Reverse-engineering Self-selection into YouTube Video Categories

CAPP 30255: Advanced Machine Learning for Public Policy

Ta-Yun Yang & Patrick Lavallee Delgado*

20 April 2020

1 Introduction

When forced to choose an identity, how do we reconcile everything about ourselves into the confines of a label? And despite everything that seems to separate some of us, how is it that we still identify the same? Broad categories are convenient for individuals to sort their preferences into cultures, politics, and other social phenomena, but messy because the ambiguity of language allows each to attach his or her own meaning to those labels. While there may exist quantitative markers with which to explain why an individual subscribes to one category and not another, language in other self-expression is interesting because word choice demonstrates additional stated preference for substantiating the same.

YouTube offers a version of this challenge. As a user upload video content, he must describe the work in the title, description, tags, and category. The user may use any words he wishes, so long as he remains consistent with YouTube's community guidelines, but he is limited by the categories available and he must choose exactly one category. These categories are broad and several may apply to a video. For example, a funny video of a dog playing volleyball could conceivably exist in the "Pets & Animals", "Sports", and "Comedy" categories. Which does the user choose and why? We attempt to answer these questions using the rest of his own description of the video to recreate his decision.

Our data collects the descriptions and activity statistics of the 200 most-watched videos in a week for every week between January 2017 and May 2018. The data has two levels: words in videos and videos in categories.

2 Related work

Our challenge is fundamentally an exercise in *text segmentation*, grouping language into coherent topic clusters using the lexical cohesion that arises from the semantic relationships between words. An early attempt to group text with shared meaning is lexical chaining [9],

*Candidates, MS Computational Analysis and Public Policy, {tayuny, pld}@uchicago.edu.

which links nearby words on whether they exist in related thesaurus categories and evaluates the strength of the resulting chains on frequency and density. This inspired years of work on unsupervised text segmentation, which uses the frequency and co-occurrence of words to identify topic boundaries in a text. Among the first of these algorithms is TextTiling [6], which compares the lexical similarity of adjacent sentence groups from the words those sentences share, and finds a topic boundary where the similarity of words between those groups is low. The choice in topic boundary from several possibilities improves with Latent Semantic Analysis (LDA) [3], which uses principle component analysis to cluster the frequency of co-occurring words in order to reveal semantic dissimilarities between sentence groups.

Beyond the neatness of written text, this area of research also explores language with multiple participants at different times. Addressing this variation in meeting transcripts is the Lexical Cohesion-based Segmenter (LCSeg) algorithm [4], which identifies lexical chains on word frequency alone and compares the cosine similarity of lexical chains among adjacent sentence groups to identify potential topic boundaries in a manner similar to TextTiling. Other work extends LCSeg to asynchronous conversations in emails threads and blog comments [7], which draws paths to sequential fragments in different texts and consolidates topic clusters that LCSeg identifies with those that have a high cosine similarity among their sentences. A generative approach is the TopicTiling algorithm [10], which uses latent Dirichlet allocation (LDA) [2] to estimate topic-word and topic-document probability distributions from a corpus of text that can associate topics to words in a document; in an extension of TextTiling, it finds a topic boundary where the cosine similarity of topics between sentence groups is low.

Recent advances in text segmentation use word embeddings, a departure from one-hot encodings that represent each word as its own dimension towards dense encodings that represent words as vectors in space. Training continuous vector representations of words with domain-specific text becomes its own task [8]. Among algorithms that demonstrate this improvement is GraphSeg [5], which draws a semantic relatedness graph of each sentence in a document to all others and creates segments composed of the adjacent sentences in maximal cliques; It calculates similarity of sentences as the sum of cosines of the embedding vector of each word in one sentence to that of each word in the other.

The discipline builds to deep neural networks, perhaps the best approach to the problem we pursue. In a very similar case recovering the subsection titles of Wikipedia articles, the Sector algorithm [1] learns an embedding of latent topics from ambiguous headings, segments the the document into coherent sections, and labels each section with a learned topic. It uses bidirectional long short-term memory (BLSTM) networks to generate the dense topic embedding matrix. For the text segmentation phase, Sector traces movement in the topic embedding vector space over sentence groups and identifies a new section where that movement is fastest. For the section label decoding phase, it identifies the heading whose words have the strongest association with the topic embedding matrix. In a further departure from previous work, the end-to-end framework in Sector recasts text segmentation as a semi-supervised task: learning the semantic similarities of the text to define the scope of topics that in turn segment the text.

References

[1] Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184, 2019.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. Latent semantic analysis for text segmentation. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001.

[4] Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan, July 2003. Association for Computational Linguistics.

[5] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 125–130, Berlin, Germany, August 2016. Association for Computational Linguistics.

[6] Marti A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[7] Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, 47:521–573, 2013.

[8] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, pages 1–12, January 2013.

[9] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

[10] Martin Riedl and Chris Biemann. TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea, July 2012. Association for Computational Linguistics.