10-2010

Collecting Legacy Corpora from Social Science Research for Text Mining Evaluation

Bei Yu
Syracuse University, byu@syr.edu

Min-chun Ku
Syracuse University, mky@syr.edu

Collecting legacy corpora from social science research for text mining evaluation

Bei Yu
School of Information Studies
Syracuse University
byu@syr.edu

Min-chun Ku
School of Information Studies
Syracuse University
mku@syr.edu

ABSTRACT

In this poster we describe a pilot study of searching social science literature for legacy corpora to evaluate text mining algorithms. The new emerging field of computational social science demands large amount of social science data to train and evaluate computational models. We argue that the legacy corpora that were annotated by social science researchers through traditional Qualitative Data Analysis (QDA) are ideal data sets to evaluate text mining methods, such as text categorization and clustering. As a pilot study, we searched articles that involve content analysis and discourse analysis in leading communication journals, and then contacted the authors regarding the availability of the annotated texts. Regretfully, nearly all of the corpora that we found were not adequately maintained, and many were no longer available, even though they were less than ten years old. This situation calls for more effort to better maintain and use legacy social science data for future computational social science research purpose.

Keywords

Computational social science, evaluation, corpora, text categorization, topic clustering

INTRODUCTION

The new emerging field of Computational Social Science aims to use computational models to analyze large amount of data to "reveal patterns of individual and group behaviors" (Lazer, *et al.*, 2009). A subarea in computational social science is to use machine learning and natural language processing techniques to automatically analyze large amount of text, especially user-generated content on the Web, in order to understand the topics, perspectives, mood, personalities, and many other aspects that humans manifest in language.

The advances of text mining techniques provide potentially powerful tools for automatic annotation of large text corpora (Cardie and Wilkerson, 2008). However, these tools were not initially designed for social science research purpose and thus have not been extensively evaluated in corresponding tasks. The lack of empirical knowledge on the tools' reliability, validity, and best practice poses great challenges for non-expert users. One big obstacle in extrinsic evaluation is the lack of benchmark data sets. Creating these data sets costs high and the effort has not been very rewarding for both computer science and social science researchers. At the same time social scientists constantly use Qualitative Data Analysis (QDA) methods to manually annotate small text corpora with reliability check. We argue that these high-quality legacy corpora are ideal data sets to evaluate text mining methods, such as text categorization and clustering.

Text categorization is a family of supervised learning techniques that automatically assign texts to pre-defined categories. Originally developed for information organization purpose, text categorization techniques are now used to categorize emotions, opinions, ages, genders and other human characteristics with various levels of success. More empirical evidence is needed to guide novice user to choose appropriate tools (Yu, 2008).

The unsupervised text clustering techniques, such as LDA and pLSI, automatically estimate the main themes in a large corpus and the themes each document involves. A theme is usually represented by a weighted list of words that may or may not directly make sense to humans. Also due to the numerous ways to tune the parameters, evaluating the validity of text clustering result has been a persistent problem for both experts and users (Chang, 2009).

Traditional QDA, such as content analysis and discourse analysis, shares similar goals with text categorization and clustering in that they all aim to annotate texts based on various properties. For example, political scientists sometimes examine the valence of news articles as positive, negative, or neutral toward a presidential candidate. This is exactly an application of sentiment classification, a kind of text categorization. QDA has produced numerous manual annotations with high reliability scores, which are considered "gold standard" in text mining evaluation. Conversely, evaluating text mining algorithms in new tasks

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

improves our understanding of the reliability and validity of the techniques.

This pilot study aims to explore the possibility of collecting these "gold-standard" corpora by searching social science literature and directly requesting data sets from authors. If this proves to be a viable approach, we may expect to automate this process and build a large repository of benchmark corpora with minimal effort. In this pilot study, we first searched articles that involve content analysis and discourse analysis in leading communication journals, and then contacted the authors regarding the availability of the annotated texts. We present the two steps in the next two sections, followed by conclusion.

SELECTION OF CANDIDATE SOCIAL SCIENCE DATA

In this pilot study we choose the communication field with specific focus on health communication and computer-mediated communication in that one of their foci is user-generated content on the Web, which is often publicly accessible. We consulted the *Journal Citation Reports* of *ISI Web of Knowledge* to select leading journals based on their impact factors. We selected four journals: *Journal of Communication* (current impact factor 2.266), *Journal of Health Communication* (2.057), *Journal of Computer-Mediated Communication* (1.901), and *Discourse Studies* (1.116).

We use queries "content analysis" and "text analysis" to search articles from 1995 to present in hope of finding existing use cases of text mining applications because this is the period when text mining techniques enjoy great advances. Some social scientists might have attempted to adapt new techniques in their research. Table 1 shows the number of articles retrieved from the four journals.

Journal	"content analysis"	"text analysis"
JoC	85	13
JoHC	55	13
JCMC	58	13
DS	12	33

Table 1. The number of retrieved articles.

We then narrowed down the list after carefully reviewing the data collection and processing sections in each article, including the unit of analysis (word, phrase, paragraph, or document), the coding scheme, and the inter-coder reliability measure. We excluded review articles and others without clear descriptions of the above details. To focus on textual data and English language, we also excluded the articles that dealt with non-English data, multimedia data and pictorial data without transcripts. Since text categorization and clustering are often conducted at document level, we included studies with document-level coding only.

At the end of the selection process we identified ten candidate corpora, which include news articles, blog posts, user comments, emails, newsgroup discussions, and personal letters. The smallest data set consists of 46 documents, the largest one 2316 documents, and the others ranging from 200 to over 1,000 documents. Surprisingly all of these studies were published between 2006 and 2010, with nine from US and one from UK.

DATA REQUEST

We emailed the ten first authors to express our interests in using their datasets to evaluate text mining algorithms. Nine authors replied with various responses. Only one author (from U.K.) attached the data set directly with the reply. One author said the data were private. One author said the data were not ready to share. One author pointed us to the principle investigator of the parent project. The rest five authors said they no longer kept the copies of data, but some kept their codes. We then sent the second round of emails to request information to reconstruct the data sets ourselves, such as the URLs to web pages, the titles of news articles to be retrieved from Lexis-Nexis, etc., only to find these metadata were no longer maintained either.

CONCLUSION

Regretfully, collecting legacy corpora from previous social science research does not seem a viable approach because nearly all of the corpora we found were not adequately maintained. Half of them are no longer available, even though they were created less than 10 years ago. A number of reasons might contribute to this situation. First, social scientists may not have adequate resources to deposit and maintain the corpora. Second, social scientists might not be aware of the data's value to computer and information scientists in new inter-disciplinary research. This situation calls for more effort to better maintenance and use of legacy social science data for future computational social science research purpose.

REFERENCES

Cardie, C. and Wilkerson, J. (2008). Text annotation for political science research. *Journal of Information Technology and Politics, 5*(1), 1-6.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D. (2009). Reading tea leaves: how humans interpret topic models. *NIPS, 22*, 288-296.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A-L., Brewer, D., Christakis, D., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy D., Alstyne, M. (2009). Computational social science. *Science, 323*, 721.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing, 23*(3), 327-343.