Detection of Personal Experiences or Storytelling

by Zequn Che

Submission date: 08-May-2020 10:09PM (UTC-0400)
Submission ID: 1320003850
File name: Detection_of_Personal_Experiences_or_Storytelling.txt (22.02K)
Word count: 3332
Character count: 19109

Detection of Personal Experiences or Storytelling

Group 6

Hong Zhu1, Guoxing Yao2, Wanyue Xiao1, Zequn Che1

School of Information Studies1 & School of Engineering2, Syracuse University, Syracuse, NY 13244 USA

Abstract

Personal experience sharing, or storytelling, is one of social strategies used commonly in daily conversation for serving the purposes of intimate relationship construction, strengthening persuasion level, or emotion sharing. Different from formal evidence in academic research, personal experiences are depicted in daily language and become more and more popular and persuasive as social media is rampant. Compared with self-disclosure, the importance of personal experience, however, is seldomly investigated by researchers in the field of psychology and linguistics. Therefore, to gain more valuable understanding, this paper will investigate the effect of personal experience sharing on the online collaborative community. Based on the quantitative analysis of Tagged Reddit Submission Dataset, this study will find the appropriate classification method to predict the existence of personal experience and summarise potential features that are crucial for the detection result.

Introduction

Ensuring the accuracy, timeliness, and worthiness of online resources has always been an intractable issue since people who are able to access the internet are capable

of leaving comments, editing online recourse repositories, and sharing personal opinions. Due to the convenience brought by the powerful Information and Communication Technology (ICT) and affluence of online information, the web-based collaborative community, a learning atmosphere or context from which people could gain lessons through a sustainable learning processes (Jobring, 2002), enable human beings to find the desirable or precise answer no matter how nebulous the questions are. Especially in some popular online learning communities, such as Reddit and StackOverFlow, knowledge sharing is incredibly frequent since every day thousands and hundreds of topics will be posted or responded by other users who share the common interests. Such an interactive online learning method is being reckcon as offering more efficiency than self-study and self-reflection (Carlen & Jobring, 2005). Hence, it is inevitable that humans, in particular for young people, are more relying on these online collaborative communities instead of traditional, face-to-face communication (de Moor, 2013). Under such a circumstance, gaining accurate knowledge is more important than ever since the low reliability of online information affects both the quality of people's daily works as well as their professional researchers.

One approach examining the information authenticity is to check if the editor shows some common ground, opinions, or evidence in the corresponding knowledge field (Clark & Sampson 2007). Such an approach is as known as rationel which is a comprehensive process containing "reasoning of the individual designer and the discourse among participants in a design project" (Shipman et al., 1997). Rationalizing

is beneficial to people's communication, collaboration, and learning experience through sharing contextual information and achieving common ground (Xiao 2011; Xiao 2014; de Moor, 2013) given that people who are fully conscious of the shared content are more likely to build intimacy and trustiness during the process of relationship construction. Additionally, studies have shown that online information involving reasoning, expertise sharing, and previous experience referral is perceived as professional from the academic perspective (McGrew et al., 2018). Since experience sharing enlightens people for extracting the similarity and patterns of previous cases (Linn, 1982), justification which embedded experience from previous cases is more reliable and easy to comprehend. Therefore, the detection of experience referral could be a key indicator when it comes to the assessment and evaluation of online resources, especially in the online discussion community.

Experience is a sensory awareness arising by existing events from which human beings could extract empirical understanding. Specifically, a sequence of events have happened or are happening to the creatures in the universe, including humans, animals, plants, could be summarised as experience. Therefore, experience is nothing but a complex and constructed reality (Fox, 2008). The method of sharing or expressing previous experience is called storytelling. According to the Narrative Paradigm proposed by Fisher in 1985, as the fundamental element of human expression and communication, logos (or words) could be formed in a logical, persuasive, and argumentative way and then transformed to a reflection of knowledge, truth, and reality (1985). More importantly, this paradigm mentioned that a good story

could be more convincing than a good argument (Fisher, 1985). A well structured, logical, and comprehensive storytelling process provides extraordinary compelling reasons to people who are seeking constructive suggestions and answers. For instance, people tend to use their previous experience to strengthen the reliability and then to achieve the goal of statement justification in daily conversations or some special circumstances requiring techniques of reasoning. Hence, the techniques of mastering personal experiences demonstration or storytelling are crucial and practical for reasoning. However, this conclusion is based on the assumption that personal experience shared in the public is rational instead of being subjective and partially understanded.

Compared with personal experience, the topics related with self-disclosure have been widely discussed. Hence, it is worthwhile to mention the difference between personal experience and self-disclosure since the difference could be confusing and obscure for people who do not understand those terms thoroughly. Personal experience is narration which reveles a previous story or a series of events from which people could obscure knowledge related to the present situation (Akinsanya & Bach, 2014) while self-disclosure is a social phenomena or a human behaviour which disclose private information to others (Ravichander et al., 2018). Firstly, the content is different. Personal experience focuses on actual events that happened previously while content of self-disclosure could also contain emotion sharing. Secondly, People revealed their own experience to enhance the information reliability while in the area of self-

disclosure, people revealed personal information to enhance interpersonal relationships and improve social support (Bak et al., 2014; Balani & Choudhury, 2015).

This paper is organized as below. Section 2 will describe background knowledge of NLP. Section 3 will summarize the dataset used and models for machine learning and statistics analysis. Section 4 will evaluate the performance of models established in section 3 and make a general conclusion.

Research Background

NLP enables computers to undertake a series of text analysis related assignments, ranging from the basic sentence-parsing to sophisticated language translation and error detection. Before introducing the model design section, we first will explain several NLP Basic Concepts.

Basic Concepts

1. Avcailable Packages

NLTK (The Natural Language Toolkit) is a popular Python package which is used for NLP in Python. The nltk package contains a lot of corpus. It is helpful in NLP analysis tasks. Another classic NLP text processing package is Stanford CoreNLP. As a software which provides many technology tools of human languages based on Java, Stanford CoreNLP could be used on detecting past tense of sentences with relatively higher accuracy. In the following analysis, we will use this package in python.

2. Corpus

In linguistics, corpus is a large and structured set of texts. It is used to do statistical analysis. In NLP, corpus is a collection of texts which helps people for analysing the documents from individual word level.

3. Tokenization

Tokenization is a task of cutting a character into pieces and throwing away the certain characters, like punctuation. In other words, tokenization is to disassemble the long sentence into small meaning parts, such as short sentences and words. It is very useful to deal with word frequencies, which we used in analyzing personal experience.

4. POS tagging

Part-of-speech tagging, which is also called POS tagging, is a powerful tool to mark up the words by their characteristics and to indicate its syntactic role (such as noun, adverb, plural, etc.) (Collobert et al., 2011). We used Part-of-speech tagging to get the past tense sentences. The underlying reason is that when the user posted the comments under the topic, the experience mentioned in the comments had already happened. After scanning the datasets, one could observe that the common expressions in their sentences, for instance, is “I did something or I have done something”. Specifically, those expressions could be “I got”, “I hosted”, “I have conducted”, “I’ve searched”, and “I had finished”. Some special cases like “My father and I have discussed” also occurred. Since Stanford CoreNLP does not offer a good solution for subjunctive mood, we switch to the past tense, that is, Stanford CoreNLP marks the sentences to have personal experience as long as there is a combination of “I/We/My/Our” and verb in past tense.

5. TF-IDF

TF-IDF, the abbreviation of term frequency–inverse document frequency, is a tool to show how important a word is to a collection of texts. It is usually used as a weighted factor.

Model Construction Approaches

1. Statistical NLP Methods

In the early stage of the 20th century, the only approach to design a Natural Language Processing system was to manually construct a rule-based model which consists of semantic components and language grammars (Marcus, 1995). Under such a circumstance, it would be difficult for logistics to summarise the pattern that occurs in a large text dataset. Being pushed by the desire to address the problem, mathematics and statistics have been utilized in the tasks of text processing (Manning et al., 1999; Marcus, 1995). Compared with current state-of-the-art techniques, a set of statistical techniques are still being widely and frequently employed by NLP researchers in real-world problems. However, it does not imply that rule-based systems will be forgotten by the public. On the contrary, a rule-based system works better than a statistical model in a small-size dataset (Santaholma, 2007; Saracevic & Dalbello 2001).

2. Machine Learning

The mighty computational text processing ability enables machine learning to become the most prevailing model construction approach in the context of Natural Language Processing (Olsson, 2009; Beggelman & Smychkovich 2009). There are two types of machine learning method, which are supervised machine learning employing fully preprocessed dataset and unsupervised machine learning involving model

construction without annotation or pre-tagging. Some typical supervised NLP machine learning algorithms are Support Vector Machine, Hidden Markov Models, Conditional Random Field, and Maximum Entropy (Marquez & Salgado, 2000). In this research, Guassin Naive Bayes, a tool of the Naive Bayes Family, will be considered.

3. Deep Learning

Deep learning, which is a subset of advanced machine learning in the area of Artificial Intelligence, utilizes multiple processing layers to manage unsupervised unstructured or unlabeled data and produce results from the output layer. Recently, deep learning made its name for the impressive performance of data processing and analysis. Consequently, increasing attention has been given to Deep Learning and more NLP researchers show willingness to use this method to tackle the most challenging and complicated NLP problems, including text classification, language modeling construction, and machine translation. Given that deep learning is a continuing growing field of study, a variety of optimization models are available. Generally, there are 5 major methods which deserve future exploration in NLP. Those methods are Embedding Layers, Multiplayer Perceptrons (MLP), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Recursive Neural Networks (ReNNs) (Young et al., 2018). Deep learning, however, will not be considered in this research project since building a deep learning model on a small dataset is highly possible to cause overfitting issues (Socher et al, 2012).

Research Methodology

Based on the dataset available, deep learning or neural networks are out of our focus since their performances would be passable if the dataset are not scaled. We perform statistics and machine learning classification tasks. Naive Bayes classifier is chosen with its efficiency.

Figure X: Flowchart of tasks in our project

Dataset Description

The dataset is from a research paper, 'The Art of Justifying in Social Media: Insights from Reddit "Change My View" Submissions', JASIS 2020 under review, with Submission column and Personal Experience column. There are 330 submitted comments retrieved from Reddit Discussion Posts and corresponding personal experience marked manually to indicate whether there is personal experience, valued 1 if there is and 0 otherwise. When we observe all data records one by one, we find that personal experience is sentence level based since most personal experience description is constrained in a few successive sentences. On the other hand, there are many anomalous data, for instance, many http url data exist. There are 8468 sentences after removing ones with invalid data.

Models Design Process

1.Word frequency analysis

First, we perform word frequency analysis to get the general image of the dataset. We analyze the frequency of the top 100 words, removing the stop words and punctuation

marks. Figure 1, 2 and 3 shows the top 100 words from all datasets, non personal experience data and personal experience data, respectively.

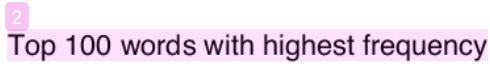

Figure 1 Top 100 words with highest frequency from dataset as a whole

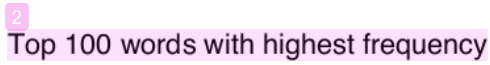

Figure 2 Top 100 words with highest frequency from dataset without personal experience

Figure 3: Top 100 words with highest frequency from dataset with personal experience

In this section, we also attempt to identify whether some words uniquely exist in personal experience, but most words are commonly used and overlapping on each of them..

2. Classification by Naive Bayes Classifier

As talked in Section 2, we utilize the Naive Bayes classifier to predict the words with personal experience. To begin with, we remove the invalid data including url, stop words, punctuation marks, to generate the feature sets. Then, words from the dataset are vectorized in binary. 10% of the dataset are taken as test cases while 90% as the training set. We employ F1, Recall and Precision to evaluate the classifier performance

	TF-IDF	Word Bag(Count of Vector)
F1	0.5546	0.5546

Accuracy	0.5859	0.5859
Recall	0.7857	0.7857

Table 1: F1, Accuracy and Recall values from Naive Bayes classifier

Step 3: Past Tense Detection

In addition to Naive Bayes classifier, we also perform statistical analysis by comparing the manual marked values with automatic marked values. In this step, the Stanford CoreNLP library is utilized to detect past tense sentences. Stanford CoreNLP can parse a sentence and tag each part by displaying a tree, for instance, in Figure 4. In our project, we use regular expressions to match the sentences with VBD and PRP words with the regular expression.

```
pattern = re.compile(r'(\(PRP i\).*\(VP \(VBD .*)|(\(PRP we\).*\(VP \(VBD .*\))|(\(PRP\$ my\).*\(VP \(VBD .*)|(\(PRP\$ our\).*\(VP \(VBD .*)', re.IGNORECASE)
```

If the pattern matches, we label the personal experience with 1, otherwise 0.

Figure 4: An example of Stanford CoreNLP library to parse a sentence

Finally, we compared the personal experience marked by the Stanford CoreNLP library and manually. It shows the Stanford CoreNLP library label the personal experience at accuracy of 72.155% in accord with the manual marked label.

Result Evaluation and Limitations

In the Naive-Bayes classification model, we get a not bad result with pretty high F1 values. There is risk during the process that the training set is converted by tf-idf(term frequency–inverse document frequency) and Bag of Words(count vector) because both training sets are over sparse with widespread 0's with a tiny proportion of non zero. After carefully checking and comparing the data sets, we verify that we get the almost the same F1 values, Accuracy value, and Recall values.

Secondly, in the method of past tense detection with the Stanford CoreNLP library, we get better results. It shows the tool we built has good practicability on detecting past tense which is for detecting personal experience in natural language. In brief, the personal experience detection tool we built works well on detecting personal experience texts based on the method of finding the texts by using past tense detection. But there is a risk that the Stanford CoreNLP library can not parse all sentences since it does throw a parse exception. Also, our past tense detection is based on the regular expression. It is risky if the interval between I/We/My/Our is big, it is possible that the verb in past tense is not connected to I/We/My/Our but other parts of the sentence.

Conclusion

In this paper, we developed a program to detect the personal experience or storytelling sentences. For better understanding the definition of personal experience, we studied from a lot of previous papers. Based on the data we have, we decided to judge if the

sentence is a personal experience by different methods. In our actual operations, count vectors and TF-IDF are not good for our data. Detecting the personal experience by past tense is a better method which has much higher accuracy in our operations. In conclusion, detecting the personal experience texts by detecting past tense sentences is a feasible and reliable measure.

Reference List

1. Akinsanya, A., & Bach, C. (2014). Narrative analysis: The personal experience narrative approach. In ASEE 2014 Zone I Conference.

2. Bak, J., Lin, C. Y., & Oh, A. (2014). Self-disclosure topic model for classifying and analyzing Twitter conversations. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1986-1996.

3. Balani, S., & De Choudhury, M. (2015). Detecting and characterizing mental health related self-disclosure in social media. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp.1373-1378.

4. Beggelman, M., & Smychkovich, Y. (2009). Machine Learning Systems and Methods for Improved Natural Language Processing. U.S. Patent Application 12/264, pp.668.

5. Clark, D. & Sampson, V. (2007), Personally- Seeded Discussions to Scaffold Online Argumentation, International Journal of Science Education, 29:3, pp.253-277.

6. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12(Aug), pp.2493-2537.

7. Carlen, U., & Jobring, O. (2005). The rationale of online learning communities. International journal of web based communities, 1(3), 272-295.

8. de Moor, A. (2013). Creativity meets rationale: Collaboration patterns for social innovation. In Creativity and Rationale, pp. 377-404.

9. Fox, K. (2008). Rethinking experience: What do we mean by this word "experience"?. Journal of Experiential Education, 31(1), pp.36-54.

10. Fisher, W. R. (1985). The narrative paradigm: In the beginning. Journal of communication, 35(4), pp.74-89.

11. Jobring, O. (2002), The Research Project of Online Learning Communities English Homepage, Göteborg, Sweden, available online: http://www.learnloop.org/olc/eng.htm [7th July 2004].

12. Linn, M. C. (1982). Theoretical and practical significance of formal reasoning. Journal of research in Science Teaching, 19(9), pp.727-742.

13. McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. Theory & Research in Social Education, 46(2), pp.165-193

14. Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

15. Marquez, L., & Salgado, J. G. (2000). Machine learning and natural language processing.

16. Marcus, M. (1995). New trends in natural language processing: statistical natural language processing. Proceedings of the National Academy of Sciences, 92(22), pp.10052-10059.

17. Shipman, F. M., & McCall, R. J. (1997). Integrating different perspectives on design rationale: Supporting the emergence of design rationale from design communication. AI EDAM, 11(2), 141-154.

18. Socher, R., Bengio, Y., & Manning, C. D. (2012). Deep learning for NLP (without magic). In Tutorial Abstracts of ACL 2012, pp.5-5.

19. Santaholma, M. E. (2007). Grammar sharing techniques for rule-based multilingual NLP systems. In Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA).

20. Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing.

21. Ravichander, A., & Black, A. W. (2018). An empirical study of self-disclosure in spoken dialogue systems. In Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, pp. 253-263.

22. Saracevic, T., & Dalbello, M. (2001). In: Proceedings of the American Society for Information Science and Technology,(2001), vol. 38, pp. 209-223. A survey of digital library education. Proceedings of the American society for information science and technology, 38, pp.209-223.

23. Xiao, L. (2011). A shared rationale space for supporting knowledge awareness in collaborative learning activities: An empirical study. In 2011 IEEE 11th International Conference on Advanced Learning Technologies, pp.71-75.

24. Xiao, L. (2014). Effects of rationale awareness in online ideation crowdsourcing tasks. Journal of the Association for Information Science and Technology, 65(8), pp.1707-1720.

25. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), pp.55-75.

Detection of Personal Experiences or Storytelling

ORIGINALITY REPORT

2%	2%	1%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	machinelearningmastery.com Internet Source	1%
2	Liu Liu, Bin Li, Lijun Bu, Tian-tian Zhang, Xiaohe Chen. "Chapter 17 Automatic Acquisition of Chinese Words' Property of Times", Springer Science and Business Media LLC, 2013 Publication	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		