Text Coherence Analysis on Social Media

Aman Raj (170101006), Tushar Bhutada (170101073),
Udbhav Chugh (170101081), Mayank Baranwal (170101084)

CS 565 Project - Group 8

Abstract

In the era of social media, text coherence is critical in enabling effective communication. However, the unique challenges posed by social media such as unstructured text make it difficult to assess the coherency of users' posts correctly. To combat these issues, we propose generating new datasets that reflect the conversational nature of social networking platforms. Furthermore, we put forth a novel model layout to better assess the coherency of these raw posts and describe its characteristics.

1 Introduction

Coherence is a subdivision of linguistics that explores the characteristics of semantically meaningful text. A coherent passage is one where all the parts have a clean transition and are heading in the same direction. The different arguments should flow smoothly from one sentence to the next. Ensuring coherence is key to enabling easy comprehension of an idea. It is, thus, an essential quality for humans from all walks of life. Without coherence, readers can find text choppy as significant gaps may be present between ideas. Thus, coherence is key to ensuring effective communication, regardless of the platform.

1.1 Coherence in Social Media

Social media is a medium that provides even the smallest voices with an extraordinary reach: a clear message has the potential to be heard across the world. While most users write logical sentences, an argument is perceived as coherent only if the ideas flow smoothly from one paragraph to another. Furthermore, users must ensure that there isn't an unreasonable drift from the original theme.

More often than not, with social media posts, users will start a discussion on a particular topic but finish on an entirely different note. As a result, the post can be ineffectual in conveying the intended message. Not only does this subtract from the user's end goal, but it can also be used to confuse others. A malicious actor can choose to obscure the facts by convoluting a topic beyond comprehension. Thus, it is necessary for social media users to not only be coherent but also have the capability to recognise an incoherent argument.

1.2 Project Aim and Scope

The primary goals of this project are two-fold:

- To generate a new dataset that can be used to train models for text coherence analysis in a social media context.
- To provide a novel model architecture for coherence analysis on social media text and analyse its performance on text coherence applications in general.

1.3 Potential Challenges

Social networking platforms pose a host of challenges from a text coherency analysis standpoint such as:

- A large variation in themes and a potential lack of structure can make it difficult for models to discern the characteristics of coherency.
- Users often employ heavy usage of colloquialisms, making it hard to identify and classify keywords and entities.
- Often, social media posts can be only a few sentences long. Defining and assessing text coherency for short texts can be difficult as there are very few entities.

2 Prior Works

2.1 Overview of Existing Datasets

Two major corpora are widely used for text coherence analysis [BL05]. The first is a collection of aviation accident reports written by officials from the National Transportation Safety Board and the second contains Associated Press articles from the North American News Corpus on the topic of earthquakes. The dataset consists of ordered pairs of alternative renderings (x_{ij}, x_{ik}) of the same document d_i, where x_{ij} exhibits a higher degree of coherence than x_{ik}. Existing datasets are synthetically generated using an original document (x_{ij}), taking its different permutations and forming the pair (x_{ij}, x_{ik}) with x_{ij} being the original document (and is more coherent), for each permutation.

2.2 Overview of Existing Models

We now briefly describe the current methods used for analysing text coherence. A few entity-based papers associate adjacent sentences through entities mentioned as noun phrases [BL05; EC08]. Other lexical models connect sentences based on semantic relations between words in sentences [MS16]. Some coherence analysis models employ deep learning with recursive and recurrent neural networks for computing semantic vectors for sentences [LJ17; MS18]. Additionally, a convolution neural network architecture to capture text coherence has also been proposed [Cui+17].

3 Proposed Method

3.1 Dataset Generation

To acquire an extensive dataset for training and testing of text coherence in a social media context, we aim to generate data synthetically. In particular, the candidate set consists of a source social media post and permutations of its sentences. The underlying assumption is that the original sentence order in the source document must be coherent, and so we should prefer architecture configurations that rank it higher than other permutations. Since the coherence relation between different permutations is unknown, our corpus will include only pairwise rankings that comprise the original document and one of its permutations. Given k original documents, each with n randomly generated permutations, we can generate $k \cdot n$ pairwise rankings for training and testing. This approach of dataset generation for social media texts is similar to what has been used in standard text coherence analysis applications [BL05]. We aim to use Kaggle's Reddit Top 1000 or a similar dataset for generating synthetic data for social media text coherence analysis.

3.2 Proposal for Text Coherence Model

At this point, we are exploring a few approaches that we think are suitable for this problem. We now enumerate some of the features that we believe are necessary to capture text coherence and plan to include them in our model.

- We believe a model must capture the transitions from one sentence to another. Large social media posts may cover a variety of topics, making it all the more necessary for the flow between sentences to be smooth. Hence, we believe coherence is best judged by taking a sliding window approach where we consider a few consecutive sentences at a time.
- We plan to extract relevant entities/words in this sliding window that allow us to capture the relevant themes. Existing named entity recognition models may be used for this regard.
- No matter how large a corpus, there is always a chance it does not cover some rare, unused words. As such, we intend to utilise vector embeddings of words instead of the words themselves when identifying the theme of texts. Furthermore, these embeddings allow us to quantify the similarities between different words. To ensure the text is coherent, we will explore various options like clustering or using neural networks on the word embeddings.

4 Individual Contribution to Proposal

From the suggested themes, we picked a few that intrigued us the most to explore. Mayank and Aman examined Word Embeddings and Entity Recognition related areas. Udbhav and Tushar explored the topics of Text Coherence Analysis and Conversational Agents (ConvAI). After a joint discussion, we finally decided to move ahead with the subject of Text Coherence as we felt we could contribute a new corpus for social media coherence analysis and form an architecture that encompasses multiple fields.

Datasets and Architecture: Udbhav and Tushar researched the existing datasets used in text coherence analysis and the methods used in their synthesis. Mayank and Aman examined the different social media corpus and techniques to synthesise new corpus for social media. For the model architecture, all four investigated potential approaches and collectively decided on the features that we plan to explore in the proposed text coherence model.

References

[BL05] Regina Barzilay and Mirella Lapata. "Modeling Local Coherence: An Entity-Based Approach". In: volume 34. Jan. 2005. https://doi.org/10.3115/1219840.1219858.

[Cui+17] Baiyun Cui, Yingming Li, Yaqing Zhang, and Zhongfei Zhang. "Text Coherence Analysis Based on Deep Neural Network". In: (Oct. 2017). https://doi.org/10.1145/3132847.3133047.

[EC08] Micha Elsner and Eugene Charniak. "Coreference-inspired Coherence Modeling". In: *Proceedings of ACL-08: HLT, Short Papers*. June 2008, pages 41–44. https://www.aclweb.org/anthology/P08-2011.

[LJ17] Jiwei Li and Dan Jurafsky. "Neural Net Models of Open-domain Discourse Coherence". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Sept. 2017. https://doi.org/10.18653/v1/D17-1019.

[MS16] Mohsen Mesgar and Michael Strube. "Lexical Coherence Graph Modeling Using Word Embeddings". In: *Proceedings of the 2016 of the Association for Computational Linguistics*. June 2016, pages 1414–1423. https://doi.org/10.18653/v1/N16-1167.

[MS18] Mohsen Mesgar and Michael Strube. "A Neural Local Coherence Model for Text Quality Assessment". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Oct. 2018. https://doi.org/10.18653/v1/D18-1464.