Focused Contrastive Training for Test-based Constituency Analysis

Benjamin Roth, Erion Çano

Digital Philology @ Data Mining and Machine Learning lab, **University of Vienna**

Overview

- Self-training of grammaticality models for linguistic tests
- Linguistic test
  - Reformulation of a sentence according to a rule (a *test transformation*)
  - If reformulation is grammatical, the rule has detected a phenomenon of interest
  - Usually done by linguists…
    … but we want to automate the process and replace intuition about grammaticality by a **constrastively trained model**
- Here: Constituent tests
  - Constituent: Part of a sentence that functions as a unit
  - Pronominalization test: Replace part of sentence with pronoun
- Findings
  - Improved quality by selecting the right **positive** examples
  - Characteristics of **positive** examples should match the characteristics of negatives
  - Strong improvements when using our proposed scheme of ***focused contrastive training***

Pro-Forms

- **Pronouns:** it, ones, this, that, they, I, we, you
- **Pro-PPs (preposition+pronoun):** of it, for it, in it;
- **Pro-VPs:** did so, do that, does that
- **Pro-sentences:** it is, that it is
- **Pro-adverbs:** there, this way

Grammaticality Model: Contrastive Training & Application

- Train a **grammaticality** model (**contrastive training**)
  - **Positive samples:** Sentences from corpus
  - **Negative samples:** Corrupted sentences (a random subspan is replaced by a proform)
- Find out whether a subspan is a constituent:
  - **Replace** by pro-form
  - **Predict** grammaticality score
  - **Aggregate** scores over all pro-forms (maximum, average, voting)

Focused Contrastive Training

- Grammaticality model should **focus** on cases where a linguistic test has been applied
  - Only keep those **positive** instances that could have the result of a test
    ➔ Occurring sentences that contain one of the pro-forms
  - **Generate negatives**, such that global statistics (occurrences of pro-forms) match the selected positives
    ➔ Replace random subspans proportionally with pro-forms
  - **Mark** the occurrence of a pro-form by extra symbols
- Example:
  Other fund managers were similarly sanguine .
  Other fund managers <S> did so <E> .

Data, Experiments

- **CoLA**. For supervised training, we use the grammaticality annotations of 10,657 sentences from Warstadt et al. (2019).
- **WSJ.** For contrastive training, we use sections 02-21 of the WSJ portion of the Penn Treebank which total in 37,374 sentences.
- **RoBERTa.** For all experiments, we fine-tune a pre-trained RoBERTa-base (Liu et al., 2019) model from the huggingface transformer library

Results

- Better than a supervised general grammaticality model at detecting constituents
- Identified a set of strong pro-forms
- *Focused constrastive training* much better than naively picking positives and generating negatives

	sentence	pred.?	const.?
orig.	Mr. Sim figures it will be easier to turn Barry Wright around since he 's now in the driver 's seat .		
repl.	Mr. Sim figures it will be easier to turn Barry Wright around **this way** .	Yes	Yes
repl.	Mr. Sim figures it will be easier to turn Barry **it** 's seat .	No	No
orig.	Other fund managers were similarly sanguine .		
repl.	Other fund managers **did so** .	Yes	Yes
repl.	Other fund **of it** sanguine .	No	No
orig.	On Saturday night , quite a few of the boys in green and gold salted away successes to salve the pain of past and , no doubt , future droughts .		
repl.	On Saturday night , quite a few of the boys in green and gold salted away successes to salve **did so** .	No	Yes
repl.	**This way** green and gold salted away successes to salve the pain of past and , no doubt , future droughts .	Yes	No

pronoun set	accuracy
this way	0.7500
this way, did so	0.7763
this way, did so, of it	0.7883
this way, did so, of it, it	0.7996
it, ones, did so	0.6848

data	scheme	score-full	score-greedy	Cao et al.
dev	non-focused	0.7075	0.8188	0.7243
test	non-focused	0.7142	0.8399	0.7470
dev	focused	0.7907	0.8433	0.7518
test	focused	0.7922	0.8560	0.7602

Table 3: Comparison of different pronouns sets used for scoring (for the model trained on the *full* pronoun set).