlet Allocation (LDA) models for Twitter topic discovery can be improved by aggregating tweets into a document. In these studies, a "document " consists either of all tweets under the same authorship (Hong and Davison, 2010; Weng et al., 2010), all tweets published in a particular period, or all tweets sharing a hashtag (Mehrotra et al., 2013). These criteria are useful for finding topics, into which tweets can be classified, but our purpose requires a different degree of "documentness." Our study deals with a fixed topic and is interested in whether or not only tweets relevant to the topic can be pooled. All tweets merged into the same document as constructed in the previous studies are not necessarily coherent or related to the same topic because it is not usually expected that ordinary users devote their Twitter accounts to a single topic. In this study, we will develop more detailed criteria for the aggregation of tweets by combining authorship with time intervals and adopting features such as sentiment consistency and discourse markers.

A method of using discourse markers for microblog data was proposed by Mukherjee and Bhattacharyya (2012). They noted that a dependency parser, on which opinion analyses using discourse information (Somasundaran, 2010) are usually based, is inadequate for small microblog data, and instead used a "lightweight" discourse analysis, considering the existence of a discourse marker on each tweet. The list of discourse markers used in their study was based on the list of conjunctions representing discourse relations presented by Wolf et al. (2004). This method was successful for sentiment

[3] Some Korean users sarcastically call this a "saga" of tweets.

analysis on Twitter data assuming that the relevance of each tweet to a certain topic was already known. We will take a similar approach of using discourse markers, but with a different assumption and for a different purpose. In our study, we treat unknown topic relevance of tweets with missing query terms by aggregating them with a topic-marked tweet using discourse markers.

3 Features

3.1 Properties of Tweet Serialization

Multiple tweets are likely to be consistent with a topic if they form a discourse as in the following situations, with examples of tweets in Korean translated into English. In each tweet, topic words are in boldface.

Conversation This is the most typical case.

> **U1:** Wow the neighborhood theater is packed; will ***Snowpiercer*** hit ten million?
>
> **U2:** @U1 My parents and my boss are all gonna watch, and they watch only one film a year. This is the measure for ten million.

Comment after retweet Users retweet and comment.

> **U3** RT @U4: Today's quote. "It is stupid to concentrate on symbolic meaning in Wang Kar Wai's *Happy Together.* That would be like trying to find political messages and signs in ***Snowpiercer***." — Jung Sung-Il
>
> **U3** Master Jung's sarcasm.....☆

On-the-spot addition Because a published tweet cannot be edited, users can elaborate or correct it only by writing a new tweet or deleting the existing tweet.

> **U5** Is Curtis the epitome of **Director Bong**'s[4] sinserity
>
> **U5** Sincerity, shit

True (intentional) serialization Some users begin to write tweets with a text of more than 140 characters in mind. They arrive at the length limit and continue to write in a new tweet.

> **U6** (1) Watched ***Snowpiercer***. It was more interesting than I thought. It felt more like black comedy than SF. On another note, I was surprised by several oddities, making the film feel more like a Korean film with foreign actors in it rather than Director Bong's Hollywood debut.
>
> **U6** (2) In many ways the film was "nineties"... like watching *The City of Lost Children* all over again... and the trip from the tail-car to the first car, though I expected some kind of level-up for each car,
>
> **U6** (3) the world connected car to car was not an organic world (a sideways pyramid?) but worlds too separate car by car, and the front-car people were so lifeless that I was surprised. The scale of the "charge" after 17 years felt shrunken.

If this is a characteristic feature of Korean Twitter data, this may be due to reasons such as personal writing style, the writing system of the Korean language, and Korean Web platforms. First, it may be simply because these users prefer formal language and are reluctant to use short informal expressions even in Web writing. Second, it is possibly because CJK writing systems including Hangul, the Korean alphabet, have more information per character than the Roman alphabet (Neubig and Duh, 2013). Since a 140-character text in Hangul has generally more information than that in the Roman alphabet, a Korean (or Japanese) user can more readily tweet about content which an English (or other European) language user would consider too long to write about on Twitter. Third, for many Korean users Twitter is the most available medium for publishing their opinions online, as a number of standard blogs have been replaced

[4]Director of the film *Snowpiercer*

by microblogs. Some users divide a long public text into multiple length-limited tweets simply because they do not have a blog to write in.

While Internet slang and abbreviations are common in tweets, "Serializers" tend to use 1) fully-spelled forms (unlike "reducers"), 2) usually without hashtags and emoticons, 3) which are all visible on `twitter.com` itself (unlike "expanders"), so it is not guaranteed that all serialized tweets will contain the topic word, as in the examples above. This implies that some tweet segments in a single discourse may not be retrieved even if the discourse is relevant to a given query. Search results may include a partial document for which it is difficult the full version of which is difficult to find.

3.2 Extralinguistic Criteria

Two tweets are more likely to be a part of a larger document consisting of a series of tweets if

Reply-relation one of them is a reply to the other,

Temporal proximity they are published immediately one after the other, or

Continuation markers they share such markers as numbers, arrows >> and continuation marker '(continued).'

Figure 1 shows examples of each case.

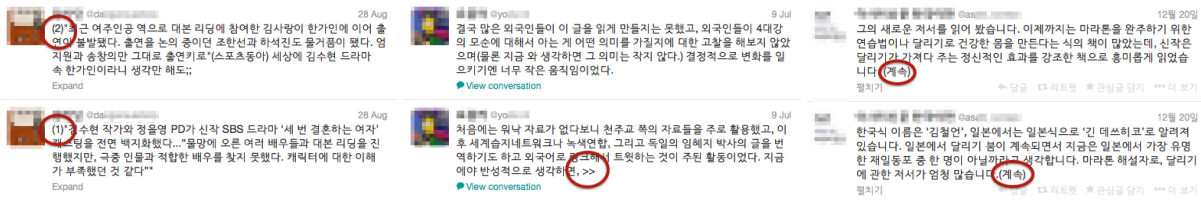

Figure 1: Serialized tweets with numbers, an arrow, or a continuation marker '(continued)'

3.3 Linguistic Clues

Semantic similarity to the query In order to determine the relatedness of two documents, the similarity between their term distributions is mainly considered. Based on this idea, one of our baseline methods will represent each tweet as a bag-of-words vector and retrieve a tweet containing no query term if its tf-idf weighted vector has a high cosine similarity with at least one vector from a tweet containing a query term.

Discourse markers Users may add a discourse marker when writing a new sentence in a new tweet. If a tweet begins with a marker that indicates continuation of a discourse, it is likely to be a part of a larger document. A sentiment analysis in Twitter by Mukherjee and Bhattacharyya (2012) adopted discourse relations from Wolf et al. (2004). In this paper, we use linguistic characteristics described by Kang (1999) in order to classify Korean texts, listing their English translations in Table 1. The *discourse marker* feature refers to whether or not any marker on the list occurs in the first N words (set $N = 5$) of the tweet.

4 Experiments

4.1 Data

We collected 173,271 tweets posted or retweeted by 105 Korean users, including film critics, film students, and amateur cinephiles from 27 July to 26 September 2013. Out of the 105 users, 17 users who had mentioned the film *Snowpiercer*[5] most often were singled out. In addition, the highest overall occurrence of the keyword was found to be between 1 to 15 August, probably due the film's release on 31

[5] `http://www.imdb.com/title/tt1706620/`

Demonstratives	*this, that, it, here, there*
Proverbs	*be so, do so*
Discourse	*well, now*
Conj-Reasoning	*because, so, therfore, thus, hence*
Conj-Conditional	*then, as long as, in the case, under*
Conj-Coordinate	*and, nor*
Conj-Adversative	*but, yet, however, still, by contrast*
Conj-Discourse	*meanwhile, anyway, by the way*

Table 1: List of selected Korean discourse markers used for classifying text types in Kang (1999), translated into English

July in South Korea. Then we kept all 8,543 tweets posted by those 17 users from the period between 1 to 15 August 2013, in order to construct a labeled data set. This set includes 189 tweets that explicitly contain the word *Snowpiercer*. Each tweet in the filtered set was labeled as *related* or *not related* to the movie by three annotators who were Twitter users already following most of the above 17 users and thus aware of the context of most tweets, and a tweet was considered relevant if two or more of the annotators agreed. Inter-annotator agreement was evaluated by using Fleiss's kappa statistic flei:71, which was $\kappa = 0.749$ ($p \approx 0$). Table 2 shows the annotation results.

	Related	Not related	Total
Explicit	173	15	188
Not explicit	207	8,148	8,355
Total	380	8,163	8,543

Table 2: The number of annotated tweets classified by explicitness and relatedness

Table 2 shows that 8163/8543 = 95.55% of the tweets in the dataset are not relevant to the movie *Snowpiercer*. Additional topics are induced from 7–9 manually collected seed words among the 200 most frequently occurring nouns in the dataset, in which each tweet text was POS-tagged by the Korean morphological and POS tagger Hannanum[6]. Induced topics and their seed words are listed in Table 3.

Topic	Seed words
Movie	Movie, Snowpiercer, director, The Terror Live, actor, stage, audience, film, theater
Literature	Story, book, writing, author, novel, character, work
Gender/relationship	Men, women, female, marriage, male, wife, lover
Politics	Politics, state, Park Geun-hye, government, president, party, Ahn Cheol-soo

Table 3: Four topics from manually collected seed words

As described in 3.1, it should be noted again that hashtags are not always useful for finding information in Korean tweets, particularly in this dataset. Among the seed words above, only *Snowpiercer* was ever used as a hashtag, and happened only three times (twice in English and once in Korean). Only nine types of hashtags occurred more than twice in the full dataset (they are presented in Table 4 with their respective frequencies). This predicts that hashtag-based tweet aggregation would not be very useful to find tweets relevant to *Snowpiercer* or one of the four induced topics.

Table 5 shows the number of tweets containing seed words for each topic, where a tweet is allowed to belong to more than one topic. Since only 1853/8543 = 21.69% of the tweets explicitly contain a topic or seed word, it is not plausible that each of the remaining 80% tweets belongs to one of the four topics. Many of the tweets may be related to a topic which was of a too small portion to be induced, or to no topic at all. So, instead of classifying all of the tweets into the given topics, the experiment seeks to retrieve any tweet that is relevant to a certain topic, which allows each tweet to belong to more than one topic at once. In every experiment we regarded tweets that contain a topic or seed word as relevant to the topic, and restricted the test set to those tweets which did not contain them.

[6] `http://sourceforge.net/projects/hannanum/`

#make_people_cry_with_a_story_of_two_words	13
#lgtwins	10
#quote	7
#changing_zero0_to_fatty_makes_things_totally_depressing	6
#EBSbookcafe	4
#today_i_feel	4
#blow_the_whistle_on_chun_doo-hwan	3
#chosundotcom	3
#the_name_of_your_bias_followed_by_the_name_of_the_food_you_just_ate_feels_nice	3

Table 4: Korean hashtags occurring more than twice in the dataset, translated into English

Movie	Literature	Gender	Politics	Total
716	452	379	306	1853

Table 5: Number of tweets including at least one of the seed words for each induced topic

4.2 Measures

For all models, the authors judged the relevance of each of the retrieved tweets for induced topics until ten relevant tweets were retrieved. In the *Snowpiercer* case, precision scores were calculated for all recall scores. We built a ranking retrieval system for each model and evaluated its performance by average precision. For models including a randomizing process, we used the mean of average precisions over 1,000 replicated samples. Precision was computed at every percentile of recall levels for *Snowpiercer* case and after each retrieved relevant tweet (up to top 10) for induced topics. In sum, the performance of a model m was defined in two ways as

$$\text{meanAP@percent}(m) := \frac{1}{1000}\sum_{i=1}^{1000}\text{AP@percent}(m_i)$$

and

$$\text{meanAP@10}(m) := \frac{1}{1000}\sum_{i=1}^{1000}\text{AP@10}(m_i)$$

, where m has 1,000 replicates $m_1, \cdots m_{1000}$ whose measures are

$$\text{AP@percent}(m_i) := \frac{1}{100}\sum_{j=1}^{100}\text{prec@}j\%(m_i)$$

and

$$\text{AP@10}(m_i) := \frac{1}{10}\sum_{k=1}^{10}\text{prec@}k\%(m_i).$$

When m is a tf-idf model, which has a unique ranking without replication, average precision was used.

4.3 Baselines

Query matching method The most obvious baseline method for this study is the current Twitter search system that treats topic words and seed words as queries and finds documents, or tweets, that are relevant to the topic. Since only tweets not containing the query terms remained in the test set, there are no tweets matching them. As the set of retrieved tweets is empty, relevance rank is randomly assigned to each tweet of the test set.

Tf-idf weighting method One may predict that a tweet is likely to be relevant to a topic if it shows a similar word distribution to some explicitly relevant tweets. Under this assumption, we represented each tweet as a tf-idf weighted vector (Salton and Buckley, 1988) after removing all punctuation marks and user-mention markers (`@username`). Stopwords were not removed and tf-idf values were length-normalized. Relevance of each tweet in the test set was defined as the maximum of its cosine-similarities with all tweets containing a query term.

4.4 Tweet Serialization

Examples of Tweet Serialization in Section 3 indicate clues between related tweets other than distributional similarity. When 1) a tweet is a reply to another one, 2) two tweets are written one after another by the same user, 3) one tweet following another includes some discourse marker, or 4) two tweets share a marker, such as numbers, they can be considered to be serialized into a single document rather than being two separate ones. Tweets serialized together are treated as a single document, and if this document contains a a tweet with a query term, then all tweets lacking it but belonging to the same the same document are retrieved. All retrieved tweets are first ranked in random order, followed by the others also in random order.

We suggest four criteria for Tweet Serialization:

Reply Two tweets are serialized if one is a reply to the other.

Continuation markers Two tweets are serialized if they are written successively by the same user and share a marker, such as a number or a phrase "(cont.)"

Discourse markers Two tweets are serialized if they are written successively by the same user, the latter contains one of the discourse markers listed in Table 1 in its first 5 words, and neither of them is a reply to another user.

Time Two tweets are serialized if they are written successively by the same user within a given interval and neither of them is a reply to another user. The upper boundary for intervals is set in one of the following ways:

Constant 30 or 60 seconds

User-specific Users may show different densities in their tweets, depending on their tweeting environment. Distribution of time intervals between successive tweets over users is presented in Table 6. The smallest 5% and 15% quantiles were selected, corresponding to 30 and 60 seconds respectively.

Quantile	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15	U16	U17
0%	3	19	2	1	1	5	10	3	2	3	9	3	2	3	3	2	16
5%	20	42	18	16	13	30	21	18	13	8	43	23	18	15	13	12	110
10%	33	45	25	35	20	43	38	38	28	13	71	35	28	35	21	21	130
15%	47	52	33	57	30	56	67	57	40	23	89	51	37	61	27	40	161
20%	62	67	41	79	41	73	92	74	53	31	111	65	50	84	33	58	197
25%	81	86	55	100	55	92	145	95	69	43	138	84	68	105	38	77	275
50%	237	298	164	322	151	242	1060	297	167	159	297	317	178	258	90	266	725

Table 6: Time intervals (in seconds) by cumulative percentile between consecutive pairs of tweets for each user

For all criteria, Tweet Serialization is transitive, that is, if t_i and t_j are serialized and t_j and t_k are serialized, then t_i and t_k are serialized. Table 7 shows the distribution of serialization sizes (number of serialized tweets) over criteria. Time value of 60 seconds serializes most tweets, as many as (8543-6464)/8543=24.33%, while continuation markers serialize only (8543-8511)/8543 = .37%. Assuming all serializations are correct, the relevance of retrieved documents is judged.

4.5 Results

The average precision values of all models are summarized in Table 8 (means calculated over recall levels) and Figure 2 (means calculated over 1,000 replications). In both Tables 8 and 9, differences between the tf-idf weighting model and each of the Serialization methods were statistically significant according to t-test. Figure 2 compares the results of the serialization methods, among which *continuation marker* model has the highest precision over 0.8 at the 1% recall level, and *Time with 15% quantile* has the average precision score showing the slowest decrease. Even though for all serialization methods average precision values converge to zero as recall levels increase, each of the method gets higher precision rates than baselines until some part of relevant tweets are retrieved.

Size	Repl.	Disc.	Cohe.	T:30s	T:60s	T:5%	T:15%
1	8137	8169	8511	7314	6464	7845	6849
2	88	166	6	465	664	298	610
3	34	14	2	76	149	31	109
4	9	0	0	6	40	1	19
5	3	0	1	5	12	1	8
6	5	0	0	1	6	0	2
7	3	0	0	0	3	0	0
8	1	0	0	2	2	0	1
9	2	0	1	0	0	0	0
10	0	0	0	0	0	0	0
11	0	0	0	0	1	0	1

Table 7: Distribution of serialization size (number of serialized tweets) under each criterion

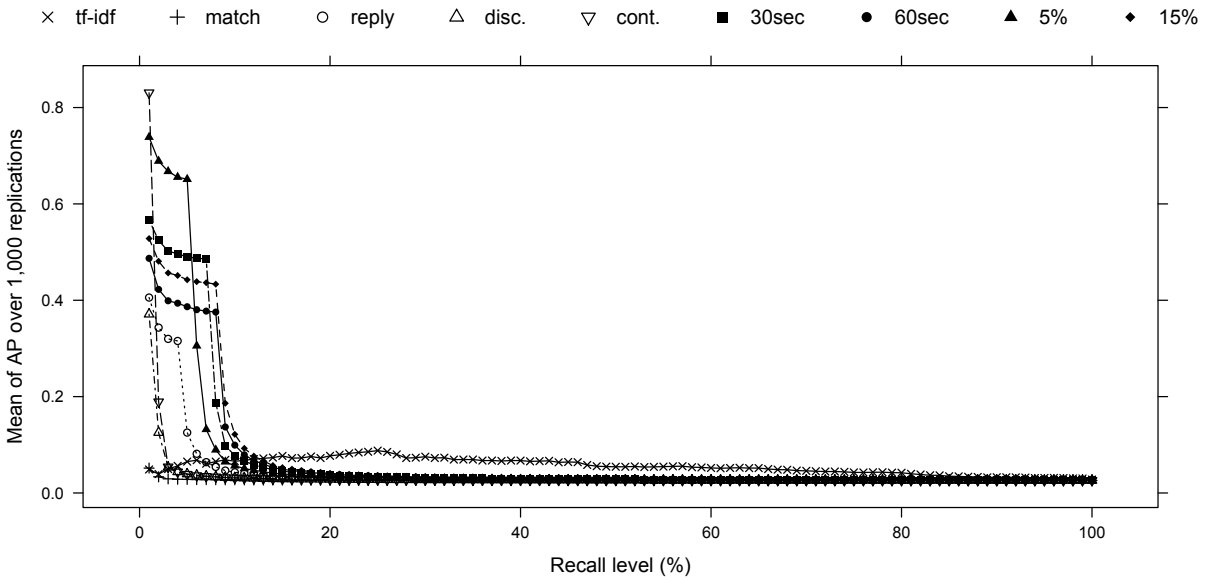

Figure 2: Means of average precision rates of all methods for the topic *Snowpiercer*

Recall	Baselines:		Repl.	Disc.	Cont.	Time difference threshold			
level $\leq$	Match	tf-idf	rela.	mark.	mark.	30sec	60sec	5%	15%
5%	.0342	.0518	.3019	.1266	.2313	.5158	.4178	**.6804**	.4720
10%	.0309	.0588	.1798	.0801	.1324	.3916	.3459	**.4050**	.3976
25%	.0284	.0695	.0920	.0494	.0702	.1824	.1665	.1847	**.1894**
50%	.0273	.0685	.0602	.0382	.0486	.1062	.0986	.1070	**.1103**
100%	.0268	.0556	.0434	.0322	.0375	.0666	.0628	.0669	**.0687**

Table 8: Means of average precision rates (at recall level up to 5%, 50%, and 100%) on various serialization criteria for the topic *Snowpiercer* (Results in boldface represent the best results among the methods.)

Serialization methods also perform better than the tf-idf baseline for induced topics, as shown in Figure 3 and Table 9. In particular, *Reply* and *Discourse markers*, which were far from the best for *Snowpiercer*, serve well for other topics such as *Movie* in general, *Politics*, and *Gender/Relationships*.

The precision of *Reply* for the topic *Movie* is exceptionally high, partly because the data were initially collected from users who were interested in films. *Reply relation* is dependent on the choice of the data, in that it is determined by interaction between users, not by a single user's tweets. If data are collected from users friendly with each other, *Reply* will serialize many tweets. On the contrary, if data contains some users while leaving out their friends, replies to these friends are not serialized by *Reply* criteria.

Discourse markers give a precision of higher than 50% for the topic *Politics*, which is likely to be discussed in more formal expressions using various conjunctions.

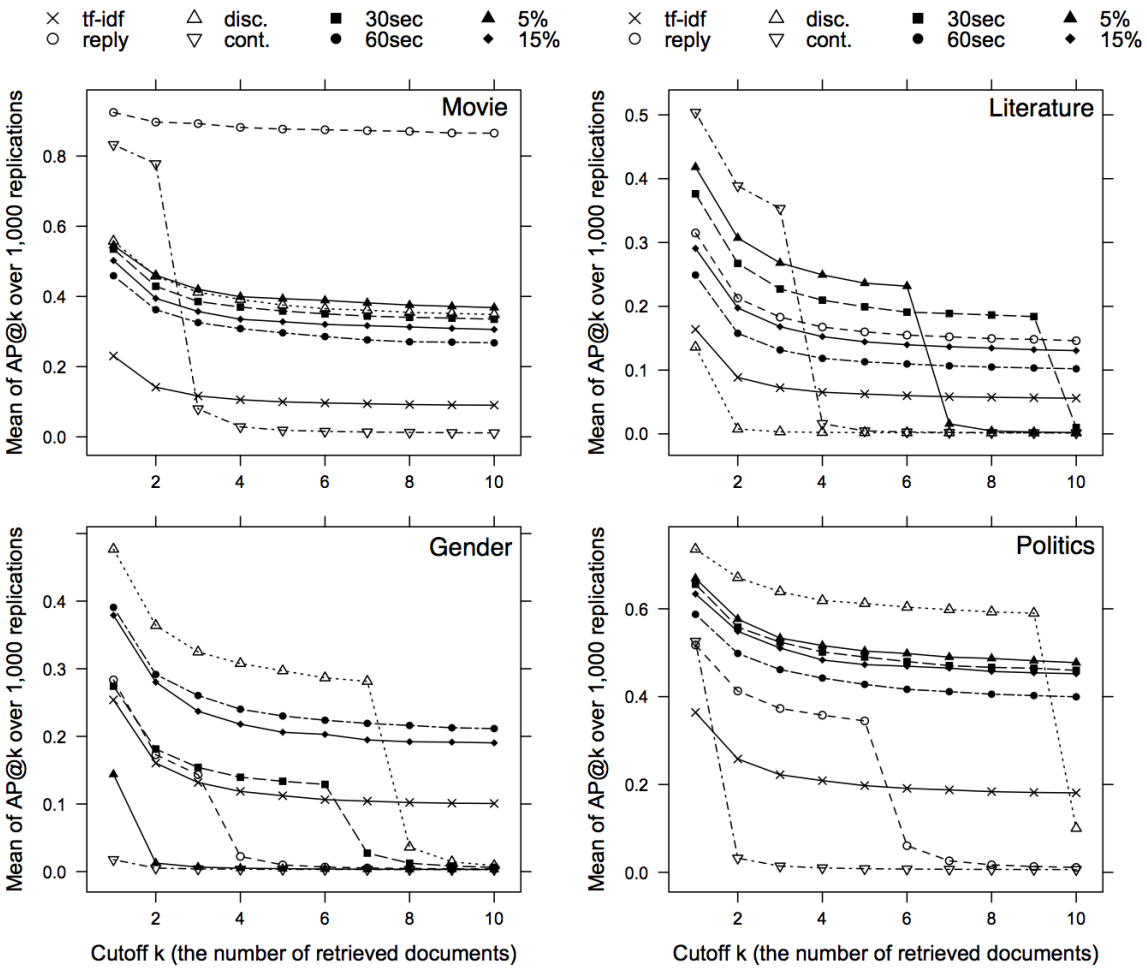

Figure 3: Means of average precision rates of all methods for the induced topics

	Baselines:		Repl.	Disc.	Cont.	Time difference threshold			
	Match	tf-idf	rela.	mark.	mark.	30sec	60sec	5%	15%
Mov.	.0134	.1139	**.8855**	.3925	.1791	.3787	.3123	.4161	.3435
Lit.	.0026	.0759	.1804	.0171	.1293	**.2005**	.1287	.1719	.1601
Gen.	.0048	.1287	.0653	.2424	.0050	.1092	**.2476**	.0187	.2297
Pol.	.0090	.2176	.2135	**.5762**	.0625	.5072	.4453	.5234	.4948

Table 9: Means of Average Precision rates at cutoff $k = 10$ of baselines and different serialization criteria for induced topics (Results in boldface represent the most accurate results of the topic among the methods.)

In the topics *Literature* and *Gender/Relationships*, average precision scores are at most 25%, which possibly results from the fact that the seed words for these topics consist of general terms only, while those of the other two topics include proper nouns such as movie titles or politicians' names. This is less a problem of the topic itself but rather one of data selection, which focused on users tweeting about films, and so the set of seed words will vary according to differences in data collection.

5 Conclusion

In this paper, we found that tweets with an implicit topic can be found more effectively by considering whether or not they are serialized with some tweet containing the overt keyword. Our experiments show that Tweet Serialization can be detected using various criteria such as reply relations between users, presence of discourse or continuation markers, and temporal proximity under the same authorship. Our

original purpose was to find as various opinions on a given topic as possible, but we expect the methods used here will be helpful for other tasks, including topic discovery and sentiment analysis, by setting more exact document boundaries in microblog data. The method we proposed is for Korean Twitter data, where tweet serialization is observed frequently, particularly among influential users, but it is also applicable to other languages with similar phenomena.

In future work, we will investigate methods for the evaluation of the results of Tweet Serialization and combine tf-idf methods with Tweet Serialization criteria. Furthermore, we aim at verifying the applicability of the results of this study with regard to more various users and more topics.

References

Paul Burstein. February 7, 2013. Older Tweets in search results. *The Official Twitter Blog.* https://blog.twitter.com/2013/now-showing-older-tweets-in-search-results.

Jack Dorsey. September 25, 2007. Tracking Twitter. *The Official Twitter Blog.* https://blog.twitter.com/2007/tracking-twitter.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin.* 76(5): 378–382.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. *SOMA 2010: The Proceedings of the First Workshop on Social Media Analytics.* 80–88.

Beom-mo Kang. 1999. *Hankukeui theksuthu cangluwa ene thukseng* [Text genres and linguistic characteristics in Korean]. Korea University Press, Seoul, Korea.

Esteban Kozak. November 19, 2013. New ways to search on Twitter. *The Official Twitter Blog.* https://blog.twitter.com/2013/new-ways-to-search-on-twitter.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment analysis in Twitter with lightweight discourse analysis. *COLING 2012: The 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers.* 1847–1864.

Rishabh Mehrotra, Scoot Sanner, Wray Buntine and Lexing Xie. 2013. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. *SIGIR '13; The 36th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 889–892.

Graham Neubig and Kevin Duh. 2013. How much is said in a Tweet? A multilingual, information-theoretic perspective. *AAAI Spring Symposium: Analyzing Microtext, Volume SS-13-01 of AAAI Technical Report*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. 24(5): 513–523.

Swapna Somasundaran. 2010. *Discourse-level Relations for Opinion Analysis*. Ph.D Thesis, University of Pittsburgh.

Biz Stone. August 22, 2007. Searching Twitter. *The Official Twitter Blog.* https://blog.twitter.com/2007/searching-twitter.

Biz Stone. December 23, 2008. Finding Nemo — Or, name search is back! *The Official Twitter Blog.* https://blog.twitter.com/2008/finding-nemo%E2%80%94or-name-search-back.

Biz Stone. February 18, 2009. Testing a more integrated search experience. *The Official Twitter Blog.* https://blog.twitter.com/2009/testing-more-integrated-search-experience.

Biz Stone. April 03, 2009. The discovery engine is coming. *The Official Twitter Blog.* https://blog.twitter.com/2009/discovery-engine-coming.

Biz Stone. April 30, 2009. Twitter search for everyone! *The Official Twitter Blog.* https://blog.twitter.com/2009/twitter-search-everyone.

Twitter. April 4, 2011. Discover new accounts and search like a pro. *The Official Twitter Blog.* https://blog.twitter.com/2011/discover-new-accounts-and-search-pro.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: Finding topic-sensitive influential twitterers. *WSDM '10: Proceedings of the Third ACM International Conference on Web Search and Data Mining*. 261–270.

Florian Wolf, Edward Gibson and Timothy Desmet. 2004. Discourse coherence and pronoun resolution. *Language and Cognitive Processes*, 19(6): 665–675.