Computational Analysis of Collective Intelligence – Towards Automatic Detection of Rationales in Online Deliberations

TARANEH KHAZAEI, University of Western Ontario
LU XIAO, University of Western Ontario

1. INTRODUCTION

Web-enabled collective intelligence platforms allow crowds to gather and contribute in solving the problems that may be difficult or impossible to solve by even the smartest individuals and fastest machines. Various analytical techniques are developed to understand how intelligence emerges from within social interactions and to determine various factors that may influence the collective intelligence phenomenon. We conducted a systematic review of the prior studies to examine the state-of-the-art computational approaches that study and support collective intelligence in large-scale user-contributed text of discourse. Our analysis suggests that the majority of the previous works have been focused on the environments that are primarily designed to facilitate social interactions. However, little effort has been made to understand task-oriented online platforms such as deliberation tools and idea management systems. A new underexplored direction is the development of computational techniques to study these task-oriented environments by detecting and analyzing the "traces" of collective intelligence. It is expected that there are various factors that influence such "traces", including the characteristics of the participants, the kinds of intelligence tasks, and the design of the environments. We also assume that participants' rationales exchanged in these environments, also referred to as justifications in the argumentation, are among the kinds of "traces" that reflect the collective intelligence of the deliberation crowd.

In this paper, we report our effort of identifying rationales in written discourse through detecting rhetorical relations. Rhetorical relations are paratactic or hypotactic relations that hold between spans of text, explaining the construction of coherence in discourse [Tabaoda 2006]. For years, researchers have been focused on building robust theoretical foundations for rhetorical relations. However, automatic identification of such relations among text spans has remained a difficult and a challenging task. Prior studies suggest that three rhetorical relations are commonly present in the rationales: CIRCUMSTANCE, EVALUATION, and ELABORATION [Xiao 2013]. As a starting point, we study lexical cues that signal these three relations. In general, various kinds of cues can signal the existence of a relation, including lexical cues, mood, modality, and intonation [Tabaoda 2006]. We aim to gain insight into the ability of lexical cues in the automatic detection of these rhetorical relations. We also hope to understand whether the effectiveness of cue-based approaches varies based on the nature of the relation and the underlying text genre.

We provide initial results for the following tasks: (1) understanding the features of corpus-based cues (2) exploring the value of such lexical cues in the detection of different rhetorical relations and (3) analysis of the potential differences and similarities among the cues extracted from two different corpora that belong to two different text genres. The experiments are conducted on two human-annotated corpora in the form of (RST): the RST corpus [Carlson et al. 2001] and the SFU review dataset [Taboada et al. 2006]. The altered version of TF-IDF proposed in [Biran & Rambow 2011] is used to process the two corpora and to extract a set of n-grams as potential lexical cues. By analyzing such corpus-based cues, we move beyond the fixed cue lists that are normally used in the prior

literature. Such cue sets are commonly bound to a few well-defined syntactic categories and are referred to as discourse markers. Moreover, to understand the potential value of using lexical cues in the identification of rhetorical relations, the list of top cues extracted from each corpus is applied to the other corpus to extract the instances of the corresponding relation. The classification measures of precision, recall, and F are then calculated and analyzed.

2. METHODOLOGY

2.1. Underlying Corpora

We used two human-annotated corpora as our underlying datasets for the experiments: the RST corpus and the SFU review dataset. Both corpora are annotated in the RST framework and are constructed using the RSTTool (http://www.wagsoft.com/RSTTool). The RST corpus, which has been made available by the Linguistic Data Consortium over the years, includes 385 Wall Street Journal articles and covers more than 178,000 words.

The SFU review corpus is a collection of 400 review documents from movie, book, and consumer products. This dataset contains over 303,000 words and was collected in 2004 from the Epinions Web site (http://www.epinions.com/). To ease the process of genre-specific comparisons and analysis in the rest of the manuscript, we refer to the RST corpus as the news corpus and the SFU corpus will be referred to as the review corpus.

2.2. Lexical Cue Extraction

In order to extract the cues for each corpus, the approach proposed in [Biran and Rambow 2011] is followed. We first collect all the text spans that are linked with the relation of focus in the underlying corpus and form a relation document containing all the relation instances. Similarly, we create non-relation documents, containing all the text spans that participate in any relation except for the relation of focus. Forming these two document sets allows us to use each corpus either as a training set to extract the cues, or a test set to analyze and evaluate the use of cues in the relation extraction process.

For each relation document, we extract all the n-grams (up to tri-grams) that appear in the document and calculate an altered version of TF-IDF for each of the n-grams. The IDF measure is still calculated based on the number of documents that contain the n-gram and the total number of documents in the corpus. However, using the regular TF measure could result in a bias in favour of the n-grams that appear more than once in a relation instance. Therefore, the TF metric is calculated based on the number of relation instances that contain at least one instance of the n-gram. The extracted n-grams are then sorted based on the altered measure of TF-IDF in a descending order. This list is then filtered not to include any pronouns. Modal and auxiliary verbs are also excluded from the lists.

The sorted list of lexical cues extracted from each corpus is applied to the other corpus to extract the corresponding rhetorical relations. The measures of precision, recall, and F are then calculated for each of the cues independently. Based on the experiment results, it could be seen that for all of the relations and for both datasets, those cues ranked after 120 had zero performances. Therefore, only the first 120 cues ordered by TF-IDF are involved in our analysis process. Samples of the cues for the CIRCUMSTANCE relation are *when, now, since, while, until,* and *once.* For the EVALUATION relation, the example cues are *good, high, well, nice,* and *impressed.* For ELABORATION, sample cues are *who, which, where,* and *as if.*

3. EXPERIMENT RESULTS

To ease the process of analysis, we created various visual representations of the experimental results. The analysis of the visual encodings of the TF-IDF measure revealed that TF-IDF is consistently

lower for the lexical cues extracted from the reviews. This finding can be attributed to the fact that news text is a well-structured formal writing, whereas online reviews are relatively less structured and informal. It was also discovered that the ELABORATION cues have a relatively lower TF-IDF for both datasets, indicating that regardless of the genre, ELABORATION may not be well signaled by lexical cues. This metric is higher for both CIRCUMSTANCE and EVALUATION; however, the difference between the cues extracted from the two corpora is more considerable for the EVALUATION relation. These results may indicate that lexical cues might be more genre-specific for EVALUATION.

The precision results for CIRCUMSTANCE is consistent with the TF-IDF finding since lexical cues extracted from the news set have a relatively high precision score. As our visualizations revealed, the precision for the EVALUATION drops considerably after the first couple of cues, confirming that EVALUATION cues can be very specific to the underlying genre. The results for the ELABORATION, however, are not consistent with the TF-IDF metric as the precision is relatively high and is considerably higher for the lexical cues extracted from the review collection. This unexpected result can be attributed to the high percentage of ELABORATION instances in the news dataset (more than one-third of the corpus is annotated with ELABORATION). The recall metric has a very similar trend across all the relations and datasets. For every cue, the instances classified as false negative consist of those explicit relations that are signaled by any lexical cue other than the cue of focus as well as implicit instances that are not signaled at all. A relatively large proportion of relations are normally implicit, which can cause the number of true positives and false negatives signaled by other cues to have little influence on the recall results. Therefore, it can be concluded that implicit relations mainly characterize the recall measure.

Finally, the results from the calculation of F-score, as the most reliable performance metric, showed that except for the ELABORATION, the majority of the lexical cues extracted from the news dataset are performing better in the extraction of rhetorical relations. This better performance for these two relations is consistent with the TF-IDF results. As well, a large proportion of the news dataset being labeled as ELABORATION can explain the better performance of the extracted cues from the review set on the news corpus. Comparison of the results for the three relations indicates that the top CIRCUMSTANCE cues have a better performance compared to the top cues from the other two relations, while the difference is less considerable between EVALUATION and ELABORATION. In essence, the analysis results indicate that the cue-based approaches can be quite effective in detecting CIRCUMSTANCE. However, the ability of lexical cues in relation identification is limited for ELABORATION. For the EVALUATION relation, genre-specific factors can play a more significant role.

4. CONCLUSION

As the first approach toward the identification of rationales in online corpora, we attempted to detect specific rhetorical relations. A rationale is an explanation of the reasons underlying decisions, conclusions, and interpretations. Prior studies on rationale articulation and sharing suggest that it contributes to quality control, knowledge management, and knowledge reuse [Xiao 2014]. However, there have only been a few attempts in applying computational techniques to identify rationales from ill-structured text such as online discourse. Our study contributes to the research effort in this emerging area by demonstrating the potential and limitations of using rhetorical relations to detect rationales. For details of this work, please refer to [Khazaei & Xiao 2015].

REFERENCES

O. Biran & O. Rambow. 2011. Identifying justifications in written dialogs, *In Proceedings of the IEEE International Conference on Semantic Computing*, pp. 162–168.

L. Carlson, D. Marcu, & M. E. Okurowski. 2001. Building a discourse tagged corpus in the framework of Rhetorical Structure Theory, *In proceedings of the SIGdial Workshop on Discourse and Dialogue*, pp. 1–10.

T. Khazaei & L. Xiao. 2015. Corpus-based analysis of rhetorical relations: A study of lexical cues, *In Proceedings of the IEEE International Conference on Semantic Computing Conference*, pp. 417–423.

M. Taboada.2006. Discourse markers as signals (or not) of rhetorical relations, *Journal of Pragmatics* **38**(4), 567–592.

M. Taboada. C. Anthony & K. Voll. 2006. Methods for creating semantic orientation dictionaries, *In Proceedings of the Conference on Language Resources and Evaluation*, pp. 427–432.

L. Xiao .2013. Do members converge to similar reasoning styles in teamwork? A study of shared rationales in small team activities, *In Proceedings of the iConference*, pp. 524-530.

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