Prompt based learning for temporal relation extraction in medical domain

Proposer(s) / Proposatzailea(k):

Oier Lopez de Lacalle
Aitziber Atutxa

Contact / Kontaktua: oier.lopezdelacalle@ehu.eus

Description / Deskribapena

Building Information Extraction (IE) systems for real-world applications is very costly and has suffered from data-scarcity problems, due in part to the expertise and time required to annotate training data at a large scale with sufficient consistency, but also due to poor transfer between domains: IE annotations depend on the schema used in each domain, and moving to new domains requires a new schema, new annotation guidelines and manual annotation of new data. In many cases, there is some information overlap between schema, but performing transfer learning to leverage such overlap (i.e. learning from **multiple sources**) can be difficult: it often requires manually mapping labels between schemas, which is typically brittle, cumbersome and requires costly domain expertise.

In order to save annotation effort, recent work recasts IE tasks as Textual Entailment tasks (Sainz et al., 2021). For instance, (Sainz et al., 2021). manually verbalize each relation type in the Relation Extraction (RE) dataset TACRED (Zhang et al, 2017) to generate hypotheses for each test example, and then apply an entailment model to output the relation type of the hypothesis with highest entailment probability.

In this project we want explore if entailment models are able to model temporal relationships of event in medical domain. For that we will relay on the work presented in (Sainz et al., 2021) in which we'll need to construct templates for verbalizing each temporal relation and deploy domain specific entailment model.

Experiments will be run on public clinical histories from E3C corpus (Magnini et al., 2020), which is composed by general clinical statements, presenting the reason for a clinical visit, the description of physical exams, and the assessment of the patient's situation. Data is provided in five different languages: English (EN), Italian (IT), Spanish (ES), French (FR) and Basque (EU) offering the opportunity to design verbalization templates for multiple languages.

References

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In Proceedings of the 2021. Conference on Empirical Methods in Natural Lan guage Processing, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics

Magnini, B., Altuna, B., Lavelli, A., Speranza, M., Zanoli, R.: The e3c project: Collection and annotation of a multilingual corpus of clinical cases. In: CLiC-it (2020)

Goals / Helburuak

The **main objective** of the project is to explore and implement a entailment based prompt learning approaches to extract temporal relations of events. The key objectives are the following:

1. Adapt a2t framework (Sainz et al. 2021) to run on E3C corpus.
2. Design and evaluate verbalization templates for temporal relation extraction.
3. Train domain specific entailment model for medical texts.

Requirements / Betebeharrak

None

Framework / Esparrua

Python, Tensorflow/Pytorch, a2t framework (https://osainz59.github.io/Ask2Transformers/a2t/index.html)

Tasks and plan / Atazak eta plana

- Analyze the E3C dataset.
- Analyze the state of the art in temporal relation extraction, and related topics
- Adapt a2t framework to E3C dataset.
- Design and implement templates for temporal relation extraction
- Exploration of multiple pretrained NLI models.
- Test and evaluate the implemented algorithm on E3C.
- Analyse the output of the system to 1) perform an error analysis and 2) purpose possible improvements.
- Write up the report.