Transfer Learning for Causal Sentence Detection

Manolis Kyriakakis, Ion Androutsopoulos, Joan Ginés i Ametllé, Artur Saudabayev
Causaly Inc. (https://www.causaly.com)

Department of Informatics, Athens University of Economics and Business, Greece
(http://nlp.cs.aueb.gr)

Introduction

Causality is very important to biomedical research. A wide range of biomedical questions, from what causes a disease to what drug dosages should be recommended and which side effects might be triggered, center around detecting particular causal relationships between biomedical entities. In natural language processing, causality detection is often viewed as a type of relation extraction, where the goal is to determine which relations (e.g., cause-effect), if any, hold between two entities in a text. Causality can be expressed in many ways, from using explicit lexical markers (e.g., "smoking causes cancer") to markers that do not always express causality (e.g., heavy smoking led to cancer" vs. "the nurse led the patient to her room") to no explicit markers ("she was infected by a virus and admitted to a hospital").

We focus on detecting causal sentences, i.e., sentences conveying at least one causal relation. This is a first step towards mining causal relations from texts. Once causal sentences have been detected, computationally more intensive relation extraction methods can be used to identify the exact entities that participate in the causal relations and their roles (cause, effect). To bypass the scarcity of causal instances in relation extraction datasets, we exploit transfer learning (ELMO and BERT).

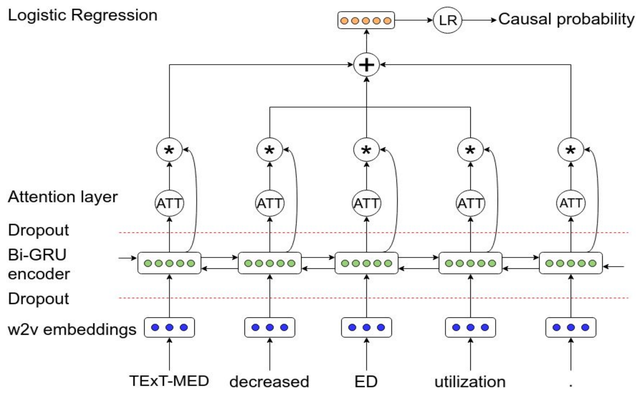
We believe this is the first work to:

- focus on causal sentence detection as a binary classification task,
- consider causal sentence detection in both generic and biomedical texts,
- explore the effect of transfer learning in this task.

Methods

- **LR (n-grams):** LR classifier with TF - IDF n-gram features (word n-grams, n = 1, 2, 3).
- **BIGRUATT**: The model views each sentence as the sequence of its pre-trained w2v embeddings. A sentence is represented by the weighted sum of the (concatenated forward/backward) states of the BIGRU chain. The weights are produced using a linear self-attention mechanism. The weighted sum is fed to a logistic regression (LR) layer to estimate the probability that a sentence is causal.
- **BIGRUATT+ELMO:** Similarly to BIGRUATT, a sentence representation is constructed using a linear self-attention mechanism over BIGRU's hidden states. The BIGRU is fed with the concatenation of the pre-trained and contextualized ELMO embeddings of each sentence.
- **BERT+LR:** A logistic regression (LR) layer is added on top of BERT to estimate the probability that an input sentence is causal. The LR layer is fed with the embedding of the 'classification' token, which BERT produces for each sentence. The entire network is fine-tuned on causal sentence detection training data, with a small learning rate that aims to prevent catastrophic forgetting.
- **BERT+BIGRUATT:** Same as BIGRUATT, but uses the context-aware word embeddings that BERT produces at its top layer as the input to BIGRUATT, instead of w2v embeddings. Similarly to BERT+LR, the entire network is fine-tuned with a small learning rate.

BIGRUATT Model



Datasets

Similarly to Li and Mao (2019) we converted SemEval, CausalTB and EventSL to causal detection datasets. Furthermore, we developed a biomedical causal detection dataset (BioCausal-Large) from PubMed sentences, a subset of which (BioCausal-Small) we made publicly available (https://archive.org/details/CausalySmall). The average inter-annotator agreement on a sample of 300 sentences was 79.36%. Cohen's Kappa was 0.56.

-SemEval (1325 causal, 2500 non-causal sentences)
-CausalTB (244 causal, 500 non-causal sentences)
-EventSL (77 causal, 200 non-causal sentences)

- New **BioCausal-Small** (1113 causal, 887 non-causal sentences)
- New **BioCausal-Large** (7562 causal, 5780 non-causal sentences)

Evaluation (generic datasets)

	SemEval		CausalTB (Small)		EventSL (Small)	
Model	*F1*	*AUC*	*F1*	*AUC*	*F1*	*AUC*
LR (n-grams)	76.22	87.50	36.36	65.02	42.86	73.55
BIGRUATT	90.64	96.57	69.98	74.38	63.65	70.36
BIGRUATT+ELMO	**92.81**	97.03	75.08	82.06*	66.55	77.31
BERT+LR	91.55	96.94	**80.55**	82.26*	72.35	78.15*
BERT+BIGRUATT	91.45	**97.61**	80.06	**84.27***	**73.09**	**84.17***

In the AUC columns, stars indicate statistically significant ($p \leq 0.05$) differences (two-tailed approximate randomization tests) compared to BIGRUATT. AUC scores are the main ones to consider, since they examine performance at multiple classification thresholds. F1 scores were computed for a classification threshold of 0.5.

Evaluation (biomed datasets)

Dataset	BioCausal-Small		BioCausal-Large	
Model	*F1*	*AUC*	*F1*	*AUC*
LR (n-grams)	77.49	87.65	79.21	86.54
BIGRUATT	85.97	93.91	85.84	93.71
BIGRUATT+ELMO	**87.32**	**94.95**	86.77	94.64*
BERT+LR	85.64	90.75	**87.33**	92.77
BERT+BIGRUATT	85.87	93.75	87.09	**94.70**

Transfer learning (ELMO, BERT) improves the AUC of BIGRUATT by a wide margin in the two smallest datasets (CausalTB, EventSL), which contain only hundreds of instances. In the other three datasets, which contain thousands of instances, the AUC differences between transfer learning and plain BIGRUATT are small.

Conclusions

- Transfer Learning significantly boosts causal sentence detection performance on small datasets.
- With a few thousands of training sentences (both in generic and biomedical texts) BIGRUATT reaches a performance plateau, then increasing the dataset size or employing transfer learning does not significantly improve performance.
- BERT+BIGRUATT consistently performed better (in terms of AUC) than BERT+LR. Thus, using deeper task-specific architectures on top of BERT may be a promising alternative to the common practice of adding only shallow task-specific models on top of BERT.

References

- Li and Mao, 2019, Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*.
- Hendrickx et al., 2009, SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.
- Mirza et al., 2014, Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL Workshop on Computational Approaches to Causality in Language*.
- Caselli et al., 2017, The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*.