Abstract

Retrieval of Twitter messages without an explicit query term by means of serialization and discourse segmentation

Park, Suzi
Department of Linguistics
The Graduate School
Seoul National University

This thesis describes a phenomenon where multiple tweets constitute a single discourse segment, and builds two rule-based models to detect whether two consecutive tweets under the same authorship convey a single message. Given the length limit of 140 characters, a tweet should be interpreted as an element of a larger unit rather than an individual document. Considering such a larger unit as a discourse segment and a tweet as an utterance, this study makes the following assumptions based on Centering Theory:

(a) A tweet has at most one topic.

(b) In non-initial tweets of a discourse segment, a topic word is realized as an anaphora, in particular a zero form in Korean.

(c) Coherence between two tweets written by the same author is considered only if there is no tweet between them.

(d) In two consecutive tweets, a topic is preferred to be continued.

To predict tweet serialization and discourse segmentation, two criteria were used: temporal proximity and discourse markers. Temporal proximity shows whether the time interval between two tweets is less than a threshold level, which can be a constant or user-specific value. Discourse markers are classified into continuation markers and shift markers. Continuation markers include web-specific ones such as '>>', '(continued)', and numbers, and linguistic ones such as conjunctions and referring expressions. Shift markers include web-specific ones such as 'RT' and URLS, and linguistic ones such as interjections and temporal adverbs. These factors are treated differently in two different models. The Strict Serialization (SS) model regards two tweets as serialized only if their interval is extremely short or they have a continuation marker. On the contrary, the Serialization Plus Discourse Segmentation (SPDS) model, following the assumption (d) that continuation is preferred to shifting, considers two tweets as serialized if their interval is not too long, and terminates a discourse segment only if the current tweet has a shift marker.

To verify whether the proposed models are useful, an information retrieval task

is implemented. It is predicted by the assumption (b) and observed in the data that topic words were implicit in some tweets in discourse segments consisting of multiple tweets. The current search system cannot retrieve such tweets and thus fails to satisfy users' information need to find diverse opinions in Twitter. When finding discourse segments compiled by the proposed models, the system can retrieve tweets that belong to the same discourse segment as some explicitly relevant one, without retrieving too many irrelevant tweets. Consequently, the proposed models achieve higher means of precision rates than those of the Query Matching model and TF-IDF Weighting model. Furthermore, since the SPDS model outperforms the SS model, the principle of unmarkedness of topic continuation seems to be also valid for social media. Lastly, this thesis also discovers that linguistic markers such as interjections, which have been typically treated as stopwords in information retrieval, are useful for discourse segment detection.

Keywords: Centering theory, Discourse marker, Information retrieval, Social media, Twitter

Student Number: 2012-20031

Contents

List of Tables

List of Figures

1 Introduction

The major purpose of this thesis is to detect discourse segments among unstructured short text messages in social media such as Twitter which have been treated mainly as independent documents, and to improve the performance of natural language processing (henceforth NLP) tasks by finding such discourse segments. It stargs start with the delineation of its subject matter.

1.1 SUBJECT

In the middle of 2010s, *social* has become a productive adjective, coining terms such as *social commerce*, *social (network) game*, and *social television*, in a somewhat different sense to from its meaning in *social democracy* or *social security*, rather reminiscent of its Latin etymology *socius* (meaning "one who accompanies another, a companion, comrade").[1] This usage was derived from *social media*, which, according to Merriam-Webster Dictionary, refers to "forms of electronic communication (as Web sites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (as videos)."[2]

This definition attracts linguists' attention, because "information, ideas, and personal messages" are typically expressed in human language and even "other content

[1] P.G.W. Glare. (1968) *Oxford Latin Dictionary*. Oxford University Press. p. 1779.
[2] `www.merriam-webster.com/dictionary/socialmedia`

as videos" includes descriptions and tags. Messages in social media provide plentiful examples of human language use for theoretical linguistics as well as NLP.

Though there are various various text-based social media services such as Facebook (`facebook.com`) and Line (`line.me`), this study will focus on Twitter (`twitter.com`), not only because of its popularity, but also because of the following characteristics:

Openness. While users in most social networking services such as Facebook allow only the selected people to read their messages, Twitter's users do not need to be authorized to read messages, or tweets, of others. This spreadability of tweets explain Twitter's impact on various fields such as politics.

Length limit. The most distinctive feature of Twitter is the number 140, the maximum number of characters that a single tweet can consist of. This limit takes the pressure of writing a long and serious article off users.

On the basis of these characteristics of Twitter data, goals of this thesis will be described in the next section.

1.2 PURPOSES

1.2.1 Detection of discourse segments in Twitter data

The first goal of this thesis is to detect discourse segments among tweets. Even though each tweet is mainly presented as a part of the "timeline" of users who

subscribe, or "follow"' the author of that tweet, in NLP it is typically treated as an independent document because it has a permalink in the form of `twitter.com/<user>/status/<tweet>`. This treatment may overlook an important property of social media, where messages are interchanged to share information in users' community and then expected to be connected each other. Hence, a tweet needs to be considered as an utterance which can be an element of a discourse segment. NLP researchers have proved that it is useful to aggregate tweets sharing some aspects into a larger unit. While their studies have typically used structured non-textual features such as timestamps, geotags and hashtags, this thesis will focus on finding linguistic and nonlinguistic markers from unstructured text in tweets, with the existence of discourse segments presupposed. In short, its purpose is not to classify individual tweets, but to reconstruct discourse segments produced by Twitter users.

1.2.2 Retrieval of tweets without an explicit query term

The second goal of this thesis is to improve Twitter search. The current Twitter search system retrieves tweets that contain a given query term explicitly. However, in case when multiple tweets constitute a discourse segment on a topic, not every tweet of the segment should contain the topic word in the full form. Consequently, there can exist tweets that cannot be found by the current search engine even though they are relevant to a given query. This would block attempts to collect diverse opinions in Twitter. Discourse-segment-level search can be a solution to this problem, retrieving

all tweets in a discourse segment that matches the given query term.

1.3 STRUCTURE

The remaining part of this thesis is structured as follows. Chapter 2 explores previous studies on Twitter data in the field of NLP, focusing on those of them which are related to the subject or methods of the thesis. Chapter 3 investigates major concepts in centering theory for discourse segments and their compatibility with characteristics of Twitter, an open social media with length limit. Chapter 4 describes a variety of real examples of discourse segments from Korean Twitter data and proposes two models to detect discourse segments consisting of multiple tweets. Chapter 5 applies discourse segmentation to an information retrieval task, finding tweets that are only implicitly relevant to a given query. Finally, Chapter 6 concludes this thesis.

2 Previous Work

2.1 GENERAL NLP STUDIES

Social media has become an important subject in NLP as a source of a training set for various tasks (Feng et al., 2013) as well as a source of a main test set.

2.1.1 On social media data

Studies on social media include diverse tasks from the processing of noisy texts to topic/event/item detection, to named entity recognition, and to sentiment analysis.Most of these tasks answer one of the following three questions:

- How is it shown that something is being talked about?
- What is talked about?
- How is it talked about?

The fourth possible question, namely Which messages talk about it?, has attracted little attention, partly because retrieval of short documents such as tweets seems a trivial task, for which a query matching search is sufficient. In Chapters 4 and 5, counterexamples observed in Korean Twitter data will be introduced.

2.1.2 Using discourse knowledge

Traditional computational discourse theories, such as centering theory and rhetorical structure theory, have provided crucial insight into tasks including anaphora resolutions. Recently, a series of Somasundaran's works using transition states between adjacent discourses (2007; 2008a; 2008b; 2009; 2009a; 2009b; 2010) contributed to sentiment analysis. Discourse knowledge is useful for extracting information that is not explicit lexically, but usually relies on a detailed annotation and a dependency parser. These components are not readily available for Twitter data, where a vast number of documents are being added every moment and most of them are too noisy to be parsed automatically. The following section introduces various studies that have attempted to overcome these difficulties and will be introduced in the following section.

2.2 TASK-SPECIFIC STUDIES

2.2.1 Finding a proper unit for unstructured short texts

In general, pooling messages into a document or thread is a valid method for dealing with diverse types of unstructured short texts such as emails, blogs, short text messages (SMS), and chats (Huang et al., 2011; Joty, 2013). It has been reported that a tweet is too short to be a complete document and inappropriate for models using co-occurrence of words; it has been proved that an LDA (Blei et al., 2003) topic

model for Twitter can be improved through aggregation of multiple tweets written by the same user, published at the same moment, or sharing a hashtag (Weng et al., 2010; Mehrotra et al., 2013). Topic detection on disasters also has benefited from a tweet-pooling scheme based on location information and word distribution (Lee, 2012; Kumar et al., 2014). Even though these criteria for tweet aggregation were useful for compiling documents of sufficient size to discover a topic, it is not guaranteed that all tweets pooled into the same document belong to the same topic, because users are unlikely to devote their account to a single topic and can talk about arbitrary topics even at the peak of popularity of a certain topic. Retrieval of tweets for a given topic requires more coherent "documentness." Thus, multiple tweets need to be considered as serialized only if authors and readers recognize them as forming a discourse segment, usually by aid of discourse markers which will be discussed in the next subsection.

2.2.2 Discourse markers in Twitter data

2.2.2.1 Twitter-specific part-of-speech tagging

Twitter-specific discourse markers were first described by Gimpel et al. (2011). They stipulated that environments for discourse markers include 'continuation of a message across multiple tweets.' According to them, users indicate continuation across their tweets by '(...)' and '>>.' As their task was a development of POS tags for

Twitter, they did not describe or utilize these markers further. This thesis raises a question whether Twitter search can be improved using linguistic discourse markers as well as Twitter-specific ones.

2.2.2.2 Linguistic markers

Linguistic discourse markers have been employed for Twitter data by Mukherjee and Bhattacharyya (2012). The authors applied discourse markers from Wolf et al. (2004) instead of a dependency parser, which is commonly used in discourse-based opinion analysis (Somasundaran, 2010) and is not appropriate for unstructured tweets. Their discourse features increased the accuracy of a polarity classifier with prerequisite knowledge of the topic of each tweet, and captured discourse information within a tweet rather than between two tweets. In contrast, this thesis aims at discovering each tweet's relevance to a given topic with the aid of discourse relations between two tweets, under the assumption that we are aware of the topic of only tweets containing the topic word.

2.2.3 Classification of tweets without an overt topic word

2.2.3.1 Item detection

Classification of tweets having none of the given topic words was attempted in an item detection task using idf (inverse document frequency) features in SVM learning by Cremonesi et al. (2013). While their study dealt with official and fan accounts of

a fixed list of movies and TV programs, the tweets for this thesis will be obtained from personal personal account of users who can tweet about arbitrary topics.

2.2.4 Summary

The section above explores how social media such as Twitter has been dealt with in NLP studies and which computational tasks benefit from using discourse knowledge. The following issues have been found previously:

- multiple tweets can correspond to a single document and be pooled;
- discourse analysis can be facilitated by finding discourse markers;
- tweets can be relevant to a topic without containing the topic word explicitly.

They will be discussed in detail in the remained part of this thesis. Despite these overlaps, this thesis includes two novel points:

- discourse markers will be used for detecting a discourse segment composed of multiple tweets;
- detecting such segments will be shown to improve retrieval of tweets that do not contain a query term explicitly but are relevant to the query term.

The following chapter will examine the theoretical background for capturing discourse segments among tweets.

3 Centering Theory for Discourse Segment Detection

This thesis aims at building models for detecting tweets that constitute a discourse segment based on assumptions, rules, and concepts of Centering Theory (Grosz et al., 1995). This chapter will introduce the theory and investigate an applicability of its postulates to Twitter data.

3.1 OVERVIEW

In Centering Theory, an utterance has *centers*, and is linked to other utterances in the same discourse segment by its centers. A set of forward-looking centers is assigned to each utterance; each non-initial utterance has a single backward-looking center. Among the theory's assumptions regarding the relationship between discourse coherence, inference load, and choice of a referring expression, the following three in particular provide a proper account for topic behavior in consecutive tweets:

(1) Claims of centering Theory (Grosz et al., 1995, 210–211)

   a. Each utterance has exactly one backward-looking center.

   b. The forward-looking centers are partially ordered according to a number of factors. Ranking of elements in the current utterance determines the backward-looking center of its consecutive utterance.

c. The backward-looking center for the current utterance is chosen from the set of forward-looking centers of the immediate previous utterance. In this sense the backward-looking center is strictly local. The backward-looking center of the current utterance cannot be taken from other prior sets of forward-looking centers.

Grosz et al. (1995) also defined three types (2) of transition relations between consecutive utterances and stated a constraint (3) on the preference among these types.

(2) Center transition types (Grosz et al., 1995, 210)

a. CONTINUATION: The backward-looking center of the current utterance is the same to that of its following utterance, and this entity is the most highly ranked element of the set of forward-looking centers of that following utterance.

b. RETAINING: The backward-looking center of the current utterance is the same to that of its following utterance, but this entity is not the most highly ranked element of the set of forward-looking centers of that following utterance.

c. SHIFT: The backward-looking center of the current utterance is not the same to that of its following utterance.

(3) Rule 2 on Center Movement (Grosz et al., 1995, 215):

Sequences of continuation are preferred over *sequences* of retaining; and sequences of retaining are to be preferred over sequences of shifting.

Following Brennan et al. (1987), some researchers subcategorize SHIFT into SMOOTH-SHIFT and ROUGHT-SHIFT, but this distinction will be not used in this thesis. The next subsection covers four issues of tweets based on these claims and rules.

3.2 MAJOR CONCEPTS USED IN THIS THESIS

3.2.1 Uniqueness of the backward-looking center

The uniqueness of the backward-looking center (1a) is reflected by the most salient characteristic of Twitter data: a tweet is too short to contain multiple topics. Similar to the sentiment consistency of a tweet assumed by Feng et al. (2013), the uniqueness of a topic in a tweet is can be safely assumed.

In this thesis, it is not intended to identify the topic of each tweet. Instead, the focus will fall on tweets that are retrieved or or considered to be relevant when a search query is given in an information retrieval task. Such a query term is expected to refer to a specific entity such as an event or an item e.g. *World Cup*, rather than be a general term e.g. *sports*. Thus it is plausible that if a tweet is relevant to a query term, then the query term refers to the topic of the tweet. According to the hypothesis mentioned in the previous paragraph, a tweet and its topic will be considered to correspond to an utterance and its backward-looking center respectively.

3.2.2 Highest rank of zero pronouns as centers

Forward-looking centers of an utterance are ranked in partial order (1b). Among them, the most highly ranked one is said to be the preferred center (Brennan et al., 1987). Kameyama gave a linguistic hierarchy of nominal expressions to determinine the preferred center:

(4) Nominal Expression Type Hierarchy (EXP ORDER) (Kameyama, 1998, 92):
Given a hierarchy [ZERO PRONOMINAL > PRONOUN > DEFINITE NP > INDEFINITE NP], an entity realized by a higher-ranked expression type is normally more salient in the *input* attentional state.

(5) EXP CENTER (Kameyama, 1998, 93):
An expression of the highest-ranked type in EXP ORDER normally realizes the Center in the output attentional state.

Like Japanese (Walker et al., 1994, 1998), Korean has zero anaphora, so an expression realized as a zero pronoun is most likely to be the preferred center. However, it is extremely difficult for computers to distinguish whether an invisible form is a zero pronoun or absent, especially in languages with relatively free word order. However, Kameyama's proposal shows at least a possibility for the reference of an expression which is not overt in an utterance to be realized as the center of that utterance. This possibility will be the basis of the task in Chapter 5, which attempts to find tweets that are implicitly relevant to a given query.

3.2.3 Locality of coherence

Strict locality (1c) in the choice of the backward-looking center requires adjacency between two utterances. Even though Roberts (1998), building upon the theories of Heim (1982) and Kamp (1981), introduced a new definition of locality under the observation that "utterance adjacency is not always necessary in anaphoric relations," adjacency condition has been used until recent work (Karamanis et al., 2009; Wei, 2014), not limited in centering theory (Redeker, 1990; Passonneau and Litman, 1997). In this thesis the term *locality* will also be used to refer to adjacency.

3.2.4 Preference of center continuation

Finally, the preference relation CONITUNUATION>>SHIFT (3) should be noted. Even though continuation and shift are indicated by various cue phrases (Thanh et al., 2004; Forbes and Miltsakaki, 2002; Hirschberg and Litman, 1993; Passonneau and Litman, 1997; Grosz and Sidner, 1986; Fraser, 1999; Joty, 2013; Bestgen and Vonk, 1995, 2000; Bracewell et al., 2012; Grosz et al., 1995; Redeker, 1990), such phrases are not only optional, but also more likely to be omitted in unstructured writing. By assuming the preference of transition states, a default value can be assigned to cases without an explicit marker. If there is no sufficient for any transition state, it is natural to consider a center to be continued.

3.3 SUMMARY

This chapter has investigated major concepts in centering theory and their compatibility with characteristics of Twitter, an open social media with a length limit. Consequently, when assuming that a tweet and its topic correspond to an utterance and its center, it is plausible to expect that consecutive tweets share a topic and constitute a discourse segment. The following chapters will describe various aspects of discourse segments observed in Korean Twitter and then explore features for a model to detect discourse segments based on given query terms.

4 Tweet Serialization and Discourse Segmentation

4.1 TWEET SERIALIZATION

4.1.1 Phenomenon

When an author writes on a topic in multiple tweets,[3] readers generally recognize that the tweets are "serialized" (Park and Shin, 2014).

Planned serialization: If an author intends to make a long document out of 140-character blocks,[4] he or she usually indicates their cohesion with some markers such as a number (see Example 8), an arrow, '(continued)', a hashtag, a bracketed heading, or a reply to self.

Unplanned serialization: In the case that an author posts tweets one after another off the top of his or her head, readers can perceive continuation mainly because of short intervals between the tweets.

Between these two extremes, various types of tweet serialization are observed in Korean data, as presented in (6–10):

(6) Conversation

[3]This tweeting style has provoked debates since at least 2008. See the following blog post: Glenn Murray. November 10, 2008. Twitter Etiquette: Should One Message Span Multiple Tweets? *Divine Write (Blog)*. `www.divinewrite.com.au/social-media/twitter-etiquette-should-one-message-span-multiple-tweets`

[4]Some Korean users sarcastically call such documents as a "saga" of tweets (대하트윗). Twitter search results (`twitter.com/search-home`) for "대하트윗" reveal various reactions to dozens of tweets.

a. @u1: 헐 동네극장 아침부터 개터져나감; 설국 천만 가나요; (Aug 03 2013 01:37:09)

@u1: Wow the neighborhood theater is packed; will *Snowpiercer* hit ten million?

b. @u2: @u1 일 년에 영화 한 편 보는 울 엄마 아빠나 회사 국장님 부장님이 보려는 거 보면요. 천 만의 척도. (Aug 03 2013 01:44:49)

@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.

(7) Comment after retweet[5]

a. RT @u: 오늘의 명언. "왕가위의 [해피투게더]에서 상징적 의미에 집중하는 것은 바보같은 짓입니다. 그건 마치 [설국열차]에게서 정치적 메세지와 기호들을 찾으려 애쓰는 것과 같은 것이죠." — 정성일 (Aug 13 2013 14:16:58)

RT @u: 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

b. 정성생님 돌려까긔..... (Aug 13 2013 14:17:11)

Master Jung's sarcasm.....

(8) Intentional serialization

a. (1)[설국 열차] 봤다. 예상보다 재미있었다. SF보다 블랙 코미디의 일종으로 느껴졌다. 그것과 별도로, 보는 내내 몇가지 이질적인 점 때문에 되게 당황스러웠는데, 이게 봉 감독의 할리우드 진출작이라기보다는 외국 배우 나오는 '한국영화'라는 느낌. (Aug 02 2013 16:53:47)

(1) Watched *Snowpiercer*. It was more interesting than I thought. It felt more

[5]From now on, all tweets in the same example are written by the same user.

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.

b. (2) 여러가지 의미에서 '90년대적' 인 영화라는 느낌도 들었는데... 뭐랄까 [잃어버린 아이들의 도시] 같은 영화를 다시 보는 느낌이라고 해야 하나.... 그리고 꼬리칸에서 맨 앞칸까지 이르는 여정이, 칸칸별로 일종의 레벨업이 이뤄질 걸로 예상했는데, (Aug 02 2013 16:58:17)

(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,

c. (3) 칸칸이 이어지는 그 세계가 거대한 유기적 세계(수평으로 이어지는 피라미드?)가 아니라 너무 칸별로 개별적인 세계이고, 앞칸 사람들은 아무 존재감이 없어서 그것도 놀랐다. 17년을 기다려야 했던 그 '진격'의 스케일이 확 줄어드는 느낌. (Aug 02 2013 16:59:57)

(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.

(9) On-the-spot correction

a. 커티스는 봉감독님 진전성의 화신인가 (Aug 05 2013 13:53:47)

Is Curtis the epitome of Director Bong's sinserity

b. 진정성 시발년아 (Aug 05 2013 13:53:57)

Sincerity, shit

(10) Free-style addition

a. 설국열차보고나옴 (Aug 05 2013 12:28:47)

I watched *Snowpiercer*

b. 썩 괜찮은 영화였닿 별 세개 (Aug 05 2013 12:36:50)

Wasn't bad lol I give it three stars

Discourse types found among serialized tweets are classified with respect to the number of participants:

Dialogue: Two or more users exchange tweets in reply to each other (Example 6);

Quotation: One user quotes other's tweet and adds one's own comment in the next tweet (Example 7);

Monologue: One user continues to talk in successive tweets (Examples 8–10).

Among these three types, monologues and quotations will be the main subject of this study for the following reasons:

i) There is more need for developing detection methods for them in Twitter. Dialogues can be identified immediately and automatically because each tweet has a parameter `in_reply_to_status_id`, whose value is what the tweet is in reply to.[6] On the other hand, monologues and quotations do not have such a direct link so it is challenging to recognize such types in the timeline the way users do.

[6] `dev.twitter.com/docs/api/1.1/post/statuses/update`

ii) They are recognized more clearly when tweets in a dialogue tweets are separated.[7] Within a user's tweet stream, a tweet in reply to another may intervene between two tweets in the same monologue as the user can receive a mention and give an answer while writing a series of tweets, without loss of coherence of the series. If the intervening one is removed, continuity of the tweets in a monologue or quotation can be preserved.

4.1.2 Constraints

Given a tweet, it is ineffective and inefficient to search the whole Twitter data for other tweets which are serialized with it. The scope of candidates can be narrowed down by proper constraints. Such constraints will be convenient to verify if they are based on parameters of each tweet.

4.1.2.1 Authorship

Each tweet has a parameter `user` and its subparameter `user_id`, whose value is the Twitter user identifier of its author.[8] For monologues and quotations, where the value of the parameter `in_reply_to_status_id` is `NULL`, tweets are considered as serialized only if they were published consecutively by the same user.

[7] In Twitter, the default profile page of each user (`twitter.com/<screen_name>`) all user's tweets except for replies. Tweets in reply to other users appear on an additional page (`twitter.com/<screen_name>/with_replies`).

[8] `dev.twitter.com/docs/api/1.1/get/users/lookup`

4.1.2.2 Temporal proximity

As discussed in subsection 3.2.3, judging coherence of two consecutive tweets will be meaningful only if they meet locality condition. For locality to be compatible with tweet serialization, it should be defined not only as adjacency (Redeker, 1990; Grosz et al., 1995; Passonneau and Litman, 1997; Roberts, 1998; Karamanis et al., 2009; Wei, 2014), but also as temporal proximity, because one's yesterday's last tweet and today's first tweet are unlikely to be serialized. Since each tweet has a timestamp as a value of the `created_at` parameter,[9] time interval between two tweets can be calculated automatically. After calculation it is necessary to determine how to scale the value and how to specify its direction.

Scalability: real-valued vs. boolean-valued

The degree to which two consecutive tweets are related to each other will be predicted by using their temporal proximity, which will be measured as a function of their time difference. That function can be chosen to be a continuous monotonic one (such as linear, logarithmic, and so on) or a discrete finite-valued one (such as boolean). In this thesis, both pairwise relatedness and temporal proximity will be treated as binary-valued so that

i) annotators will label each tweet with "relevant (to a given query)" or "irrelevant," and

[9] `dev.twitter.com/docs/entities`

ii) two tweets will be considered to satisfy a temporal proximity condition if they are consecutive, belong to the same author and their interval is smaller than some threshold value.

Setting a proper threshold level of time difference will be discussed in section 4.2.

Direction: two-way vs. one-way

In this subsection temporal proximity has only been defined as a property of a pair of tweets, calculated from its time interval. Instead, it can be considered as a property of a single tweet and its direction. Since tweets are strictly ordered by timestamp, each tweet has two adjacent tweets—one precedes it and the other follows it, so temporal proximity is divided into two kinds—backward and forward. This distinction will be useful if similar values of time difference should be interpreted differently according to direction. In the next section, two models will be proposed with two different functions defining temporal proximity:

i) a function of time difference of each pair of tweets (See 4.2.1.1 Strict serialization), and

ii) a function of directional time difference of each tweet (See 4.2.1.2 Serialization Plus Discourse Segmentation).

4.2 DISCOURSE SEGMENTATION OF SERIALIZED TWEETS

4.2.1 Two strategies for discourse segments detection

4.2.1.1 Strict serialization

The first model proposed in this paper is called Strict serialization (SS). In this single-stage model, which was earlier introduced in Park and Shin (2014), tweets are essentially individual documents. Two consecutive tweets are considered to be serialized only if there is firm evidence of continuation. Such evidence includes the following:

Immediacy: Continuation between two tweets is indicated by a very short interval. A threshold level for "shortness" is set as the following two values:

i) a constant value of 30 or 60 seconds and

ii) a user-specific threshold at the 5% or 15% quantile reflecting individual tweeting styles.

Continuation markers: When a topic is maintained over several utterances—tweets, users can type a discourse marker as naturally as they speak or write it. Moreover, active authors use some web-specific expressions to inform their readers that the current tweet is continued from the previous one or will be continued in the next one. Types and examples of continuation markers will be specified in subsection 4.3.2.

Reply: A tweet is considered to be serialized with its replies.

There seem to be few tweets that satisfy any of these conditions, but once a tweet satisfies it, then the tweet is highly likely to be serialized. Therefore, the SS model is expected to have high precision but low recall.

4.2.1.2 Serialization Plus Discourse Segmentation

The Serialization Plus Discourse Segmentation (SPDS) model, which is a double-stage one, considers two consecutive tweets to be serialized *as far as they are not too distant.* It can find more serialized tweets and hence achieve a higher recall rate than that of the SS model. However, to increase recall without a fatal loss of precision, the SPDS model should further find discourse segments that are coherent enough. Coherence cannot be assured by temporal proximity alone, because in Twitter there is little limit on topics and users move from one topic to another more freely than in other speech or writing situations. In order to establish a more coherent discourse segment, in its second stage the SPDS model cuts the serialized tweets into discourse segments *if there is possible evidence of topic shift.* The most reliable evidence for topic shift is the presence of a new topic word, but it is exceedingly difficult to check it without a list of possible topics. Instead, shift markers that signal a change in topic will be investigated.

The requirement for temporal proximity in the SS model is relaxed in the SPDS model because of the preference of center continuation discussed in the subsection

3.2.4, as well as classical principles cited in the study of Bestgen (1998) such as 'given-new contract theory' (Clark and Haviland, 1977), 'nextness principle' (Ochs, 1979), and 'principle of continuity' (Segal et al., 1991, 32). The threshold value for proximity will be defined as the median, or 50% quantile (cf. 5–15% in the SS model) of all time differences between immediate pairs of tweets from the same user.

Direction of temporal proximity is also a matter to consider. In the SPDS model, temporal proximity will be only forward, measured on tweets which follow the one that contains an initial topic word, according to the claim (12) that centers of the current utterance determine that of the next utterance. A series of tweets in (11) is an example where backward proximity with a previous tweet fails. Even though the interval between the tweet (11e) and its preceding tweet (11e) is as short as 41 seconds, they do not constitute the same discourse segment because (11a–d) are tweets about the film *Oldmen Never Die*[10] while (11e–f) are about the film *Snowpiercer*[11]. This thesis will claim that this shift is indicated with a discourse marker such as 'RT' in (11e), which will be discussed in the next subsection.

(11) a. 죽지 않아 봤습니다. 유산상속 받으려고 수구꼴통 할배 밑에서 4년 동안 죽어라 일하다 지친 손자가 여자를 데려와 할배를 복상사시키려 한다는 이야기. 코미디처럼 들리겠지만 전통적인 필름 느와르의 공식을 그대로 따릅니다. 컴컴해요. (Aug 02 2013 07:36:10)

[10] www.imdb.com/title/tt3086950
[11] www.imdb.com/title/tt1706620

Did *Oldmen Never Die*. A grandson working his ass off for 4 years under his reactionary grandfather for his fortune, and finally trying to get him a girl to induce coition death. Sounds like a comedy but it follows the traditional film noir formula. Dark.

b. 장르 공식이 노골적으로 드러나 있고 소재가 되는 인물들의 사고방식이 뻔하기 때문에 결말이 훤히 보이는데, 후반부엔 멀쩡한 사람들이 알아서 그 함정에 빠지는 모습이 많이 갑갑하더군요. 전 좀 짧았으면 좋았을 것 같습니다. (Aug 02 2013 07:38:19)

The genre formula is pretty blatant and the characters are stereotypical, so the ending is predictable, but towards the end it is tiring to watch decent people throw themselves into that trap. I would have liked it better if it were shorter.

c. 이 영화를 만든 사람들은 잠시 나오다마는 아버지 캐릭터에 가장 감정이입이 되어 있는 듯. 그 기준에서 수구꼴통 아버지 세대와 의욕없는 아들 세대를 보고 있는 거죠. (Aug 02 2013 07:42:30)

The makers of the film seem to be most sympathetic towards the father character that appears for a short while. They are regarding the reactionary grandfather-generation and the listless youths from that point of view.

d. 중간에 끼인 여자로 나오는 한은비는 목소리가 좋더군요. (Aug 02 2013

07:43:27)

Han Eunbi, who was cast as the woman in the middle, has a nice voice.

e. RT @u: 설국열차 리뷰 읽었어요. 솔직히 이건 좀 아닌 거 같아요. 관객은 봉준호에게 자기복제를 강요하는 게 아니라 퀄리티의 균일성을 요구하는 거라고 봅니다. (Aug 02 2013 07:44:08)

@u8: RT @u: I read the review of *Snowpiercer*. I can't really agree. I think the audience wants an evenness of quality from Director Bong, not self-replication.

f. '퀄리티의 균일성'. 윽. (Aug 02 2013 07:45:31)

"Evenness of quality." Ugh.

4.2.2 Discourse markers in tweets

Since social media is motivated by sharing information, authors are likely to use some discourse markers to help their reader understand their tweets, even while writing in a short, informal and unstructured style. According to Dascalu (2014), discourse markers indicating cohesion or coherence, even though not a necessary condition for comprehension (Sanders and Noordmand, 2000), improve understanding of readers (Degand and Sanders, 2002). In this subsection, discourse markers observed in Twitter data will be classified by source (into Web-specific & linguistic markers) and by function (into continuation & shift markers).

4.2.2.1 Web-specific markers

Continuation markers

Authors who serialize multiple tweets intentionally often mark their tweets with "(cont.)" (계속 kyeysok in Korean), ">>", or numbers, exemplified in Figures 4.1, 4.2, and 4.3 respectively. Besides ">>" reported by Gimpel et al. (2011), "(cont.)" and numbers have been observed in Korean data.

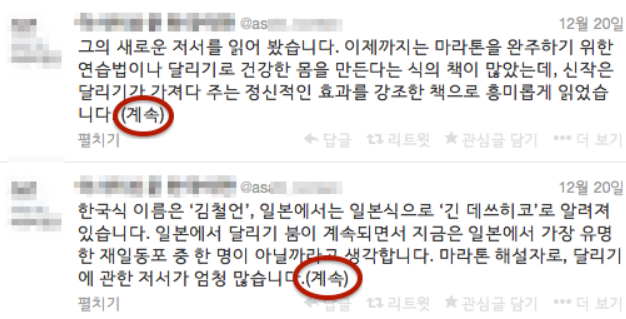


Figure 4.1: Continuation marker: "(continued)"

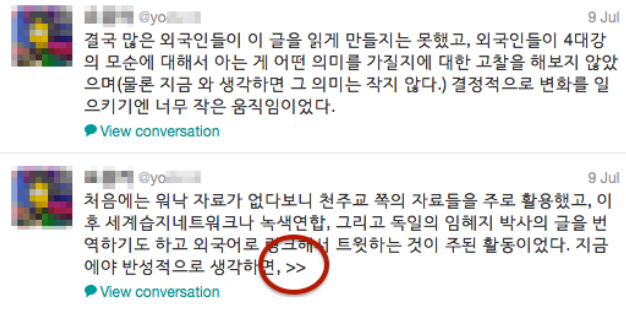


Figure 4.2: Continuation marker: arrows

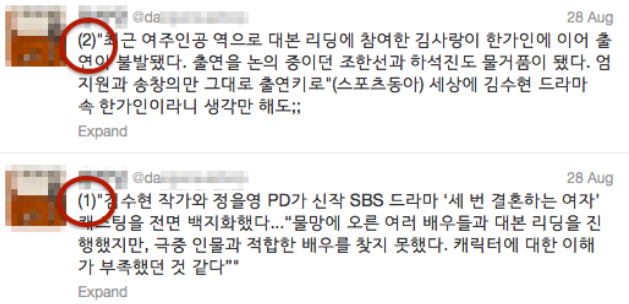


Figure 4.3: Continuation marker: numbers

Shift markers

RETWEETS A retweet, henceforth RT, can be counted as an initiation of a new discourse segment. Since tweeting is far from structured writing, it is not to be expected that users will quote others' material to support their argument in the middle of writing. Once an RT occurs in a user's stream, this RT poses a new topic rather than continues its predecessor's, and its successor is likely to be a comment on the retweeted one. In Example (12), a topic-initiating RT (a,c,e) and the user's own comment (b,d,f) to it alternate.

(12) a. RT @u1: [설국열차]의 과묵한 전사 그레이 역을 맡은 루크 파스콸리노는 액션 연기 경험이 거의 없었기 때문에 봉준호 감독의 액션 강도에 부응하고자 일찍이 제작팀에 합류, 세트장 한 켠에서 매일 같이 액션 연습을 했다고. (Aug 09 2013 05:14:17)

RT @u: Luke Pasqualino, who appears as *Snowpiercer*'s silent warrior Gray, had next to no action-film experience, and so joined the producing

team early in order to meet Director Bong's action standards, practicing in a corner of the set every day.

b. 남두수조권 최후의 계승자 (Aug 09 2013 05:14:34)

The last inheritor of Nanto Seiken.

c. RT @u2: 늘 말씀드리지만 깔 때는 최소한의 검색이라도 해보고 까는 겁니다. (Aug 09 2013 05:14:56)

RT @u2: As I always say, before you bash something, do a routine search first.

d. 최소한의 검색도 귀찮아서 아무것도 안 깝니다. (Aug 09 2013 05:15:12)

Can't be bothered to do even a routine search, so I don't bash anything.

e. RT @u3: 유빅컵 조기 소진에 따른 감사제 시작! 컵 인증하시고 이벤트 페이지 들어오셔서 인증 링크랑 알라딘 계정명을 남겨주시면 알사탕 200 개 드립니다. 인증 모아서 월페이퍼라도(웃음) 차기 증정품 아이디어도 받아요. (Aug 09 2013 05:20:02)

RT @u3: Thank-you festival for the early sellout of the Ubig cup! Register your cup, connect to the event page and leave your ID, we will give you 200 candies. Buy wallpaper by collecting registrations (lol) Accepting ideas for the next giveaway

f. 알사탕이 이백개 냠냠 잘먹겠습니다 (Aug 09 2013 05:20:17)

Thanks for the 200 candies nom nom

URLS Like RTs, URLs included in a tweet can also function as an initiator by introducing an external Web page.

4.2.2.2 Linguistic markers

Continuation markers

CONJUNCTIONS Linguistic discourse markers can be useful for capturing topic continuation or shift across tweets. First, it has been widely noted that conjunctions such as *and* connect two utterances in discourse (Grosz and Sidner, 1986; Fraser, 1999). For tweets written in Korean, the list of conjunctions from Kang (1999), shown in Table 4.1, will be used.

Reasoning	그러므로, 그러니, 그러니까, 하니까, 그니까, 고로, 따라서, 그래서, 그리하여, 그런 즉, 한 즉, 그렇기 때문에 (because, so, therefore, thus, hence)
Conditional	그렇다면, 그러면, 그렇거든, 그래야만, 그럴진대 (then, as long as, in case, under the condition)
Coordinate	그리고, 또, 더구나, 그뿐 아니라, 그뿐만 아니라 (and, nor, not only … but also, as well as)
Adversative	그렇지만, 하지만, 허나, 그러나, 그러하나, 그러하되, 그럴지라도, 그럴망정, 그럴지언정, 그래도, 그럼에도, 그렇다손치더라도 (but, yet, however, still, by contrast)
Discourse	그런데, 근데, 헌데, 한데, 한편 (meanwhile, anyway, by the way)

Table 4.1: Korean conjunctions (Kang, 1999)

REFERENTIAL EXPRESSION Second, referential expressions such as pronouns serve as indicators of local coherence (Grosz et al., 1995). For tweets written in

Korean, the list from the same book of Kang (1999), shown in Table 4.2, will be consulted.

Demonstratives	이것, 그것, 저것, 요것, 조것, 고것, 이, 그, 저, 요, 조, 거기, 여기, 조기 (this, that, it, here, there)
Pro-verbs	그렇다, 그렇게 하다 (be so, do so)

Table 4.2: Korean referential expressions (Kang, 1999)

Shift markers

INTERJECTIONS In examples (13–14), authors' attention is shifted from *Snowpiercer* to other subjects. Readers of (13) may already detect this change in (b) before seeing the new topic word *podcast* in (c) because of *ah*. This usage of interjections as a turn initiator or topic change indicator has been studied widely in previous literature (Bestgen, 1998; Montes, 1999; Norrick, 2009).

(13) a. 낼 일찍 일어나야 한다… 설국열차 생각 이만하고 자야지 O<-< (Aug 04 2013 15:32:22)

I need to wake up early tomorrow morning... I should stop thinking about *Snowpiercer* and sleep O<-<

b. 아 맞다 (Aug 04 2013 15:33:39)

Ah wait

c. 아까 신형철 평론가 팟캐스트 듣는데 무슨 거기서 추천한 이달의 책 읽고 독후감 써서 보내면 문학동네 세계문학 전집 1부터 100까지 한 사람한테 몰빵해서 준다던데 (Aug 04 2013 15:34:43)

I was listening to critic Shin Hyungchul's podcast just now and he said if you read their book of the month and write them a review they'll pick one person to give the entire Munhakdongne classics series from Book 1 to 100

For tweets written in Korean, the expressions listed in Table 4.3 were collected from crawled data (section 5.2) and will be considered as interjections.

Expression	Tweets
아 a	126
오 o	47
헐 hel	34
아니 ani	29
음 um	26
엌 ukh	23
뭐 mwe	19
후 hwu	16
헉 hek	13
앗 as	11
야 ya	10
A ㅏ Aa	2
후우 hwuwu	2
으 u	1
으음 uum	1

Table 4.3: Interjections used in this thesis; based on 8,543 tweets

TEMPORAL ADVERBS Another linguistic marker to determine discourse segmentation is a class of temporal adverbs such as *now* in (14b). Temporal expressions have been subcategorized under cue phrases for discourse segments (Grosz and Sidner, 1986; Passonneau and Litman, 1997) and have been proved to signal a thematic shift (Bestgen and Vonk, 2000).

(14) a. 아 기둥뒤에 공간 있는 영화도 또 봐야되고 설국열차도 봐야되고 마지막 사중주도 봐야되는데..... (Aug 03 2013 13:35:55)
Ah I need to watch that film with the space behind the pillar and *Snowpiercer* and *The Last Quartet*.....

b. 일단 오늘 ebs 영화나 보자.. (Aug 03 2013 13:36:08)
Now I'll watch today's EBS movie..

c. 흑흑 지난주에 ebs에서 그림자 군단 해줬는데 놓쳤다 흑흑 (Aug 03 2013 13:40:30)
sob sob I missed last week's EBS showing of *Army of Shadows* sob sob

For tweets written in Korean, the expressions listed in Table 4.4 were collected from the crawled data (section 5.2) and will be regarded as temporal adverbs.

Expression			Tweets
지금	cikum	*in the present*	36
요즘	yocum	*nowadays*	27
이제	icey	*now*	21
일단	iltan	*for the moment*	20
아까	akka	*a while ago*	12
앞으로	aphulo	*in the future*	5

Table 4.4: Temporal adverbs used in this thesis; based on 8,543 tweets

4.2.3 Algorithm of the SDPS model

Summarizing the discussion in the preceding subsections, the SPDS model is based on the following four hypotheses for the information retrieval task in the following chapter:

1. Given a set of query terms, the first tweet that contains any of them is an initial utterance of a discourse segment relevant to the query.

2. If the initial utterance has a succeeding one, the center of the former occurs most naturally as the preferred center of the latter.

3. If the successive utterance is continued, either a null anaphor or a pronoun is preferred to a full form as a topic word.

4. If a shift marker (one of RTs, URLs, interjections, or temporal adverbs) occurs in an utterance, an existing discourse segment finishes at its preceding utterance and a new discourse segment begins with this utterance.

The third hypothesis implies the existence of *implicitly relevant tweets*, which contain no query term but are relevant to the query. Algorithm 1 derives rules for the SPDS model.

Data:

- Q: set of (expanded) query terms
- $T = \{t_1, t_2, \cdots, t_m\}$: a user's tweet stream except replies
- $\theta = \text{med}_{k=1,2,\cdots,m-1}\text{time.diff}(t_k, t_{k+1})$: threshold of time difference between immediate tweets
- $T_Q = \{t_{i_1}, t_{i_2}, \cdots, t_{i_r}\}$: a user's non-reply tweets containing any query term q in Q;
  set of tweets explicitly relevant to Q
- AH: any interjection
- NOW: any temporal adverb

Result:

S : set of discourse segments relevant to Q

$\bigcup_{s \in S} s - T_Q$: set of tweets implicitly relevant to Q

```
begin
    S = ∅
    for j = 1, 2, ··· , r do
        s = {t_{i_j}}
        for k = 1, 2, ··· , i_{r+1} − i_r − 1 do
            if time.diff(t_{i_j+k−1}, t_{i_j+k}) ≤ θ then
                if (t_{i_j+k} is RT)
                or (t_{i_j+k} contains URL)
                or (t_{i_j+k} begins with AH)
                or (t_{i_j+k} begins with NOW) then
                    S = S ∪ {s}
                    go to the next j
                else
                    s = s ∪ {t_{i_j+k}}
                end
            else
                S = S ∪ {s}
                go to next j;
            end
        end
        S = S ∪ {s}
    end
end
```

Algorithm 1: Tweet serialization and discourse segmentation with respect to a given query set

5 Retrieval of Implicitly Relevant Tweets

5.1 OVERVIEW: CURRENT SEARCH SYSTEM IN TWITTER

A need for Twitter search appears to be obvious, but it is not intrinsic. When the founders of Twitter defined their service as sharing what people were doing in real time with friends,[12] it was expected that users normally read what occurred in their "curated timeline"[13,14] and rarely explored beyond it. They would be motivated to search by curiosity, and when searching they would be less interested in what was being said by users that they followed, but about what has been said by arbitrary users on topics that they were probing into. The latter has become extensive enough to browse as the number of Twitter users has increased.

Nowadays, Twitter search is steadily growing, and is extensively used not only by experts in limited fields, but also for ordinary users.[15,16] The search range has been

[12]Jack Dorsey. September 25, 2007. Tracking Twitter. *The Official Twitter Blog.* `blog.twitter.com/2007/tracking-twitter`

[13]Biz Stone. April 3, 2009. The Discovery Engine Is Coming. *The Official Twitter Blog.* `blog.twitter.com/2009/discovery-engine-coming`

[14]Biz Stone. April 30, 2009. Twitter Search for Everyone! *The Official Twitter Blog.* `blog.twitter.com/2009/twitter-search-everyone`.

[15]Biz Stone. November 14, 2006. Six More Twitter Updates! *The Official Twitter Blog.* `blog.twitter.com/2006/six-more-twitter-updates`

[16]Biz Stone. July 15, 2008. Finding A Perfect Match. *The Official Twitter Blog.* `blog.twitter.com/2008/finding-perfect-match`.

extended in diverse ways, including user accounts,[17,18,19,20] photos and videos,[21] old tweets as well as recent ones,[22] related query suggestions[23] (Mishne et al., 2013), and yet each entity can be found only if it explicitly contains a query term somewhere.

Considering the length limit of Twitter, one can expect that the full form of a topic word may be invisible in some tweets, in one of the following ways:

Reduction Topic words can be shortened or abbreviated.

Expansion Topic words can appear on an external web page linked by a URL, which can back up a short tweet (Chu et al., 2012).

Serialization Topic words can be mentioned in a preceding tweet, because some users prefer to complete their messages even though they need multiple tweets to do that (Park and Shin, 2014).

Such tweets cannot be found by a query-matching method. Since Twitter is mainly considered to be a source of various opinions rather than accurate and reliable knowledge, missing these kinds of tweets can fail to meet the needs of search.

[17] Biz Stone. August 22, 2007. Searching Twitter. *The Official Twitter Blog.* `blog.twitter.com/2007/searching-twitter`

[18] Biz Stone. December 23, 2008. Finding Nemo—Or, Name Search is Back! *The Official Twitter Blog.* `blog.twitter.com/2008/finding-nemo%E2%80%94or-name-search-back`

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

[20] Esteban Kozak. November 19, 2013. New ways to search on Twitter. *The Official Twitter Blog.* `blog.twitter.com/2013/new-ways-to-search-on-twitter`

[21] Tian Wang. November 15, 2012. Search for a new perspective. *The Official Twitter Blog.* `blog.twitter.com/2012/search-for-a-new-perspective`

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

[23] Frost Li. July 6, 2012. Simpler Search. *The Official Twitter Blog.* `blog.twitter.com/2011/discover-new-accounts-and-search-pro`

Even though it is possible for humans to reconstruct the topic word of such tweets under whichever cases, teaching it to machines requires different solutions. In case of reduction, a shortened topic word can be recovered using a dictionary or a set of rules built automatically or manually. Relevance feedback will also work because a substantial number of users search by both a full form and a reduced one. Regarding expansion, linked pages are accessible and thus searchable. In these two cases, a single tweet corresponds to a single document. This study will focus on the third case, where more than two tweets are interpreted as a single document.

5.2 DATA

5.2.1 Description

173,271 tweets were crawled from 105 Korean users from July 27 to September 26 2013 via Twitter REST API[24]. Among these, 8,543 tweets written by 17 users who mentioned the word (*Selkwukyeylcha* 설국열차 (*snowpiercer*) most frequently were selected as the test set.; the period was restricted from August 1 to 15 which was when the word was most mentioned. Three annotators labeled each tweet with relevance to the movie. The annotators were Twitter users already following most of the 17 accounts in the test set, so they were aware of the context of most tweets.[25]

[24] `dev.twitter.com/docs/api/1.1`

[25] For example, there was a tweet saying *I need to watch that film soon, if just for those two people's tweets about it* in the data set, whose author did not specify the antecedents. Since the annotators knew what *that film* was and who *two people* were, they could judge the tweet's relevance to *Snowpiercer*.

Inter-annotator agreement was evaluated by using Fleiss's kappa statistic (Fleiss, 1971), which equaled $\kappa = 0.749$ ($p \approx 0$). Each tweet was considered relevant if two or more of the annotators agreed. Table 5.1 shows the annotation results. Since there are a greater number of implicitly relevant tweets than explicitly relevant tweets, the necessity of searching for tweets that do not have a query term is confirmed. Furthermore, by observing that *Snowpiercer* covers only 380 tweets out of 8,543 tweets even on peak of its popularity, we induce it is less appropriate for retrieval to pool all tweets under the same authorship or time span into a single document, which would give an accuracy of 380/8,543=4.45% for the query term *snowpiercer*.

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

Table 5.1: Annotation

From the test data, two types of tweets are excluded: replies (to concentrate on monologues and quotations), and those already containing the word *snowpiercer* (to single out what a query-matching system cannot find). Thus, the final test data has 130 tweets out of 6,209 non-dialogue tweets that are implicitly relevant to *Snowpiercer*. The goal of this experiment is to find these 130 tweets as accurately and exhaustively as possible.

5.3 BASELINES

In all baselines and proposed models, all tweets, which originally constitute a single timeline, are listed chronologically and split by authors, as shown in 5.1. A black star (★) denotes explicitly relevant tweets, which are not counted as elements of the test set.

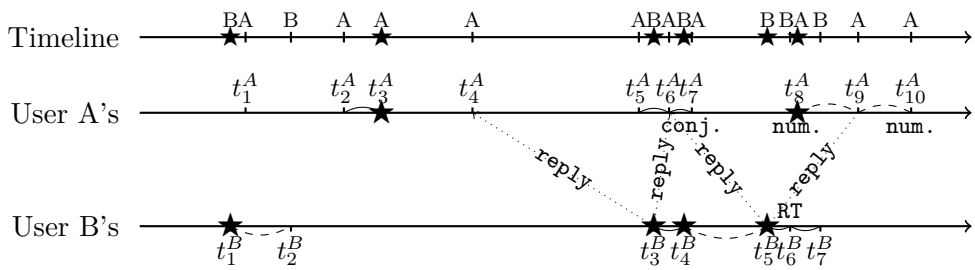


Figure 5.1: Original timeline and its author-decomposition

5.3.1 Query matching model

The Query Matching model, like the current Twitter search, retrieves a tweet only if it contains the query term *snowpiercer*. Since only tweets that do not contain *snowpiercer* remains in the test set, there are no tweets matching it. The set of relevant tweets being empty, relevance rank is randomly assigned to each of the 6,209 non-relevant tweets.

In Figures 5.2–5.7, each square (■ or □) denotes a tweet that does not contain the word *Snowpiecer*. A black square (■) refers to a retrieved tweet.

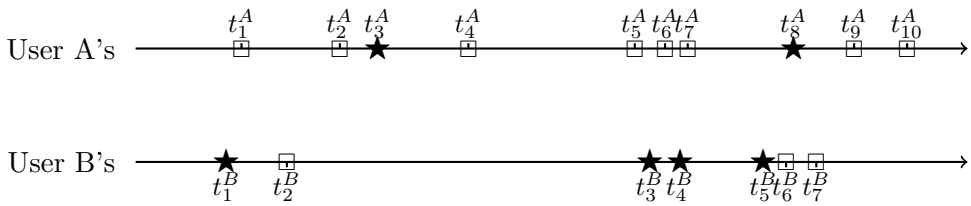


Figure 5.2: Query Matching Model

5.3.2 Tf-idf Weighting model

One may predict that a tweet is likely to be relevant to *Snowpiercer* if it shows a similar word distribution to some explicitly relevant tweets. To calculate the semantic similarity between two documents, each document is represented as a vector. Each component of this document-vector corresponds to a term, describes importance of the term in the document, and is weighted by the tf-idf (term frequency × inverse document frequency) score (Spärck Jones, 1972; Salton and Buckley, 1988). As formulated in (15), the term frequency function of a given term and a given document measures how frequently the term occurs in the document and the inverse document frequency function of a given term shows how many documents contain the term. The tf-idf score is the highest if the term occurs frequently only in a single document. If a term, like *the* in English, is used in the majority of documents, it is not little informative, which is reflected in its idf score approaching zero.

(15) Term Frequency and Inverse Document Frequency functions

$$\text{tf}(\text{term}, \text{document}) := \text{the number of tokens of the term in the document}$$

$$\text{idf}(\text{term}) := \log\left(\frac{\text{the number of all documents in the corpus}}{\text{the number of documents with the term}}\right)$$

The tf-idf score is measured for each term from the whole corpus so that all document-vectors have the same dimension.[26] Before building the list of all terms, all punctuation markers and user-mention markers (`@username`) were removed, but stopwords were retained. Moreover, for a higher tf score in a longer document to be dampened, all vectors were normalized to the same length. After these processes, the semantic similarity between two tweets was measured as the cosine similarity between two vectors, which is defined in (16):

(16) Cosine Similarity between Two Vectors

t^1, t^2: two tf-idf-weighted length-normalized tweet-vectors

$t^1 = \left(\text{tf.idf}_1^1, \text{tf.idf}_2^1, \cdots, \text{tf.idf}_n^1\right)$; $\left|t^1\right| = 1$

$t^2 = \left(\text{tf.idf}_1^2, \text{tf.idf}_2^2, \cdots, \text{tf.idf}_n^2\right)$; $\left|t^2\right| = 1$

$t^1 \cdot t^2$: inner product of t^1 and t^2

$$\text{cos.sim}\left(t^1, t^2\right) = \frac{t^1 \cdot t^2}{\left|t^1\right| \left|t^2\right|} = t^1 \cdot t^2 = \sum_{i=1}^{n} \left(\text{tf.idf}_i^1 \times \text{tf.idf}_i^2\right)$$

[26] Since a document contains only a small part of all possible terms, most elements of its vector have a value of 0.

Relevance of each tweet t in the test set was defined as the maximum of its cosine-similarities with all *snowpiercer*-containing tweets $t_q \in T_Q$:

$$\text{relevance}_{\text{tf.idf}}(t) := \max_{t_q \in T_Q} \text{cos.sim}(t, t_q).$$

This model retrieves tweets in order of the relevance score.

5.4 PROPOSED MODELS

5.4.1 Strict serialization Model

The SS model retrieves a tweet if and only if the tweet is serialized with any tweet that contains *snowpiercer*, as defined in the subsection 4.3.1.1. Temporal proximity (with threshold level at 30 seconds, 60 seconds, user-specific 5% quantiles, and user-specific 15% quantiles), reply relation, Twitter-specific and continuation markers will be used as features.

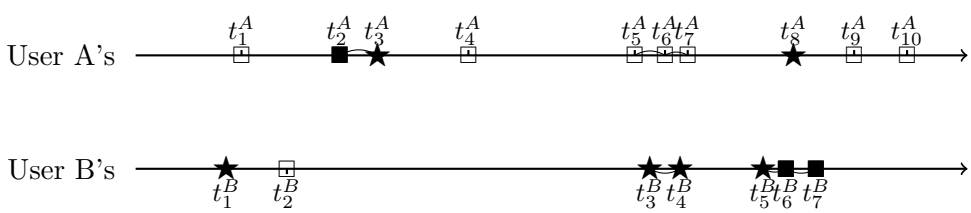


Figure 5.3: Strict serialization Model: *Time* feature

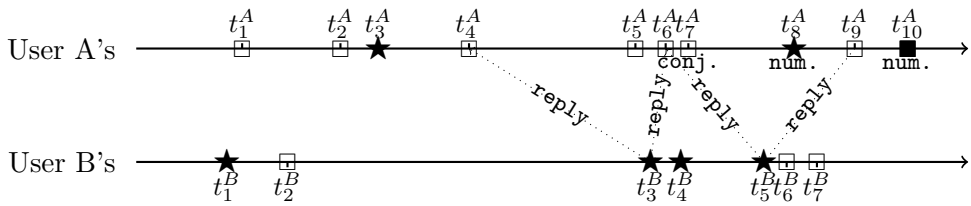


Figure 5.4: Strict serialization Model: *Discourse Marker* feature

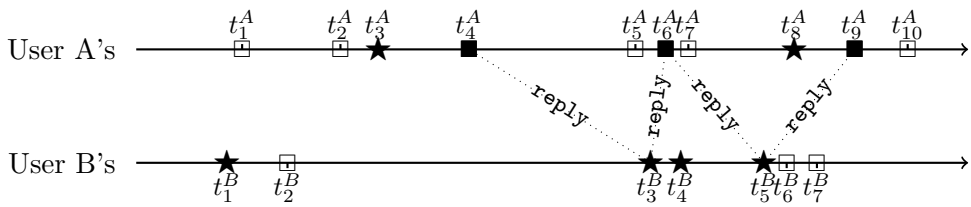


Figure 5.5: Strict serialization Model: *Reply Relation* feature

5.4.2 Serialization Plus Discourse Segmentation Model

5.4.2.1 First stage: Serialization pnly

The SPDS model identifies serialization as discourse segmentation and uses only a temporal proximity feature (with threshold level at user-specific 15% quantiles). In the data set user-specific time thresholds have a minimum of 90, a maximum of 1,060, a median of 208, and a mean of 307 seconds.

5.4.2.2 Second stage: Discourse segmentation

Tweets serialized in the first stage are segmented based on some or all of the shift markers.

RT. A tweet is regarded as a segment initiator only if it is an RT.

URL. A tweet is regarded as a segment initiator only if it contains a URL.

AH. A tweet is regarded as a segment initiator only if its text begins with an interjection from Table 4.3.

NOW. A tweet is regarded as a segment initiator only if the first five words of its text include a temporal adverb from Table 4.4.

All. A tweet is regarded as a segment initiator if any of the above four conditions are met.

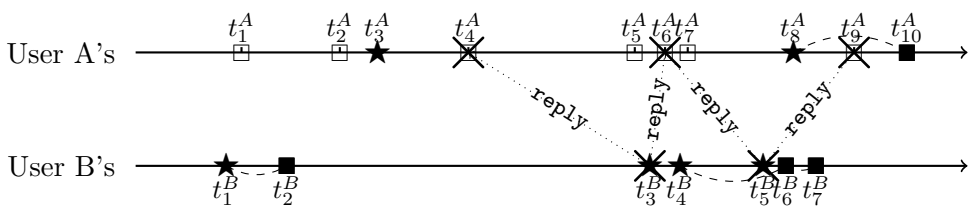


Figure 5.6: Serialization Plus Discourse Segmentation Model: Serialization Only; *monologues* only

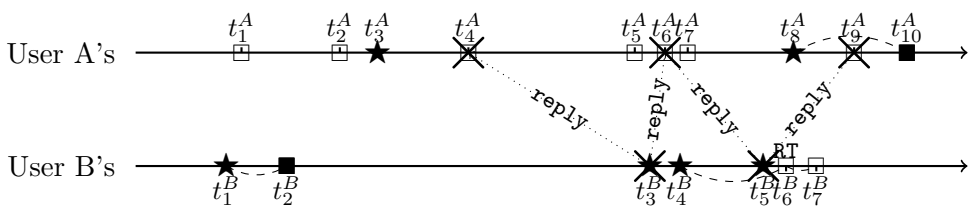


Figure 5.7: Serialization Plus Discourse Segmentation Model: *Discourse Marker* feature; *monologues* only

5.5 EVALUATION

5.5.1 Measures

First, the TF-IDF Weighting model is evaluated by average precision (AP).

(17) Average Precision:

$$\begin{aligned}\text{AP@}k\%(m_i) &= \frac{1}{k}\sum_{j=1}^{k}\text{precision@}j\%(m_i) \\ &= \frac{1}{k}\sum_{j=1}^{k}\frac{\text{the number of relevant tweets retrieved in } m_i \text{ up to } j\%}{\text{the number of tweets retrieved in } m_i \text{ up to } j\%}\end{aligned}.$$

For the other models, each set retrieval system is converted to a ranked retrieval system through 1,000 random replications, where retrieved tweets are first ranked and followed by non-retrieved ones. Then each model is evaluated by the mean of AP (μAP) over replicated samples.[27] Precision is computed at every percentile of recall levels. In sum, the performance of a model m is defined as

(18) Mean of Average Precision:

$$\mu\text{AP@}k\%(m) = \frac{1}{1000}\sum_{i=1}^{1000}\text{AP@}k\%(m_i)$$

where m has 1,000 replicates $m_1, \cdots, m_{1000}$. Conversion and evaluation process for the model is exemplified in Figure 5.8.

[27] μAP should not be confused with MAP (mean of average precision *over queries*).

Set retrieval

Retrieved (by a model)	■$_1$	■$_2$	■$_3$	■$_4$	■$_5$	□$_1$	□$_2$	□$_3$	□$_4$	□$_5$
Relevant (by the annotators)	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗

Ranked retrieval

Replicate #1 (m_1):

Rank			Recall	Precision	Average Precision
1	■$_1$	✓	1/5 = 20%	P@20% = 1/1 = 1	AP@20% = (1)/1 = 1
2	■$_3$	✓	2/5 = 40%	P@40% = 2/2 = 1	AP@40% = (1 + 1)/2 = 1
3	■$_2$	✗	2/5 = 40%		
4	■$_5$	✗	2/5 = 40%		
5	■$_4$	✓	3/5 = 40%	P@60% = 3/5 = .6	AP@60% = (1 + 1 + .6)/3 = .87
6	□$_4$	✗	3/5 = 60%		
7	□$_1$	✗	3/5 = 60%		
8	□$_3$	✓	4/5 = 80%	P@80% = 4/8 = .5	AP@80% = (1 + ⋯ + .5)/4 = .78
9	□$_2$	✓	5/5 = 100%	P@100% = 5/9 = .56	AP@100% = (1 + ⋯ + .56)/5 = .73
10	□$_5$	✗	5/5 = 100%		

Replicate #2 (m_2):

1	■$_4$	✓	1/5 = 20%	P@20% = 1/1 = 1	AP@20% = (1)/1 = 1
2	■$_1$	✓	2/5 = 40%	P@40% = 2/2 = 1	AP@40% = (1 + 1)/2 = 1
3	■$_5$	✗	2/5 = 40%		
4	■$_3$	✓	3/5 = 60%	P@60% = 3/4 = .75	AP@60% = (1 + 1 + .75)/3 = .92
5	■$_2$	✗	3/5 = 60%		
6	□$_4$	✗	3/5 = 60%		
7	□$_2$	✓	4/5 = 80%	P@80% = 4/7 = .57	AP@80% = (1 + ⋯ + .57)/4 = .83
8	□$_1$	✗	4/5 = 80%		
9	□$_5$	✗	4/5 = 80%		
10	□$_3$	✓	5/5 = 100%	P@100% = 5/10 = .5	AP@100% = (1 + ⋯ + .5)/5 = .76

. . . . . .

Replicate #1000 (m_{1000}):

1	■$_5$	✗	0/5 = 0%		
2	■$_2$	✗	0/5 = 0%		
3	■$_3$	✓	1/5 = 20%	P@20% = 1/3 = .33	AP@20% = (.33)/1 = .33
4	■$_1$	✓	2/5 = 40%	P@40% = 2/4 = .5	AP@40% = (.33 + .5)/2 = .42
5	■$_4$	✓	3/5 = 60%	P@60% = 3/5 = .6	AP@60% = (.33 + .5 + .6)/3 = .48
6	□$_3$	✓	4/5 = 80%	P@80% = 4/6 = .67	AP@80% = (.33 + ⋯ + .67)/4 = .53
7	□$_4$	✗	4/5 = 80%		
8	□$_2$	✓	5/5 = 100%	P@100% = 5/8 = .63	AP@100% = (.33 + ⋯ + .63)/5 = .55
9	□$_5$	✗	5/5 = 100%		
10	□$_1$	✗	5/5 = 100%		

$$\mu AP@20\% = (\ \ 1 + \ \ 1 + \cdots + 0.33)\,/1000$$
$$\mu AP@40\% = (\ \ 1 + \ \ 1 + \cdots + 0.42)\,/1000$$
$$\mu AP@60\% = (.87 + .92 + \cdots + 0.48)\,/1000$$
$$\mu AP@80\% = (.78 + .83 + \cdots + 0.53)\,/1000$$
$$\mu AP@100\% = (.73 + .76 + \cdots + 0.55)\,/1000$$

Figure 5.8: Retrieval conversion

5.5.2 Results

5.5.2.1 Strict serialization Model

The μAP values for baseline methods and for all features in the SS model are summarized in Table 5.2. Every feature outperforms the baseline methods up to recall level of 10%, with a significant difference according to a t-test. These differences are larger measured on a smaller recall, because as shown in Figure 5.9, high precisions values are maintained until a recall level of around 10%, where tweets that are serialized to some explicitly relevant tweets are being ranked. This means that even though the serialization cannot cover all implicitly relevant tweets, it is quite helpful in finding a part of them with high precision.

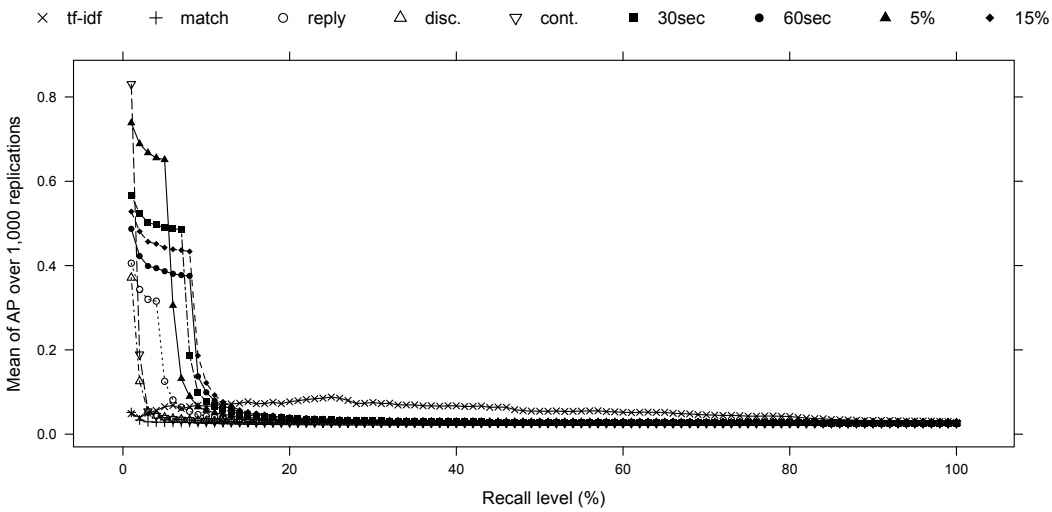
	Baselines					Time difference threshold			
	Match	tf-idf	Repl.	Disc.	Cont.	30sec	60sec	5%	15%
μAP@5%	.0342	.0518	.3019	.1266	.2313	.5158	.4178	**.6804**	.4720
μAP@10%	.0309	.0588	.1798	.0801	.1324	.3916	.3459	**.4050**	.3976
μAP@25%	.0284	.0695	.0920	.0494	.0702	.1824	.1665	.1847	**.1894**
μAP@50%	.0273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

Disc.: Linguistic continuation markers; Cont.: Twitter-specific continuation markers

Table 5.2: Strict serialization

5.5.2.2 Serialization Plus Discourse Segmantation Model

The μAP values for baseline methods and all versions of the SPDS model are summarized in Table 5.4. Figure 5.11 shows that each version outperforms the query matching model at all recall levels, and the tf-idf weighting model up to 30% recall



disc.: Linguistic continuation markers; cont.: Twitter-specific continuation markers

Figure 5.9: Strict serialization

level. All differences with baselines were statistically significant according to a t-test.

According to Table 5.4, linguistic shift markers such as AH and NOW might seem not to perform better than Serialization Only version. However, difference between the RT+URL markers model and the All markers model was also statistically significant. Though such expressions as interjections and temporal adverbs were typically treated as meaningless words and then removed in information retrieval, it is worth noting that these linguistic discourse markers play an important role in accounting for discourse segmentation of tweets.

It might be unexpected that the tf-idf weighting model was not better than the simple query-matching one. When determining whether a tweet and its successor share a topic or not, readers may depend mainly on locality shown by temporal

proximity and continuity assumed from the absence of a shift marker, rather than on word-distributional similarity between two tweets.

	Match	Serial	Serialization + Discourse segmentation					
			RT	URL	AH	NOW	RT+URL	All
μAP@5%	.0415	.3880	.4320	.4264	.4010	.4153	.4701	**.5113**
μAP@50%	.0264	.2234	.2415	.2503	.2247	.2391	.2610	**.2696**
μAP@100%	.0252	.1260	.1350	.1396	.1266	.1340	.1448	**.1488**

Table 5.3: Discourse segmentation: *monologues* only

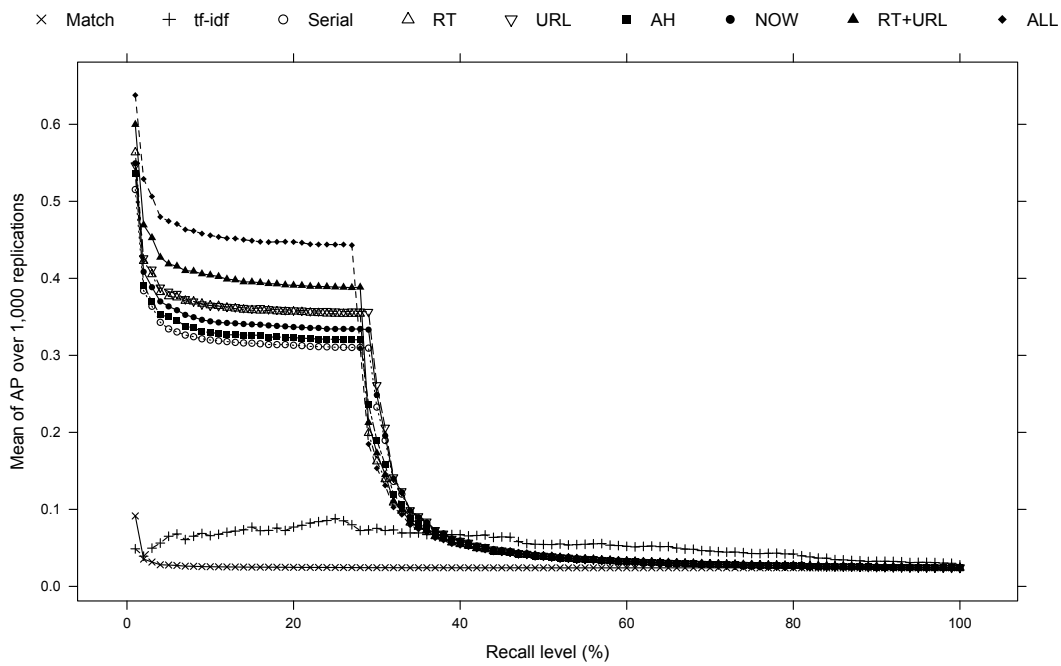


Figure 5.10: Discourse segmentation: *monologues* only

For the comparison of the SPDS model with the SS model to be valid, they need to be implemented on the same data, which means that the data for the latter model should also include dialogues. Results from automatically collecting dialogues

`in_reply_to_id_str` values are presented in Table 5.4 and Figure 5.11. μAP values achieved by the SPDS Model with all features are not only as high as those by the SS model, but also decrease more slowly as recall values increase.

	Match	Serial	Serialization + Discourse segmentation					
			RT	URL	AH	NOW	RT+URL	All
μAP@5%	.0376	.5132	.5625	.5567	.5270	.5378	.6003	**.6456**
μAP@50%	.0328	.3638	.3925	.3970	.3703	.3825	.4162	**.4453**
μAP@100%	.0323	.2019	.2162	.2186	.2051	.2114	.2279	**.2425**

Table 5.4: Discourse segmentation: including *dialogues*

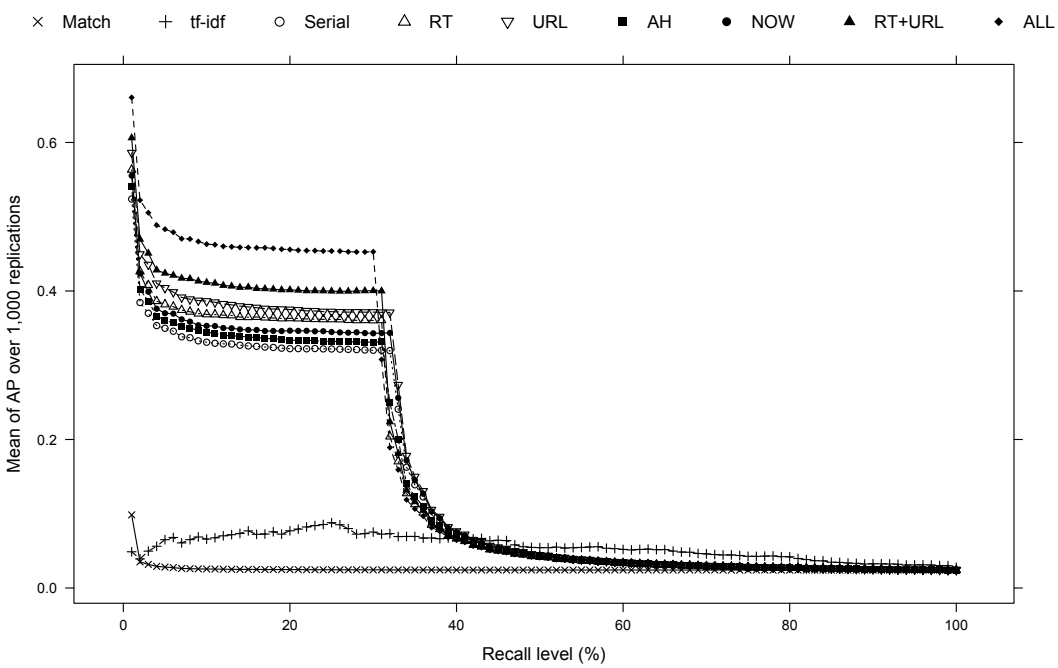


Figure 5.11: Discourse segmentation: including *dialogues*

Finally, relevance feedback was mimicked by adding *bong* and *joon-ho bong*, the name of the movie's director, to our query set and implementing Algorithm 1 again.

The words could be extracted automatically too since they frequently co-occur with *Snowpiercer*. Table 5.5 and Figure 5.12 show that this 'mock' feedback improves the performance of Serialization Plus Discourse Segmentation Model on *monologues* and *quotations*.

	Match	Serial	Serialization + Discourse segmentation					
			RT	URL	AH	NOW	RT+URL	All
μAP@5%	.0418	.4030	.4417	.4608	.4140	.4244	.4875	**.5248**
μAP@50%	.0264	.2468	.2651	.2825	.2495	.2637	.2916	**.3076**
μAP@100%	.0253	.1383	.1472	.1562	.1395	.1467	.1605	**.1683**

Table 5.5: Discourse segmentation after *relevance feedback*: *monologues* only

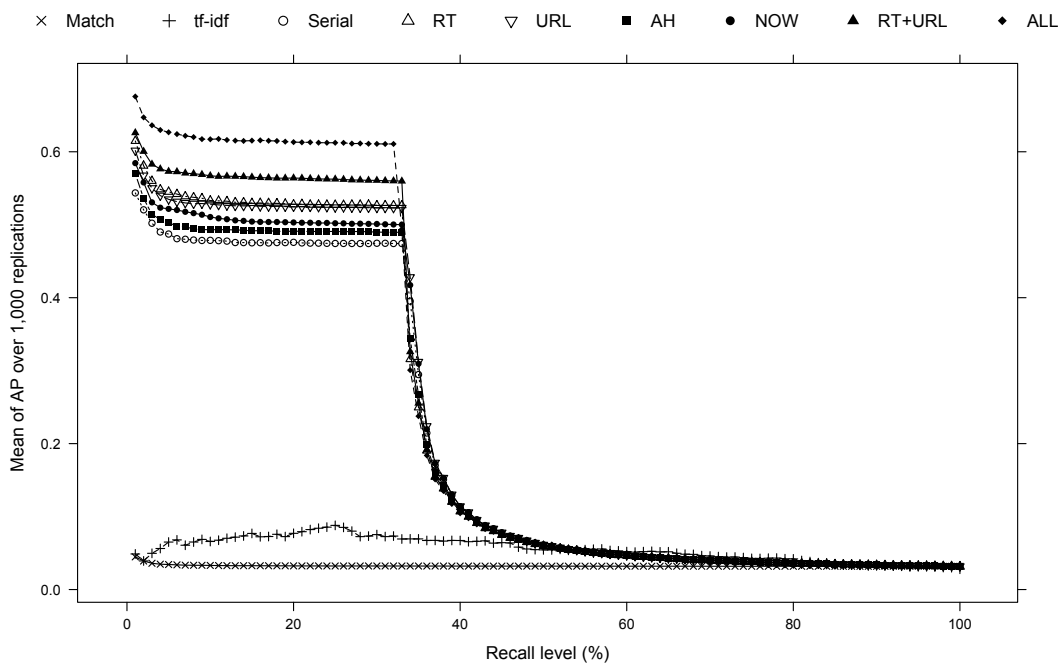


Figure 5.12: Discourse segmentation after *relevance feedback*: *monologues* only

On the other hand, when *dialogues* were added, the effect of relevance feedback

was not observed, as shown in Table 5.6 and Figure 5.13.

	Match	Serial	Serialization + Discourse segmentation					
			RT	URL	AH	NOW	RT+URL	All
μAP@5%	.0431	.3910	.4407	.4563	.4133	.4235	.4790	**.5270**
μAP@50%	.0267	.2451	.2649	.2825	.2490	.2628	.2910	**.3064**
μAP@100%	.0254	.1375	.1472	.1562	.1392	.1464	.1602	**.1677**

Table 5.6: Discourse segmentation after *relevance feedback*: including *dialogues*

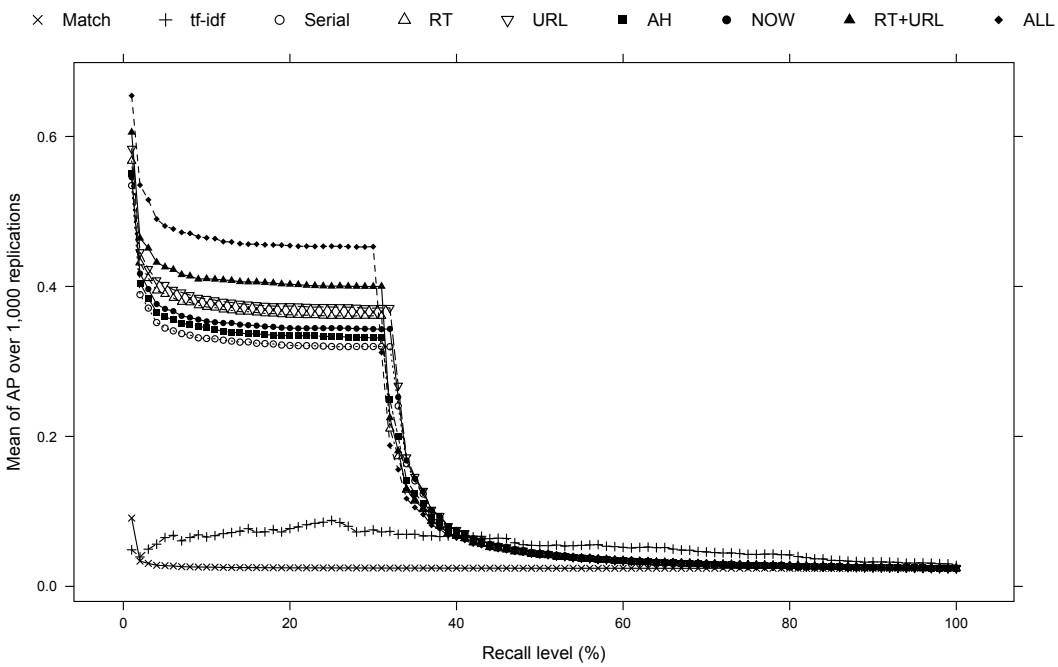


Figure 5.13: Discourse segmentation plus *relevance feedback*: including *dialogues*

5.6 DISCUSSION

5.6.1 Proper formalization of temporal proximity

The adequate threshold level and the proper direction of temporal proximity can be investigated by comparing the results of the two schemes. Table 5.7 presents μAP values for three models which used the temporal proximity feature only.

	±5%	±15%	+50%
μAP@5%	**.6804**	.4720	.3880
μAP@50%	.1847	.1894	**.2234**
μAP@100%	.0669	.0687	**.1260**

Table 5.7: Temporal proximity with respect to direction (±/+) and threshold levels

With a lower threshold level, precision values get higher, but only within a very narrow recall level. As more relevant tweets (and much more irrelevant ones) are retrieved, average precision values decrease, more radically with a lower threshold level. The actual distribution of time intervals from the data set in Table 5.8 accounts for this tendency.

Quantile	u1	u2	u3	u4	u5	u6	u7	u8	u9	u10
5%	20	42	18	16	13	30	21	18	13	8
15%	47	52	33	57	30	56	67	57	40	23
50%	237	298	164	322	151	242	1060	297	167	159
	u11	u12	u13	u14	u15	u16	u17		Mean	SD
	43	23	18	15	13	12	110		24.9	23.3
	89	51	37	61	27	40	161		54.3	31.8
	297	317	178	258	90	266	725		307.5	237.9

Table 5.8: Statistics of time intervals (in seconds) between adjacent tweets

Setting the threshold level at 5% or 15% means that two tweets are considered to belong to the the same discourse segment only if they are written in less than thirty seconds (mean of 5% threshold levels over 17 users: 24.94 seconds) or a minute (mean of 15% threshold levels over 17 users: 54.29 seconds). This is insufficient for detecting serialization because only tweets in *Comment after RT* (7, 12) and *On-the-spot correction* (9) types tend to show such immediacy. In situations such as *Intentional serialization* (8) and *Free-style addition* (10, 11), there are usually several minutes of gap between adjacent tweets. Consequently, setting the threshold level of temporal proximity at a generous amount to increase recall rate is justified empirically as well as theoretically, offering evidence to the preference of continuation to shift.

It should be noted again that a serialization model with the 5% threshold level obtained a precision rate less than 80% even at the 1% recall level (see Figure 5.9). Even in such a close neighborhood of an explicitly relevant tweet, whose distance is at most one minute, more than 20% tweets proved to be irrelevant. Such tweets include ones that were written immediately *before* a tweet that mentioned *snowpiercer*, which would not be counted as relevant under a model using forward proximity. This indicates that a constraint on direction of temporal proximity is more useful for increasing precision rate than that on an upper limit of the time span.

5.6.2 Effects of discourse markers

For each type of discourse markers, whether it is linguistic or web-specific, the experiment results show that using it is better than not using it. Fortunately, using the fixed lists of markers is a far more lightweight method than annotating discourse information manually and implementing a dependency parser. However, different types of these markers show different performance improvement. Continuation markers showed limited effect in a sense that their high precision rates maintained only while retrieving several percents of all relevant tweets, which indicates these markers have a very low coverage. This may be partly because continuation markers are redundant, given that continuation is the default transition state. It also needs to be taken into account that linguistic continuation markers were borrowed from the study for classification of written text (Kang, 1999) and hence are less suitable for unstructured text such as tweets.

Interjections and temporal adverbs, categorized into linguistic shift markers in this thesis, seem to behave differently than in oral speech. In oral speech, these expressions occur as disfluencies. On the contrary, "disfluency" in Web emerges in form of typos, while using interjections is far from a mistake. At the same time, in social media, interjections do not seem to be as improper as in formal documents. In text written in a spoken style, interjections are meaningful as well as intentional, and their discourse function should not be overlooked.

6 Conclusion

Based on the original purpose and characteristics of social media, this thesis assumed that discourse relations would exist among tweets, and presented a variety of real examples of those relations from Korean Twitter data. The way of that tweets constitute discourse segments was captured theoretically on the basis of centering framework, and then empirically by means of using temporal proximity and discourse markers. The experiments for informational retrieval task showed that the discourse-segment-level Twitter search suggested in this thesis achieved better performance than the current tweet-level search.

And yet, this thesis has at least two unsettled issues:

1. It did not examine how many tweets were serialized in the whole data. That work requires more elaborated annotations and more accurate criteria for discourse segments, which would be the subject of another study. If there are data on serialized tweets, features for tweet serialization will be able to be learned.

2. Although previous studies have shown that different interjections have different functions, this thesis treats interjections as a single category so it cannot verify which of them initiates shift or continuation.

Despite these limits, this study made several contributions as summarized below:

1. For tasks which pool multiple tweets into a single “document”, the proposed models have obtained more coherent results. Furthermore, discourse segments detected in the models will be better suited for other tasks such as topic/item/event discovery and sentimant analysis.

2. The proposed models were implemented in a lightweight way, not only by using discourse markers instead of annotation markers and dependency markers, but also by focusing on shift markers only under principles of traditional discourse theories.

3. Shift markers such as interjections, which have been typically treated as stop-words in information retrieval, proved to be useful for discourse segment detection.

Finally, as topic continuation is generally unmarked, future research shall focus on topic shift detection, including the distinction of topic shift with attribute shift, and discovery of other linguistic features such as tense and aspect.

Bibliography

Bestgen, Y. (1998). Segmentation markers as trace and signal of discourse structure. *Journal of Pragmatics*, 29(6):753–63.

Bestgen, Y. and Vonk, W. (1995). The role of temporal segmentation markers in discourse processing. *Discourse Processes*, 19(3):385–406.

Bestgen, Y. and Vonk, W. (2000). Temporal adverbials as segmentation markers in discourse comprehension. *Journal of Memory and Language*, 42(1):74–87.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bracewell, D., Tomlinson, M., and Wang, H. (2012). Identification of social acts in dialogue. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, volume 1, pages 375–90, Mumbai, India. Association for Computational Linguistics.

Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL 1987)*, pages 155–62, Stanford, California, United States of America. Association for Computational Linguistics.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–24.

Clark, H. H. and Haviland, S. E. (1977). Comprehension and the given-new contract. In Freedle, R. O., editor, *Discourse Production and Comprehension*, pages 1–40. Ablex, Norwood.

Cremonesi, P., Pagano, R., Pasquali, S., and Turrin, R. (2013). Tv program detection in tweets. In *Proceedings of the 11th european conference on Interactive TV and video (EuroITV '13)*, pages 45–53, New York, New York, United States of America. ACM Press.

Dascalu, M. (2014). *Analyzing Discourse and Text Complexity for Learning and Collaborating: A Cognitive Approach Based on Natural Language Processing*, volume 534 of *Studies in Computational Intelligence*. Springer International Publishing Switzerland.

Degand, L. and Sanders, T. (2002). The impact of relational markers on expository text comprehension in L1 and L2. *Reading and Writing*, 15(7):739–57.

Feng, S., Zhang, L., Li, B., Wang, D., Yu, G., and Wong, K.-F. (2013). Is Twitter a better corpus for measuring sentiment similarity? In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*,

pages 897–902, Seattle, Washington, United States of America. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–82.

Forbes, K. and Miltsakaki, E. (2002). Empirical studies of centering shifts and cue phrases as embedded segment boundary markers. *University of Pennsylvania Working Papers in Linguistics*, 7(2):39–57.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, 31(7):931–52.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011): Human Language Technologies: Short Papers*, volume 2, pages 42–47. Association for Computational Linguistics.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–25.

Heim, I. (1982). *The Semantics and Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts at Amherst.

Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational linguistics*, 19(3):501–30.

Huang, J., Zhou, B., Wu, Q., Wang, X., and Jia, Y. (2011). Contextual correlation based thread detection in short text message streams. *Journal of Intelligent Information Systems*, 38(2):449–64.

Joty, S. R. (2013). *Discourse Analysis of Asynchronous Conversations*. PhD thesis, University of British Columbia.

Kameyama, M. (1998). Intrasentential centering: A case study. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering : A Framework for Modeling the Local Coherence of Discourse*, chapter 6, pages 89–112. Oxford University Press.

Kamp, H. (1981). A theory of truth and semantic representation. In Groenendijk, J., Janssen, T. M., and Stokhof, M., editors, *Formal Methods in the Study of Language*. Mathematische Centrum.

Kang, B.-M. (1999). *Hankwukeui Theyksuthu Cangluwa Ene Thukseng* (Text Genres and Linguistic Characteristics in Korean). Korea University Press, Seoul, Korea.

Karamanis, N., Mellish, C., Poesio, M., and Oberlander, J. (2009). Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1):29–46.

Kumar, S., Liu, H., Mehta, S., and Subramaniam, L. V. (2014). From tweets to events: Exploring a scalable solution for Twitter streams. *Computing Research Repository (CoRR)*, abs/1405.1392.

Lee, C.-H. (2012). Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams. *Expert Systems with Applications*, 39(18):13338–56.

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. *Proceedings of the 36th Annual Conference of the ACM Special Interest Group on Information Retrieval (SIGIR '13)*, pages 889–92.

Mishne, G., Dalton, J., Li, Z., Sharma, A., and Lin, J. (2013). Fast data in the era of big data: Twitter's real-time related query suggestion architecture. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Management of Data (SIGMOD '13)*, pages 1147–58.

Montes, R. G. (1999). The development of discourse markers in Spanish: Interjections. *Journal of Pragmatics*, 31(10):1289–1319.

Mukherjee, S. and Bhattacharyya, P. (2012). Sentiment analysis in Twitter with lightweight discourse analysis. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1847–64, Mumbai, India. Association for Computational Linguistics.

Norrick, N. R. (2009). Interjections as pragmatic markers. *Journal of Pragmatics*, 41(5):866–91.

Ochs, E. (1979). Planned and unplanned discourse. In Givón, T., editor, *Discourse and Syntax*, volume XII of *Syntax and Semantics*, pages 51–80. Academic Press, New York, New York, United States of America.

Park, S. and Shin, H. (2014). Identification of implicit topic in Twitter data not containing explicit search query. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, volume forthcoming, Dublin, Ireland.

Passonneau, R. J. and Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–39.

Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, 14(3):367–81.

Roberts, C. (1998). The place of centering in a general theory of anaphora resolution.

In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering Theory in Discourse*, chapter 18, pages 359–400. Oxford University Press.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–23.

Sanders, T. and Noordmand, L. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29(1):37–60.

Segal, E. M., Duchan, J. F., and Scott, P. J. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes*, 14:27–54.

Somasundaran, S. (2010). *Discourse-Level Relations for Opinion Analysis*. PhD thesis, University of Pittsburgh.

Somasundaran, S., Namata, G., Getoor, L., and Wiebe, J. (2009a). Opinion graphs for polarity and discourse classification. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 66–74, Morristown, NJ, USA. Association for Computational Linguistics.

Somasundaran, S., Namata, G., Wiebe, J., and Getoor, L. (2009b). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods*

in Natural Language Processing (EMNLP 2009), volume 1, pages 170–79. Association for Computational Linguistics.

Somasundaran, S., Ruppenhofer, J., and Wiebe, J. (2007). Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue (SIGdial 2007)*, pages 26–34.

Somasundaran, S., Ruppenhofer, J., and Wiebe, J. (2008a). Discourse level opinion relations: An annotation study. In *Proceedings of the 9th SIGDIAL Workshop on Discourse and Dialogue (SIGdial 2008)*, pages 129–37.

Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in online debates. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP '09)*, 1:226–34.

Somasundaran, S., Wiebe, J., and Ruppenhofer, J. (2008b). Discourse level opinion interpretation. In *Proceedings of the 22th International Conference on Computational Linguistics (COLING 2008)*, pages 801–8, Manchester, United Kingdom. Association for Computational Linguistics.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Thanh, H. L., Abeysinghe, G., and Huyck, C. (2004). Automated discourse segmentation by syntactic information and cue phrases. In *Proceeding of Artificial Intelligence and Applications (AIA 2004)*, pages 2–7.

Walker, M. A., Iida, and Cote, S. (1994). Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–233.

Walker, M. A., Joshi, A. K., and Prince, E. F. (1998). Centering in naturally occurring discourse: An overview. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering : A Framework for Modeling the Local Coherence of Discourse*, chapter 1, pages 1–28. Oxford University Press.

Wei, Z. (2014). Local coherence in stream-of-consciousness discourse : A centering approach. *Canadian Social Science*, 10(1):83–87.

Weng, J., Lim, E.-P., Jiang, J., and He., Q. (2010). TwitterRank: Finding topic-sensitive influential Twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 261–70. ACM.

Wolf, F., Gibson, E., and Desmet, T. (2004). Discourse coherence and pronoun resolution. *Language and Cognitive Processes*, 19(6):665–75.

초록

트위터 게시물간 발화 연쇄 · 담화 분절 탐지 및 질의어 비포함 트윗 검색에의 활용

이 논문에서는 2000년대 중반 이후 한국을 비롯한 세계 각지에서 활발하게 이용되고 있는 사회 관계망 서비스인 트위터에서 사용자들이 작성한 여러 게시물, 즉 트윗이 하나의 담화 분절로 묶이는 현상을 기술하고, 한 사용자가 연이어 작성한 트윗 두 개가 하나의 메시지를 연쇄적으로 전달하는지를 자동으로 포착하는 규칙 기반 모형을 구축하고자 한다. 자연어처리 분야에서 일반적으로 웹 게시물 하나는 독립적인 문서 하나로 간주되었으나, 트윗은 길이가 140자로 제한되어 있으므로 하나의 문서로 완결되기보다는 더 큰 단위에 포함될 수 있다. 본 논문에서는 이러한 상위 단위를 담화 분절로, 트윗 하나를 발화 하나로 간주하고, 한 담화 분절에 속한 발화 사이의 연결에 관하여 중심화 이론에서 제시한 여러 제약을 토대로 네 가지를 가정한다.

(가) 트윗 하나의 주제는 한 개를 넘을 수 없다.

(나) 여러 트윗이 한 담화 분절을 이룰 때 두 번째 트윗부터는 주제어가 대용어로, 한국어에서는 특히 영형태로 실현된다.

(다) 한 사용자가 작성한 두 트윗 사이의 응집성은 사이에 다른 트윗이 없을 때에만 유효하다.

(라) 한 트윗에서 다음 트윗으로 넘어갈 때 주제는 이전 주제의 지속>보유>전이 순으로 선호된다.

이러한 가정을 반영하여 인접한 두 트윗이 서로 연쇄적으로 작성되었는지, 같은 담화 분절 내에 속하는지를 예측하기 위해 시간적 근접성과 담화 표지 두 가지 기준을 제시한다. 시간적 근접성은 두 트윗 사이의 시간 간격이 특정한 값 이하인지 아닌지로 측정된다. 이 한계점은 60초와 같이 상수로 설정되거나, 한 사용자가 작성한 트윗 전체에서 구한 시간차 값들의 상위 15% 분위 등 사용자마다 다르게 설정될 수 있다. 담화 표지는 기능에 따라 지속 표지와 전이 표지로, 매체에 따라 웹 표지와 언어 표지로 분류된다. 지속 표지로는 '>>', '(계속)', 숫자와 같은 웹 표지와 접속사 및 지시 표현과 같은 언어 표지가, 전이 표지로는 'RT', URL과 같은 웹 표지와 간투사 및 시간 부사와 같은 언어 표지가 설정된다. 이 두 가지 요인을 사용하여 규칙을 만드는 방식에 따라 두 가지 모형이 나온다. 먼저 '엄밀한 연쇄화'(SS) 모형에서는 각 트윗이 서로 독립적이라는 기존 처리 방식을 크게 벗어나지 않으면서, 두 트윗이 시간차가 매우 짧거나 지속 표지가 있을 때에만 연쇄화된 것으로 간주한다. 반면 '연쇄화+담화 분절화'(SPDS) 모형에서는 지속 상태가 선호된다는 중심화 이론의 가정 (라)를 토대로 두 트윗 사이의 시간차가 너무 크지 않은 한 연쇄화된 것으로 본 뒤, 전이 표지가 있을 때에만 연쇄화를 중단하고 담화 분절을 종료한다.

트윗 연쇄화 및 담화 분절화가 실제 과제에서 유용한지는 정보 검색 과제에서

확인된다. 여러 트윗이 한 주제에 관하여 담화 분절을 이룰 때 일부 트윗에서 주제어가 생략되거나 영형태로 나타나는 현상은 중심화 이론의 가정 (나)에서 예측되며 실제 자료에서도 관찰된다. 기존의 질의어 일치 방식은 이러한 트윗을 찾아내지 못하므로, 트위터에서 다양한 의견을 수집하고자 하는 사용자의 필요에 부응하는 데 부족한 점이 있다. 이 논문에서 제시하는 모형을 적용하여 연쇄화된 트윗을 담화 분절로 묶고 검색하면, 질의어를 직접 포함하는 트윗과 같은 담화 분절에 속하는 트윗을 추가로 찾을 수 있고, 이 과정에서 질의어와 무관한 트윗을 잘못 찾게 되는 비율도 낮다. 결과적으로 질의어 일치 모형 및 TF-IDF 모형보다 높은 평균 정밀도 수치를 얻을 수 있다. 또한, SS 모형보다 SPDS 모형이 더 나은 검색 성능을 보인다는 점에서, 발화가 지속되는 것을 무표적인 현상으로 보는 기존 담화 이론의 원리를 트위터를 비롯한 사회 관계망에서 사용되는 언어에도 적용할 수 있음을 확인할 수 있다. 한편 간투사나 시간 부사와 같이 기존 검색 방식에서 불용어로 처리한 언어 표현이 문서의 경계를 찾아서 더 정확한 정보를 얻게 하는 데 유용하다는 것 역시 이 논문의 또 다른 의의이다.

주요어: 담화표지, 사회관계망서비스, 정보검색, 중심화이론, 트위터

학번: 2012-20031