CS410 - Technology Review

Topic: Narrative Clustering and Generalization Technique
---- Using Subject-Verb-Object(SVO) Triplets and Motifs

Allen Zhang (jiaruiz2)

MCS, Dept. of Computer Science, UIUC

Introduction

Text data is one of the most common data types people use to document and convey information. And narrative is one of the most often used representations of text data. Therefore, it naturally becomes the center of study for many researchers. However, as the size of text collection increases along with the development of the Internet, the size of any narrative collection is also often significantly large enough for many researchers to deal with. Thus, in order to sufficiently analyze, categorize and compare various text narrative records, we need an efficient technique to standardize, generalize and cluster different narratives.

This review focuses on the technique of narrative clustering and generalization by adopting the concepts of Subject-Verb-Object (SVO) triplets and motifs. It provides an overview of the approach, illustrating using an instantiation of it from a relevant study, and discussions about its advantages and potential limitations.

Body

First, narratives tend to have dynamic structures. That is, the same narrative can have multiple framings without altering its meaning. Therefore, in order to analyze and compare multiple narratives, we need to first break down the narratives into standardized units, which are consistent for all structures and framings of narratives. To achieve this, our approach proposes that we can represent each narrative by a Subject-Verb-Object (SVO) triplets, each component representing the narrative agent, the action, and the narrative target without losing critical information of the narrative. Therefore, each narrative, no matter how different their framing is, will be broken down into the same structure: *Who-Did What-to Whom*, the last part (object) can also be extended as a noun compound, maintaining extra information of the narrative such as time and location.

A natural question following is how can we extract the SVO triplets from a narrative? A natural language toolkit called spacy provides a dependency parser we can run to extract the main verb of each sentence. Sometimes a single main verb may not be accurate enough to represent the main action of a sentence, thus we need to augment the span of the verb by "joining it with adjacent auxiliary verbs and open clausal complement verbs". [1] Next, we can extract the subject and object from the main verb we derived. Now we have a complete SVO triplet. However, not all triplets are particularly interesting, some that are too general will be trimmed.

Now that we have all triplets ready, we need to define a way to identify the similarity between two triplets. We do this by constructing a semantic space of triplets components. We use Word2Vec embedding to achieve this. The basic idea

is similar to the vector space model we discussed in the text retrieval lecture. But each triplets is represented by two connected vectors, instead of a single vector, as illustrated in this figure[1]:

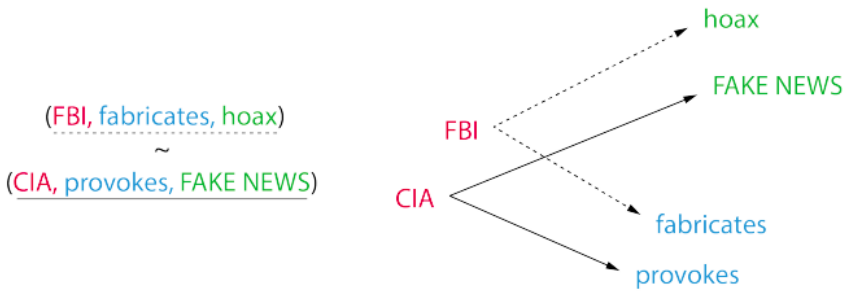

Fig. 4. Similarity between agent-action-target triplets in W2V space. We represent each triplet as a V, with the agent in the middle, and the V's arms pointing towards the action and the target. Two triplets are similar if their subjects are close together, and if they are oriented similarly.

This is from a study which analyzes conspiracy theories using this approach. The red component which connects two out-pointing vectors together is the Subject part, while two vectors are Subject-Verb relation and Subject-Object relation. Two triplets are considered similar (closer to each other in the space) if they have similar Subject, and similar Subject-Verb relation and Subject-Object relation.

When we have the vector space which contains groups of similar triplets, we can use KMeans to cluster those similar triplets together, and then generalize each cluster of triplets into a single narrative-motif.

First, we'd like to find the most representative triplets for each cluster. This can be done by calculating the distance between each triplets and the centroid of the cluster. The shortest one is what we're looking for. To generalize each component, we ask three individual human raters to find the abstract concept that triplets in each cluster have in common. For example, if we can three triplets (CIA, provokes,

FAKE NEWS), (DEA, orchestrates, disinformation campaign), and (FBI, fabricates, hoax), an appropriate generalization would be governmental agency–controls–communications, where “CIA”, “FBI” and “DEA” are all governmental agencies, “provokes”, “orchestrates” and “fabricates” all aim to control. And “news”, “campaign” and “hoax” are all sorts of communications. We call such a generalization a “narrative-motif”.

Conclusion

This technique introduced above provides an efficient approach to analyze multiple text narratives by breaking down the narratives into standardized units, clustering narratives with similar meanings and generalizing them into one single narrative-motif. It has an advantage over other techniques for this purpose because it circumvents the challenge of analyzing narratives with different framings. In addition, it makes it possible to compare and analyze a large amount of narratives at the same time.

However, it does have some limitations. For instance, a narrative may have multiple sentences, and each sentence would potentially generate a SVO triplets. Which triplets from which sentence best describes the entire narrative may vary from narrative to narrative. Hence this approach may not work as well for narratives that are too long with multiple sentences. Also, while this approach attempts to break down narratives without losing important information, certain loss of information is inevitable to avoid, which might influence the overall accuracy of generalizing the narrative.

Reference

Mattia Samory and Tanushree Mitra. 2018. “The Government Spies Using Our Webcams:” The Language of Conspiracy Theories in Online Discussions. Proc. ACM Hum.-Comput. Interact. 2, CSCW, Article 152 (November 2018), 24 pages. https://doi.org/10.1145/3274421