Entity Attention for Improved Interpretability of Text Classification

Brendan Kennedy
btkenned@usc.ed

1 Introduction

My project is to develop explainable classification models for moral sentiment in (noisy) text, first by using NER techniques (including distant supervision, e.g. Ren et al. (2015)) to provide entity-driven tokenization, then by applying attention-based techniques to learn joint, task-specific embeddings of entities and entity types for improved classification. Specifically, this process involves initializing entity and entity-type representations from an external sources (i.e. Knowledge Base) and learning representations of their *interaction* within documents, respective to a given label space.

2 Literature Survey

Prior work which supports my project falls into roughly three distinct areas: attention-based architectures (in both text and vision) which have the goal of providing explanatory power in visualizing/understanding the "reasoning" of a neural network in making a decision, based on a specific input; techniques for learning entity representations from Knowledge bases and from text corpora, and task-specific methods for classifying morality in text.

2.1 Attention-based Architectures

Explainability is an important goal of machine learning, both with regard to improving performance (by systematically understanding the causes of errors and successes) and in contributing the domains of ML application. In areas that value explanation at least as much as prediction, like medical applications, "the development of methods for visualizing, explaining and interpreting deep learning models has recently attracted increasing attention" (Samek et al., 2017, p. 1). One way to explain (or to "decompose") the predictions of a deep learning apparatus is to conduct analysis of the sensitivity of the network to changes in the inputs (see Samek et al., 2017).

Other ways, which are increasingly being developed and used, include using "attention" mechanisms to determine the relevance of different parts of inputs with specific aspects of the output. A clear example of this is using attention-based networks for aspect-level sentiment analysis (Tang et al., 2016; Wang et al., 2016). In this task, there are multiple levels of "aspect" (e.g. "food", "service") which can have differing sentiment in the same document), making it a natural application for attention. Attention is used here to determine which words are most relevant to which aspects.

In medical applications, there is a strong need to explainable neural models. In computer vision, researchers have proposed "Attention-based multiple-instance learning" (Ilse et al., 2018), wherein inputs are either naturally or artificially segmented into "chunks", and the task is framed a multi-instance learning problem (where there are no labels for instances, but labels for "bags" of instances). This attentional architecture identifies the aspects of the image which contribute to the overall label.

Also in the medical domain, but with textual input, Mullenbach et al. (2018) propose CAML, or "Convolutional Attention for Multi-label Classification". This operates similarly to the attention networks for aspect-level sentiment analysis; the meaningful difference is that the attention is applied over a CNN output, learning representations for each label.

2.2 Moral Sentiment Classification on Twitter

In the data for this project, I am using a corpus of about 30,000 Tweets which have been annotated (by at least two trained annotators) for 5 di-

mensions of "moral rhetoric" using a carefully designed coding manual (Hoover et al., 2017). These tweets have been studied in prior studies on moral sentiment with various methods, which are discussed below.

Prior studies on select subsets of the 30K corpus include the study of purity rhetoric in online communities (Dehghani et al., 2016), the study of violence in protests with an online presence (Mooijman et al., 2018), and analyzing the political polarity of blog posts (Hoover et al., 2018). The methods used in these papers to classify moral rhetoric are predominately dictionary based methods and "Distributed Dictionary Representations" (DDR) (Garten et al., 2018), which computes the cosine similarity between the spaces of concept dictionaries and documents. The dictionary used by these studies is the "Moral Foundations Dictionary"[1]

In relation to the design of the proposed study, Lin et al. (2017) improve classification performance on a subset of these posts (applying to Tweets gathered in the 2015 Baltimore Protests) using entity linking. The authors link entities in the Tweets to Wikipedia KB using "TagMe" software (Ferragina and Scaiella, 2010), which is suitable for short documents with irregular, dynamic text (like Twitter). The Wikipedia articles that are linked to by the entities in the text are then used to augment the information in the Tweets, by adding the Wikipedia abstracts as text features (for later processing) and inputting structural features from the Wikipedia page as well. Using this method, paired with a structured LSTM sentence encoder to learn textual features from both the original source and the Wikipedia abstracts, the authors were able to clearly outperform competing baselines for classifying moral rhetoric in Tweets. This is currently the state of the art in this area.

2.3 Learning entity representations from text

In my project, I am interested in learning entity representations — and representations of their interaction with reference to the label space of the particular task. Concretely, this means building on existing entity representations (e.g. those learned from Knowledge Bases) to reflect the types of interacts which correlate strongly with certain types of labels.

[1] https://www.moralfoundations.org/sites/default/files/files/downloads/moral%20foundations%20dictionary.dic

Popular approaches in learning entity and relation embeddings from Knowledge Graphs are headlined by Lin et al. (2015), who propose "TransR". TransR learns entity and relations in separate spaces, then projects entities into the relation space. Then, a translation is learned between related entities in this projected space.

In the area of learning entity representations jointly from both text and KB/KG, there are several approaches. Shalaby et al. (2018) proposes a simple adaptation of the skipgram model in order to jointly learn "concept representations" from Wikipedia text and Probase knowledge graph. By using surface representations of entities as an anchor between text and KB, Wang et al. (2014) learn representations of entities across KB and text which preserve the entity-relations. This process has three components: a knowledge model (for representing KB/KG), a text model, and an alignment model.

Recently, Newman-Griffis et al. (2018) proposes to learn jointly learning entity- and text-representations by extending the skipgram model using distant supervision. The authors, use the mapping from entities to surface forms as the distant supervision. Essentially, they add additional loss terms (to capture entities and relations) to learn entity interaction information within the text.

In another approach, Cao et al. (2017) proposes to jointly learn entities and text by adding a third component: learning "mention" representations, or a representation of how a given entity appears in the text.

3 Remaining Issues

Based on the reviewed literature, the clear problem I am seeking to solve is how to successfully combine the joint learning of text and entity representations with the attentional learning of the relationship of these representations with a given label space. By using existing attention-based approaches for multi-label learning, and by building on one of the existing approaches listed above for learning entity representations in the context of their occurrence in text, I will be working to build an attention-based neural network to use entity interaction information to improve classification and explanation results.

References

Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1623–1633.

Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General* 145(3):366.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, pages 1625–1628.

Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods* 50(1):344–361.

Joe Hoover, Kate Johnson, Reihane Boghrati, Jesse Graham, and Morteza Dehghani. 2018. Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology* 4(1).

Joseph Hoover, Kate Johnson-Grey, Morteza Dehghani, and Jesse Graham. 2017. Moral values coding guide .

Maximilian Ilse, Jakub M Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712* .

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*. volume 15, pages 2181–2187.

Ying Lin, Joe Hoover, Morteza Dehghani, Marlon Mooijman, and Heng Ji. 2017. Acquiring background knowledge to improve moral value prediction. *arXiv preprint arXiv:1709.05467* .

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour* page 1.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. volume 1, pages 1101–1111.

Denis Newman-Griffis, Albert M Lai, and Eric Fosler-Lussier. 2018. Jointly embedding entities and text with distant supervision. *arXiv preprint arXiv:1807.03399* .

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. 2015. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 995–1004.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* .

Walid Shalaby, Wlodek Zadrozny, and Hongxia Jin. 2018. Beyond word embeddings: Learning entity and concept representations from large scale knowledge bases. *arXiv preprint arXiv:1801.00388* .

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900* .

Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. pages 606–615.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1591–1601.