JOURNAL OF ASIAN SOCIOLOGY
Volume 48 | Number 3 | September 2019, 321-342
DOI 10.21588/jas/2019.48.3.002 **Special Issue**

Possibility of Discourse Analysis using Topic Modeling

Wonkwang Jo | Seoul National University

This study is an attempt to introduce topic modeling as a method for discourse analysis in order to explore new possibilities for discourse analysis. Human language data, which is used for discourse analysis, holds plenty of information, however, traditional research methods on language data have several limitations. Topic modeling, which is a statistical analysis method applied to language data, is suitable for a discourse analysis for three reasons: (1) The "topic" extracted via topic modeling contains useful information for inferring discourse. The information shows the key functions of the particular discourse. (2) Topic modeling's multiple topic assumption makes it possible to examine the dynamics of discourses. (3) Recent topic modeling techniques allow researchers to study changes in discourse over time as well as interactions between discourse and non-discursive factors. Although topic modeling methods have limitations, the shortcomings can be complemented and remedied. Furthermore, text mining, including topic modeling, is not limited to discourse analysis and can be applied to the study of various variables and concepts in social science. The social sciences must make an effort to better understand and best utilize these new methods.

Keywords: *topic modeling, discourse analysis, language data, text mining*

Human language data and research methods

Two traditional methods

This research is an attempt to study the new possibilities that topic modeling methods are introducing to discourse analysis and examine the implications of these possibilities. I argue that topic modeling can be a way to measure key information regarding discourses in a reproducible way, and that it has the potential to open a new chapter in discourse studies, as it could complement traditional methods. In order to examine this in detail, we must first look at the traditional research methods that study human language data as well as the overall features of the emerging technique called "text mining."

Recently, more and more attempts have been made to analyze human language using computers and statistical models, and they have shown considerable successes in numerous areas. These techniques and methods are called "text mining," which has been defined as "[taking] large amounts of unstructured language and quickly [extracting] useful and novel insights" (Kwartler 2017, p. 17). This is particularly well utilized in corporate strategies such as marketing, as analyses based on text mining can elicit appropriate decision making and thus bolster profits. For example, if a company were able to read consumers' desires as they appeared in various forms of human language (e.g., product reviews), the company could then engineer products which conform to the expectations and desires of consumers. This latest rise in the use of statistical analysis for human language data proves that our language contains much more information than we can imagine. But researchers' interest in human language data has existed longer than their ability to process it computationally or statistically analyze it.

Many researchers in the social sciences, for example, sociologists, have long used human language resources, such as testimonies, literature, and open questionnaires, as data. Many phenomena and variables that researchers have paid attention to are expressed in human language. Discourse, episteme, collective emotion, framing, and knowledge are representative examples. In order to measure and analyze these, researchers had to explore human language data. Given that Max Weber analyzed German media articles and subsequently published a paper on them (Dickinson 2013; Evans and Aceves 2016), we can assume that researchers have long been interested in human language and the variables expressed through it.

Although these variables are expressed in human language, they are

TABLE 1
An Example of Assessment Data on the Restaurant a, type 1

Rater ID	Taste	Customer Service	Cleanliness
1	4	3	3
2	3	4	4
3	5	5	3
4	6	7	6
5	2	2	1

TABLE 2
An Example of Assessment Data on the Restaurant a, type 2

Rater 1	That restaurant is awesome. The owner wore gold rings on each of her fingers and the place is going to be expanded to the shop next-door, which was originally a pet shop. According to something I saw on TV, the total sales of this place is about $20,000 per month.
Rater 2	This place is actually pretty good but finding the place was confusing, I think there are actually two similar restaurants. They need to be clearer with their location and restaurant names. Lunch special is good.
Rater 3	I wanted something with soup, so I decided to find a place that had noodles and ate here. The noodles themselves were very good, definitely among the best I've had, but the house special sauce and ingredients were lacking - the meat was very over cooked, and they were very stingy with the amount of noodles.

neither that clear nor distinct, thus making them difficult to measure and investigate. Emotions are expressed through language, but they are manifested in vague words which readers may have difficulty grasping the meanings and values of. Knowledge belonging to a certain era, though we can easily assume its existence and influence, is not easy to define clearly and distinctly. This is essentially due to the un-structuredness of human language. Human language materials do not have a fixed and consistent form, despite the fact that messages being expressed through the language may be the same. This is significantly different from the data traditionally collected and handled by quantitative research. See the example below.

The first table contains well-structured data, in a format which is

familiar to the social sciences. It is relatively easy to extract information, for example, the means and deviations of each variable and the correlations between variables. However, there are several limitations to this type of data. This kind of data is produced mainly through pre-determined surveys, which might affect the responses due to the framing of questions in the questionnaire. This means that the structure of the survey could contaminate the responses. It is also limiting in that only pre-determined subjects can be investigated. For example, the table above does not provide information other than for the three variables presented, even though there are a lot of other aspects of the restaurant that respondents might be willing to talk about.

The second table contains voluntary reviews left by customers. In this case, they have the advantage of being free from the researcher's influence on the respondents, because the reviews were not responses to pre-determined questions, but voluntarily presented opinions on what the reviewers wanted to address. If we could extract reliable information from this kind of data, it would be a valuable resource for examining people's attitudes or emotions towards specific objects (in this case, a restaurant). But the problem is that it is not easy to extract information from this kind of data. In the above reviews, unexpected topics appear (e.g., paying attention to the rings on the owner's fingers in evaluating the restaurant in the first review), and ambiguous attitudes are expressed (e.g., it is hard to tell if the third reviewer was satisfied). Even the simplest classifications (e.g., positive or negative) present a big challenge here. Natural language data, even when its scale is relatively small, can be called big data because of its un-structured nature (Ham and Chae 2012).

Many researchers in the humanities and social sciences, including sociologists, have traditionally used two methods to extract information from these types of unstructured natural language data. According to Kozlowski, Taddy, and Evans, these two methods could be called "interpretivist text analysis" and "systematic qualitative coding" (Kozlowski, Taddy and Evans 2018). Both are ways of making the most of human perception and intuition. This is because, as we have seen earlier, language possesses a wealth of information, but because its form is not fixed but flexible, we must use human's flexible cognition to explore it. These two methods have been used in numerous studies to produce fine results, but they have also revealed unavoidable limitations and risks. The key points and limitations of each method are as follows.

First, "interpretivist text analysis" is a method in which a well-trained researcher(s) reads the language material in depth and identifies the meaning

and characteristics of the data (Kozlowski et al. 2018). Michel Foucault, who is one of the most important researchers in the twentieth century and who greatly contributed research on 'discourse', utilized this method in his research, which continues to enjoy a profound and lasting impact on social sciences thinking. For example, according to Foucault, the way madness was perceived during the Renaissance period and the way madness is perceived in modern times are entirely disparate. In the Renaissance period, madness was identified as an inscrutable ability or force beyond human rationality. On the other hand, modern madness was identified as a degenerative disease affecting a person's cognitive abilities. Of course, the treatment and medical practice on madness are quite different in these two periods as well (Foucault 2013). Foucault detected this transition in various texts such as administrative documents and research papers from the past. Foucault identified the macroscopic transformation of the unconscious cognitive frame by reading a number of language materials in depth.

As Foucault's works illustrate well, there are masterpieces of sociological research which make use of "interpretivist text analysis" that are very compelling and thought-provoking. However, there are many limitations in terms of their research methods. First, these studies run the risk of being contaminated by the researcher's subjective bias. In other words, it is difficult to tell whether the conclusions of the study are a result of prejudice on the part of the researchers or are actual trends present in the data. Furthermore, it is not easy for other researchers to reproduce the results. Reading and processing large-scale language data itself requires a lot of time, and above all, human cognitive abilities can vary as much as they are flexible. The same data can be focused on different points and hence different conclusions can be drawn.

The second approach, "systematic qualitative coding", is a method in which multiple researchers read the same text and judge/classify the text according to several pre-determined criteria (Kozlowski et al. 2018). If the findings of multiple researchers are consistent with one another, we can be sure that the results do not stem from subjective bias. Studies analyzing media articles are a good example of research based on this method. Slater et al., for example, analyzed how cancer was reported in newspapers, TV, and magazines. First, they set certain criteria for classifying articles, then multiple coders read and classified the media articles, and finally they analyzed how similar these results were to each other (Slater, Long, Bettinghaus and Reineke 2008) In a nutshell, this approach assumes that if multiple researchers' conclusions based on the pre-determined criteria are similar to

each other, the conclusions could be considered objective judgments rather than subjective ones.

This approach secures reproducibility and objectivity in its own way, but there are still problems. In many cases, this method cannot be applied. This approach is useful when the research question is simple. However, the more complex the factors that are to be classified or judged, the more likely the researchers' judgments will vary, and naturally the overall reliability of the results will decrease. For example, it is highly likely that an analysis using systematic qualitative coding on whether news articles on cancer are about prevention or treatment would produce discernible and consistent results, because the criterion is simple, clear, and easy to apply. However, if multiple researchers are asked about the nature of metaphors for cancer, their judgments are not likely to converge, and the results will naturally become unreliable. Furthermore, kappa statistics, which are often used to assess the inter-coder reliability, may be distorted if there are very low frequency categories among the classification categories (Viera and Garrett 2005). The other problem is that the criteria of judgment must be established in advance. This means that it is difficult to employ this method in exploratory research.

Above all, the biggest problem of both approaches is that the scale of data that can be handled is limited. Both methods presuppose that humans read the text. There is a clear limit to the amount of text that human can read. A single human cannot possibly read all the hundreds of thousands of posts on Internet forums. If the goal of the study is to identify macroscopic or 'overall' trends, this becomes a fatal limitation. For example, if you are interested in the overall structure of discussions concerning MERS on Twitter, it is impossible to do the research using either of these two traditional methods.

Text mining

The development of text mining presents a new prospect for overcoming these limitations. The rise of text mining began with human language being recorded in digital form rather than on paper, and the development of technology to transform and process such digitally recorded data. Scores of language materials have begun to be digitized. Daily life and conversation started to be recorded on the Internet and began to be shared (and in many cases voluntarily!), and literature from the past began to be converted to digital form for ease of use (e.g., Google books project). Technologies have been developed that can handle such data using computers. So called "natural

language processing" techniques make it possible to assign the relevant part of speech to individual words with a high degree of accuracy, as well as extract significant morphemes from human-produced sentences.

There are two types of information that most text mining techniques extract from text. The first is the type and frequency of the words in a text. This is simple information, but valuable. Researchers such as Danner, for example, analyzed autobiographical writings left by nuns and found that those who used more positive expressions in the text were more likely to live longer than those who did not (Danner, Snowdon and Friesen 2001). That is, a significant discovery could be made from this simple information. To grasp this information, words must be extracted from sentences in the data. This process is called "tokenization" and is usually performed by morphological analyzers.

The second type of information that text mining can extract concerns the relationship between words. This relationship between words includes the co-occurrence of words in the same sentence or document, the order in which words appear, among other factors. It is important to know how individual words relate to each other in a text. As Ferdinand de Saussure made well known, the meaning and value of a word are defined by the relationship between words, not the individual words themselves (Dosse 1997). Only by considering this relational information can we know the exact meaning and nature of words. For example, many algorithms for part-of-speech tagging utilize the sequence information of words.

There are various ways in which text mining methods utilize these two kinds of information to build a model. From the perspective of statistical learning methods, text mining can be divided into two methods: those based on supervised learning algorithms and those based on unsupervised learning algorithms. Supervised learning refers to adjusting a model in order to better describe or predict a specific response variable. Regression analysis is a good example. On the other hand, an unsupervised learning algorithm is a model-building method for capturing the structure or characteristics of the data, without a clear response variable. Clustering is a good example of this (James, Witten, Hastie and Tibshirani 2013).

Text mining methods based on statistical models also use both types of learning algorithms. Various supervised learning algorithms are used for text analysis. Regression analysis, support vector machines, and neural network analysis are good examples. Research based on supervised learning algorithms mainly aims to better explain the variables of interest using language materials. Kwon et al.'s research that tried to explain the veracity of

messages using linguistic features of the messages (Kwon, Cha and Jung 2017), and Bollen et al.'s attempt to analyze changes in stock prices based on the collective sentiment observed on Twitter (Bollen, Mao and Zeng 2011) are good examples. On the other hand, unsupervised learning algorithms are used for different objectives, such as finding the dominant network structure between words, clustering documents based on linguistic similarity, and extracting topics determined by statistical models. According to Evans and Aceves, "clustering, network analysis, topic modeling, and vector space embedding" are the most common methods utilized in the unsupervised method category (Evans and Aceves 2016).

A key advantage of information extraction using text mining compared to methods utilizing human intelligence is that it is possible to process large volumes of data and yield reproducible results. Although there are differences in the size and processing efficiency of the data depending on specific models, in general, more data can be digested than humans are capable of analyzing. In addition, the results of text mining could show high reproducibility under the condition that same data and algorithms are used. Therefore, text mining is useful for academic discussions as well, as it shows the process of deriving results more clearly than methods that utilize human cognitive abilities. While text mining methods may not replace human cognitive abilities in analysis all together, they can be complementary tools for research that deals with language data.

I, especially, believe that topic modeling methods not only have all the advantages of text mining mentioned above, but their logics and features are adequate for discourse analysis. There are three reasons for this: (1) Topics produced by topic modeling contain two important pieces of information that make up the concept of discourse, which are useful for measuring discourse; (2) topic modeling methods basically extract multiple topics from data text, which are suitable for understanding the nature and dynamics of discourse; (3) using various techniques that emerged after the Latent Dirichlet Allocation (LDA), the most commonly used topic model (Gerlach, Peixoto, & Altmann, 2018), an analysis that examines historical changes in discourse is also possible.

Topic modeling and discourse analysis

Topic and discourse

Topics, one of the most important outcomes of topic modeling, provide useful information for measuring discourse. To understand this, it is necessary to clarify what discourse is. In fact, the definition of discourse varies from scholar to scholar, so I will limit the discussion of this study to the most representative and influential conceptualization of discourse, which belongs to Foucault. According to Foucault, discourse is a potential force that defines valid objects and subjects, and determines the rational mode of connection between them (Foucault 1972, 2002). Normally we think we have freedom of speech, but this is not accurate (Foucault 1971). A society often defines what can be said and how can it be said, and imposes these norms onto individuals. Foucault called this power process a discourse. Modern clinical medicine discourse, for example, defines the norms of valid discussion of diseases. First, the discourse defines valid objects of discussion. To talk about disease under the influence of modern clinical medicine discourse, one should talk about lesions, cells, and germs, not about evil spirits, as may have been typical in the past. Discourse also defines valid connections between objects. Disease should be discussed in conjunction with specific injuries, causative substances, and survival rates. Disease should not be connected to the flow of Qi(氣), evil spirits, or animal spirits. In short, discourse is the potential power which regulates valid usage of language.

Considering how discourse shapes our lives and experiences of the world, it is only natural that there has been constant scholarly interest in discourse. Social scientists have long attempted to understand and analyze human life by investigating factors outside of individual human beings (e.g., status, class, occupation, income level, and environment). If discourse regulates human actions and judgments, and changes depending on space and time, it is expected that social scientists pay attention to it. Moreover, the influence of discourse is not the kind of power which suppresses people, but the one that induces people to internalize certain standards and have certain desires (Foucault 1977, 1990). This is what Steven Luke would have called a "three-dimensional power" (Lukes 2005), which is particularly important in societies with less direct repression. In short, it is due to the importance of discourse's effects that discourse has been consistently studied, although the limitations of traditional research methods on discourse have long been a

subject of criticism.

How is discourse, which controls language usage, expressed in language materials? We should pay attention to two phenomena in language materials: (1) frequently appearing words and (2) network structures between words. If the discourse defines valid objects, then words that point to these objects would appear frequently in the language material that the discourse wields influence over. If the discourse defines the valid way of connecting the objects, then the language data will have many links between words that reflect these connections. Therefore, if the information (frequently appearing words and network structures between words) can be detected reliably, it becomes possible to infer information about the discourses from the text data in reverse.

A topic, which is the main finding of topic modeling methods, contains these two bits of information. It means that topic modeling could be utilized as a method to measure a specific discourse's expression reliably. In order to understand why topic modeling and topics have such a function, we should examine the basic logic and assumptions of topic modeling. In fact, the term topic modeling encompasses many techniques that began with Latent Semantic Indexing (LSI), but I will limit my discussion on basic characteristics of topic modeling to Latent Dirichlet Allocation (LDA), which is the most commonly used technique (Gerlach et al. 2018), for simplicity of discussion. The basic properties of topic modeling discussed below are all LDA attributes.

A topic is defined as the probability distribution of words that is most likely to generate the given text (Blei, Ng and Jordan 2003). Topic modeling has two key assumptions. (1) Each document is a bag of words; meaning that topic modeling utilizes the information on word frequency in each document, and ignores other information, such as word sequence. (2) Given data, a set of documents, is generated randomly from two kinds of probability distribution. The first distribution is topics, which are probability distributions of words, and the second distribution is a probability distribution of topics in each document. Topic modeling assumes that multiple topics exist in data and each document.

The process of a single document's creation as assumed by topic modeling could be summarized as follows. Suppose a document is made up of 100 words and there are four topics. There is information in this document about a probability distribution of topics, i.e. how much weight each of the four topics receive in the document. Based on this probability distribution, which topic each word is used in relation to is determined. Topics given a

high probability in this document will be assigned to many words and topics given a low probability are assigned to a smaller number of words. Suppose 'word number 1' is determined to come from 'topic 1.' 'Word number 1' is then randomly generated from the topic, i.e., a probability distribution of the words. Topic modeling assumes that this process recurs for every word and each document.

Therefore, the main objective of topic modeling is to find which topics and document-specific topic distributions are most likely to generate the given data. Assuming that all of the documents were created in the above manner, topic modeling tries to infer the most likely topics and topics distributions conditioning the current data. This is essentially no different from the process of inferring the probability that a coin will come out heads-up from the result of a hundred rounds of tossing. Topic modeling assumes that a set of documents is generated through a probabilistic process and aims to infer the probability structures behind the data. Whether this is a realistic assumption is controversial, but it is clear that the assumptions enable the study of language data through computers and statistical models.

What is the specific inference method of topic modeling? To briefly explain, the topic model is a hierarchical Bayesian model. In statistical language, parameters of interest that topic modeling would like to infer are the probabilities of each word in the topics and the probabilities of topics in each document. Based on Bayesian statistical inferences, topic modeling sets prior probability distributions for the parameters and computes the posterior probability distribution of the parameters, conditioned by data (i.e., observed documents). The posterior probability distribution could be written as following formula, in the case of LDA. (θ: topic proportions in each document, ϕ: topic, z: assigning topics to words, w: observed words, α: a parameter of the Dirichlet distribution that is a prior distribution of topic proportions in each document, β: a parameter of the Dirichlet distribution that is a prior distribution of topics) (Blei & Lafferty, 2009; Blei, 2012)

$$p\left(\phi_{1:K},\theta_{1:M},z_{1:M}|w_{1:M},\alpha,\beta\right)=\frac{p(\phi_{1:K},\theta_{1:M},z_{1:M},w_{1:M}\,|\,\alpha,\beta)}{p(w_{1:M}\,|\,\alpha,\beta)}$$

$$p(\phi_{1:K},\theta_{1:M},z_{1:M},w_{1:M}\,|\,\alpha,\beta)=\Pi_{i=1}^{K}p(\phi_i\,|\,\beta)\Pi_{m=1}^{M}p(\theta_m\,|\,\alpha)\left(\Pi_{n=1}^{N_m}p(z_{m,n}\,|\,\theta_m)p(w_{m,n}\,|\,\phi_{1:K},z_{m,n})\right)$$

$$p(w_{1:M}\,|\,\alpha,\beta)=\int_{\phi_{1:K}}\int_{\theta_{1:M}}\Sigma_{Z_{1:M}}p(\phi_{1:K},\theta_{1:M},z_{1:M},w_{1:M}\,|\,\alpha,\beta)$$

The first formula is the posterior probability distribution of the parameters of

interest, conditioned by data. This is a key objective of LDA. The second formula corresponds to the numerator of the first formula. The denominator (the third formula) is obtained by integrating and summing the numerator over θ, ϕ, and z. This is a process of summing the numerator (the second formula) "over every possible instantiation of the hidden topic structure" (Blei 2012, p. 81). However, since the posterior probability distribution is intractable, especially because of the denominator (Blei and Lafferty 2009; Blei 2012), topic modeling utilizes approximation methods like the Gibbs sampling or variational inference. The parameters of interest are estimated approximately.

The topic derived through this process, i.e., the probability distribution of words which are most likely to generate the data, is usually interpreted as a topic or subject that exists within given data. For example, let us assume that the following word probability distribution was derived by topic modeling analysis, using army soldiers' responses to an open questionnaire: [unity - 0.05; friendship - 0.04; comrade - 0.035, trust - 0.03, army spirit - 0.03, ……] (words listed in order of probability). From this word probability distribution, or topic, we can see that the authors of the data frequently used words such as "unity" and "friendship." Additionally, from the cluster of words with high probabilities, one can roughly guess in what context they are being used. In this case, these words emphasize cohesion, and show that cohesion is being imagined as the core of the army spirit.

I argue that the topics rendered by this type of analysis can measure discourse beyond merely a subject, which is a relatively general concept. From the topic we can identify both types of information in which discourse is expressed in language materials: (1) frequently appearing words, and (2) network structures between them. Frequently appearing words are identified from words given high probability by the topic model. In order for a word to be given a high probability, it must appear frequently in the data, even if only in a specific subset of the data. The network structure also can be seen from the topic, although it is incomplete. A set of words given high probability is a key clue. That they were given high probabilities together in one topic means that in many cases the words were often used together in one document.[1] In short, since the topic of topic modeling implies two factors in which

1 To be exact, we cannot be completely sure of the existence of networks between words accorded high probability. Depending on the situation, in theory, it is possible for words not linked to any criterion to be given a high probability of appearing at the same time. As will be explained later, this is where the need to complement Topic Modeling through network analysis and other methods becomes apparent.

discourse is expressed in the language data, topic modeling can be used as an efficient and reproducible method of measuring discourse.

The fact that a topic is a latent variable also makes topic modeling an adequate tool for measuring discourse. Discourse is expressed through language activity, and is not the language activity itself. Rather, it is a potential level of power that regulates language activities. Topics are similar to discourse in that they are also latent variables behind text. Topics are probability distributions underlying actual text, not actual words or languages. Thus, compared with simply summarizing the usage of words in text, topic modeling can be a better measure that reflects the conceptual status of discourse. Of course, the latent level of topics and the potential status of the discourse are not the exact same thing. That is why a topic is not an accurate measure of discourse, but a reference for identifying a discourse.

Multiple topics and discourses

The second reason why topic modeling is a suitable tool for discourse analysis is that topic modeling's assumption of multiple topics is appropriate for analyzing the dynamics of multiple discourses. According to Foucault, multiple discourses exist in any one era and place. They are not only different, but sometimes even contradictory to each other. Although these multiple discourses and nondiscursive elements can be connected to create some comprehensive effect, Foucault assumes the existence of various discourses and their interaction (Foucault 1972, 1990).

Therefore, if a researcher tries to measure discourses from text data, then a method that is capable of capturing multiple discourses is appropriate. Topic modeling assumes there are multiple topics in a set of documents and each document has a distribution of topics. This is a realistic assumption, because one document does not cover only a single matter. For example, in a newspaper article on breast cancer, there could be a story about how to treat cancer, but there could also be a story surrounding the doctor's work ethic. And if, as I explained earlier, the topic can be used as an observation of a discourse, the assumption that there are multiple topics can be utilized to capture multiple discourses that exist in the data.

The assumption of multiple topics inherent in topic modeling is appropriate for capturing less dominant but still distinct topics or discourses. In fact, two traces of the discourse which I mentioned before can be captured in other ways. For example, simply by counting the words which appear, one can see what the most important subject is, and by looking at the word pairs

that appear most frequently together, one can identify the most important connection structure. However, these approaches only identify the most prominent information in the data. Behind this most pronounced trend, other patterns and information may exist that are less significant but distinguishable from the most dominant one. In the case of topic modeling, this pattern is captured thanks to its assumption of multiple topics. Capturing less prominent patterns as separate topics improves the topic model's power to explain a given text, as long as the pattern shows distinct features, even if it takes up a small proportion of the whole text. Therefore, the topic modeling algorithm preferably extracts separate topics in this situation. In short, the multiple topic assumption of topic modeling makes it possible to capture less-dominant discourses that are difficult to capture by other methods.

Furthermore, using topic modeling methods developed after LDA, the interactions between discourses also can be measured, at least in part. The Correlated Topic Model (CTM) is a good example. CTM was developed to estimate correlation between topics in addition to extracting topics (Blei 2012; Blei and Lafferty 2007). When topics appear in a document, they may have a specific relationship with each other. For example, if a certain topic appears in a document, it is more likely that another specific topic will appear in the document. If this relationship between topics can be estimated, the result could be used to estimate one aspect of the interaction between discourses, which Foucault described. For example, if any two topics have a strong positive correlation, and if each topic points to discourses that the researcher pays attention to, the researchers can infer the strong relationship between the two discourses.

In sum, the assumption of the presence of multiple topics in topic modeling and the recent development of analysis techniques that focus on the relationship between topics make topic modeling methods appropriate tools for analyzing discourse.

Discourse change and topic modeling

The third reason why topic modeling is suitable for discourse analysis is that it has the potential to capture changes in discourse over time. Foucault assumed that discourse changed over time. Interactions between discourses, as well as interactions between discourses and non-discursive elements, may lead to the emergence of a new discourse, and may also lead to the loss of the existing discourse. Much of Foucault's work includes shifts in discourse shaped by the course of time. Many people have labeled Foucault's thought as

post-structuralism, because most of the structural variables presented by Foucault change due to external factors such as time (Deleuze 1988; Foucault 1977). Therefore, a method capable of detecting changes or shifts in discourse is necessary for discourse analysis.

A number of topic modeling techniques that emerged after LDA make it possible to track changes in discourse over time. For example, the Dynamic Topic Model (DTM) allows an analyst to capture changes in the composition of topics over time (Blei and Lafferty 2006; Blei 2012). Additionally, the Structural Topic Model (STM) allows analysts to estimate how the metadata of the documents analyzed, including publication time, relate to the proportion or content of extracted topics (Roberts, Stewart and Tingley 2014). That is, by incorporating the publication time of documents into the structural topic model as a covariate, it becomes possible to estimate the weight and content changes of the topics over time. These techniques can help us to study historical changes in discourses.

In particular, STM makes it possible to analyze how discourse interacts with nondiscursive elements, beyond simply considering time, by allowing researchers to take various metadata into account. For example, regarding national health policy, one newspaper may focus on the benefits of the health policy, while another may focus on the financial burden it creates. This means that the different discourses concerning health policy are linked to different newspapers. If STM is used to examine the relationship between the weight and the content of the extracted topics and the publishers of the article, the result could be used for analyzing the relationship between a specific discourse and an organizational variable.

In summary, topic modeling methods can be utilized for discourse analyses and have the potential to complement existing research methods. Because (1) a topic, which is main finding of topic modeling, contains two important information of discourse; (2) the assumption of multiple topics inherent in topic modeling makes it possible to deal with the plural discourses; and (3) recent techniques in topic modeling can detect changes in discourse over time and interactions with non-discursive elements.

Limitations of topic modeling

Topic modeling is not a perfect method and has many limitations. However, the limitations can be remedied and overcome in a variety of ways. The main limitations are as follows. First, topic modeling does not utilize the

information on relationships between words as much as would be ideal. As mentioned earlier, topic modeling considers each document a bag of words. In this process, all information about words' co-appearance in the same sentence or the order of its appearance is ignored. Even though the relationship between words is an important representation of discourse, topic modeling considers only the simultaneous appearance of words in a same document.

Second, there is no theoretical justification for the usage of specific prior distribution and no firm agreement on how to determine the various hyper-parameters that are given for topic modeling (Gerlach et al. 2018). As I mentioned earlier, topic modeling is basically a hierarchical Bayesian model, requiring prior distribution of parameters. For LDA, the Dirichlet distribution is used as a prior distribution. However, there is no special theoretical reason for using the Dirichlet distribution, except for mathematical convenience (Gerlach et al. 2018). Even if analysts accept the use of a specific distribution as a prior distribution, there are still problems. There are several numbers and vectors that analysts need to assign to the model in advance. Analysts should determine hyper-parameters that determine the shape of the Dirichlet distribution in advance (in the case of LDA). The number of topics to be extracted from the data should also be determined. Various methods have been proposed over how to determine these figures, but it is hard to say that the rationale is firm and sound.

Finally, topic modeling does not always perform well. For example, there are many reports that topic modeling does not function properly for analysis of short texts (Alvarez-Melis and Saveski 2016; Hong and Davison 2010). Topic modeling essentially utilizes information about the type and frequency of words that co-appear in a single document. If the length of documents is too short, the volume of the information also decreases, which leads to a poor model. Considering that the various text data provided on the internet generally have short word-counts (e.g., Twitter posts, product reviews), this limitation could prevent us from applying topic modeling to a vast amount of valuable data.

However, these limitations do not undermine the value of topic modeling. There are many ways to improve the model's performance. For example, the problem of poor model quality in short text could be solved by a variety of ways: (1) adopting the assumption that only one topic exists in each document, (2) pooling documents on various criteria to increase the volume of information in one document, or (3) developing new techniques tailored to short text to overcome small amounts of information (Alvarez-Melis and

Saveski 2016; Hong and Davison 2010; Jónsson and Stolee; Qiang, Chen, Wang and Wu 2017; Quan, Kit, Ge and Pan 2015; Steinskog, Therkelsen and Gambäck 2017).

If topic modeling cannot detect network information thoroughly, another adequate method could be used for the job. As noted, topic modeling does not utilize the relationship information between words completely. Semantic network analysis is the first candidate to complement these limitations. By defining language as a network and applying the rich analysis tools of network analysis, the following questions can be answered. (1) What words/concepts are linked to the main words/concepts? (2) Which words are most central when considering the relationship between words? Or which words play a particular role? (Centrality analysis and role analysis) (3) Is there a community of words that are connected particularly cohesive to each other, compared to other words? (Network cluster analysis)

Every method has its limits and shortcomings. Topic modeling is no exception. Topic modeling, however, has important advantages that traditional methods did not have. The limitations that one technique has could be complemented by another method and overcome in a variety of ways. As we saw earlier, we can improve topic modeling or complement it with a number of different kinds of techniques. By applying traditional research methods, topical modeling, and other text mining techniques together, discourse study can be carried out in more rigorous and effective ways.

The linkage between the social sciences and text mining

This study is an attempt to link topic modeling to discourse analysis. I believe that so many concepts in the social sciences and techniques in text mining must be actively connected. As I said before, social scientists can use the great potential of text mining for their research. Furthermore, through this active connection, we can deepen discussions on how text mining techniques can be adequately put to use more broadly in the social sciences.

It is not the social sciences that are leading the various methodical innovations in text mining. Rather, the fields of physics and computer engineering are making outstanding contributions. Therefore, many techniques were not originally designed to measure variables studied in the social sciences. To use these methods to measure variables in the social sciences, it is necessary to consider how the techniques relate to the variables

studied in the social sciences, and which of the techniques are most appropriate to measure such variable, instead of simply applying developed techniques.

Cluster analysis or community detection within semantic networks is a good example. As I said earlier, networks among words are critical to measuring discourse, as well as clarifying the meaning of words themselves. The analysis for finding relatively cohesive communities in a whole network can be very useful. Through this analysis, we can find detailed semantic structures that exist in the text data, which cannot be revealed on the whole network level. For example, the literature on cancer has various focuses (e.g., focus on health policy, focus on new treatments) and the focuses may be expressed by distinct networks of words. It is because the words and their networks in the health policy discussion is quite different from the words and networks stemming from the discussion of treatments. If researchers can identify distinct communities in semantic networks, they could infer detailed semantic structures in the entire data, in reverse. So, what algorithm does the analyst need to apply to identify sub-networks in the semantic network?

According to Fortunato and Hric, the frequently used community detection algorithms can be summarized into three major categories: "methods based on statistical inference," "methods based on optimization," and "methods based on dynamics" (Fortunato 2010; Fortunato and Hric 2016). There has been much discussion about which of the methods is most appropriate and useful. Benchmark tests were repeated many times. Active discussion has identified the advantages and disadvantages of each algorithm. For example, the method based on the optimization of modularity was reported to have limitations in detecting a community with a small number of nodes. This problem is called the "resolution problem" (Fortunato and Hric 2016; Yang, Algesheimer and Tessone 2016).

However, most of the findings are not based on the analysis of semantic networks. In other words, research on methodologies using semantic networks is rare. Therefore, it is not always adequate to apply algorithms that are judged to have excellent performance in benchmarks to semantic network data. This is because semantic networks have distinct characteristics from other networks. Problems related to path distance are a good example. Meanings which occur from a word network are sensitive to network distance. For example, let's assume that networks are defined by words' co-appearance in the same sentences. And let's say there are three sentences: "The monkey's butt is red," "The apple is red," and "The apple is delicious." In this case, "monkey" and "delicious" are connected by a path length of three. If

each sentence appeared very often, the connection between "monkey" and "delicious" would be identified as a fairly solid connection in most of the clustering algorithms. However, it is difficult to say that the "monkey" and "delicious" constitute a valid connection from the perspective of meaning. In short, as path distance increases, the value and meaning of semantic networks could decrease dramatically. In order to detect a community in the semantic network appropriately, techniques that take this into account are necessary.

I believe that, for now, the "method based on dynamics" is suitable for finding sub-communities in semantic networks, even if it does not perform the best on benchmarks. For example, the Walktrap algorithm in this category estimates the similarity between nodes using random walk movements with a specific number of steps and clusters based on similarity of information (Fortunato and Hric 2016). This method has the advantage of being able to overcome the resolution problem and, at the same time, allows the researcher to adjust the number of random walk-steps which are used to calculate the similarity between nodes. This is an important attribute that allows researchers to control and manage problems arising from path distance. For example, although the Walktrap algorithm has been reported to show good performance with the condition of 4 or 5 random walk-steps, the number of steps could be set to 2 or 3 to avoid detecting long-distance connections, which are not that useful in semantic networks. In addition, the results of several benchmarks show that this algorithm does not lag far behind other techniques (Pons and Latapy 2005; Yang et al. 2016).

The above conclusion regarding semantic network and community detection algorithms is my tentative judgment, and more research is needed. However, it is clear that this kind of exploration—the connection of social science concepts to text mining techniques—is necessary. Only then can we maximize the potential of text mining for social science research. As we have seen earlier, text mining offers great potential for various social science studies. In particular, the capability of processing large amounts of language data to produce reproducible results suggests the new possibility of studying language data in a brand new way. However, in order to utilize text mining methods accurately, it is necessary to connect the various techniques of text mining with social science concepts and variables, examine whether their connections are valid, and coordinate and devise methods from the perspective of social science concepts and variables. This study, which links topic modeling with discourse analysis in order to develop new possibilities of discourse analysis, is an example of this kind of attempt.

(*Submitted*: August 3, 2019; *Accepted*: September 15, 2019)

References

Alvarez-Melis, David and Martin Saveski. 2016. Topic Modeling in Twitter: Aggregating Tweets by Conversations. *ICWSM*: 519-522.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4): 77-84. doi:10.1145/2133806.2133826

Blei, David M. and John D. Lafferty. 2006. *Dynamic Topic Models.* Paper presented at the Proceedings of the 23rd international conference on Machine learning.

_____. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1(1): 17-35. doi:10.1214/07-aoas114

_____. 2009. "Topic Models." Pp. 101-124, in *Text Mining: Classification, Clustering, and Applications.* Chapman and Hall/CRC.

Blei, David M, Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of machine Learning research* 3(Jan): 993-1022.

Bollen, Johan, Huina Mao and Xiaojun Zeng. 2011. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2(1): 1-8. doi:10.1016/j.jocs.2010.12.007

Danner, Deborah D, David A. Snowdon and Wallace V. Friesen. 2001. "Positive Emotions in Early Life and Longevity: Findings from the Nun Study." *Journal of Personality and Social Psychology* 80(5): 804-813.

Deleuze, Gilles. 1988. *Foucault*. University of Minnesota Press.

Dickinson, R. 2013. "Weber's Sociology of the Press and Journalism: Continuities in Contemporary Sociologies of Journalists and the Media." *Max Weber Studies* 13(2): 197-215. doi:10.15543/mws/2013/2/5

Dosse, François. 1997. *History of Structuralism: The Rising Sign, 1945-1966* (Vol. 1). U of Minnesota Press.

Evans, James A. and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42(1): 21-50. doi:10.1146/annurev-soc-081715-074206

Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486(3-5): 75-174.

Fortunato, Santo and Darko Hric. 2016. "Community Detection in Networks: A User Guide." *Physics reports* 659: 1-44. doi:10.1016/j.physrep.2016.09.002

Foucault, Michel. 1971. "Orders of Discourse." *Social science information* 10(2): 7-30.

_____. 1972. *The Archaeology of Knowledge*. London: Tavistock Publications.

_____. 1977. *Discipline and Punish: The Birth of the Prison.* Vintage.

_____. 1990. *The History of Sexuality: An Introduction, volume I* (R. Hurley, Trans.).

New York: Vintage.

_____. 2002. *The Order of Things: An Archaeology of the Human Sciences.* Psychology Press.

_____. 2013. *History of Madness.* Routledge.

Gerlach, Martin, Tiago P. Peixoto and Eduardo G. Altmann. 2018. "A Network Approach to Topic Models." *Science advances* 4(7): eaaq1360.

Ham, Yugeun and Seungbyeong Chae. 2012. (in Korean) *Big data, gyungyoungeul bagguda [Big data changes management].* Samsung Economic Research Institute.

Hong, Liangjie and Brian D. Davison. 2010. *Empirical Study of Topic Modeling in Twitter.* Paper presented at the Proceedings of the first workshop on social media analytics.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning* (Vol. 6). Springer.

Jónsson, Elıas and Jake Stolee. *An Evaluation of Topic Modelling Techniques for Twitter.*

Kozlowski, Austin C., Matt Taddy and James A. Evans. 2018. The Geometry of Culture: Analyzing Meaning through Word Embeddings. *ArXiv e-prints.* https://ui.adsabs.harvard.edu/#abs/2018arXiv180309288K

Kwartler, Ted. 2017. *Text Mining in Practice with R.* John Wiley & Sons.

Kwon, Sejeong, Meeyoung Cha and Kyomin Jung. 2017. "Rumor Detection over Varying Time Windows." *PLoS One* 12(1). doi:10.1371/journal.pone.0168344

Lukes, Steven. 2005. *Power: A Radical View (2nd).* Hampshire: Palgrave Macmillan.

Pons, Pascal and Matthieu Latapy. 2005. *Computing communities in large networks using random walks.* Paper presented at the International symposium on computer and information sciences.

Qiang, Jipeng, Ping Chen, Tong Wang and Xindong Wung. 2017. *Topic modeling over short texts by incorporating word embeddings.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

Quan, Xiaojun, Chunyu Kit, Yong Ge and Pan Sinno Jialin. 2015. *Short and Sparse Text Topic Modeling via Self-Aggregation.* Paper presented at the IJCAI.

Roberts, Margaret E, Brandon M. Stewart and Dustin Tingley. 2014. "stm: R package for structural topic models." *Journal of Statistical Software* 10(2): 1-40.

Slater, Michael D, Marilee Long, Erwin P. Bettinghaus and Jason B. Reineke. 2008. "News Coverage of Cancer in the United States: A National Sample of Newspapers, Television, and Magazines." *Journal of health communication* 13(6): 523-537.

Steinskog, Asbjørn, Jonas Therkelsen and Björn Gambäck. 2017. *Twitter Topic Modeling by Tweet Aggregation.* Paper presented at the Proceedings of the 21st Nordic Conference on Computational Linguistics.

Viera, Anthony J. and Joanne M. Garrett. 2005. "Understanding Interobserver Agreement: The Kappa Statistic." *Family Medicine* 37(5): 360-363.

Yang, Z., R. Algesheimer and C. J. Tessone. 2016. "A Comparative Analysis of

Community Detection Algorithms on Artificial Networks." *Sci Rep, 6*, 30750. doi:10.1038/srep30750

WONKWANG JO is a lecturer in the department of sociology at Seoul National University. He was a visiting researcher in The Institute for Social Development and Policy Research, Seoul National University. He received his Ph.D. in Sociology from Seoul National University. *Address*: Department of sociology, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea. [*E-mail*: wonkwangjo@gmail.com]