Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data (EMNLP 2020)

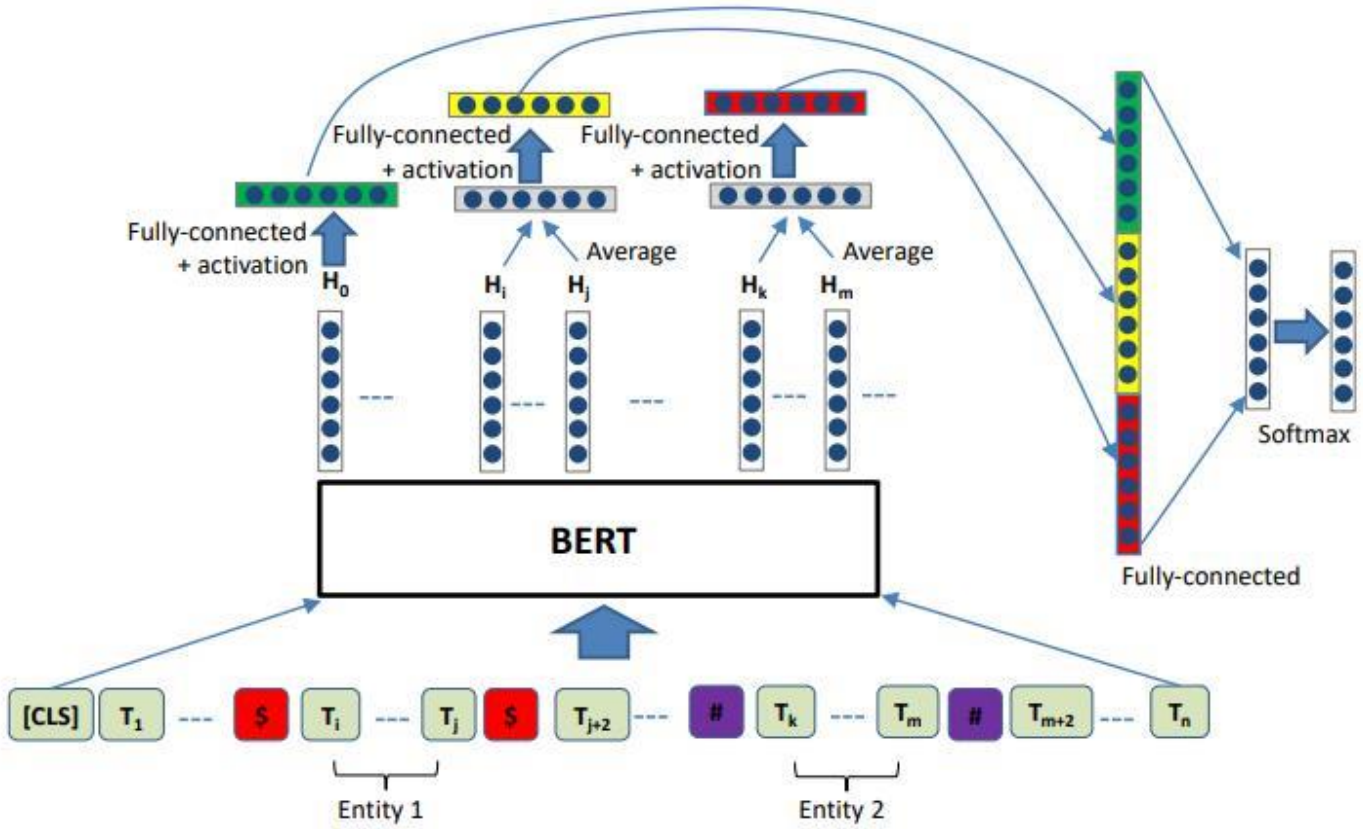
인공지능학과
백형렬

Contents

1. Introduction

Relation Extraction Task

- Input: Sentence, Subject(Entity1, e1), Object(Entity2, e2)
- Output: The Relation of e1 and e2



<모델 예시: Wu and He, 2019>

1. Introduction

TACRED Dataset (Zhang et al., 2017)

- Crowdsourcing: Text Analysis Conference Knowledge Base Population (TAC KBP) challenge에서 공개되는 corpus 중 106,264 examples를 발췌하여 라벨링.
- Labeled Dataset: Subject(e1)/Object(e2) 간의 Relation Type(42개)
- Imbalanced Dataset: "no_relation"이 전체 데이터의 78.68%

Example	Entity Types & Label
Carey will succeed **Cathleen P. Black**, who held the position for 15 years and will take on a new role as **chairwoman** of Hearst Magazines, the company said.	**Types:** PERSON/TITLE **Relation:** *per:title*
Irene Morgan Kirkaldy, who was born and reared in **Baltimore**, lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkaldy.	**Types:** PERSON/CITY **Relation:** *per:city_of_birth*
Pandit worked at the brokerage Morgan Stanley for about 11 years until 2005, when he and some Morgan Stanley colleagues quit and later founded the hedge fund **Old Lane Partners**.	**Types:** ORGANIZATION/PERSON **Relation:** *org:founded_by*
Baldwin declined further comment, and said JetBlue chief **executive** Dave Barger was unavailable.	**Types:** PERSON/TITLE **Relation:** *no_relation*

Relation	Total	Percentage	Train 2009–2012	Development 2013	Test 2014
no_relation	94001	78.68%	60179	19305	14517
org:alternate_names	1515	1.27%	893	380	242
org:city_of_headquarters	656	0.55%	437	125	94
org:country_of_headquarters	878	0.73%	540	215	123
...					
per:stateorprovince_of_death	133	0.11%	65	53	15
per:stateorprovinces_of_residence	560	0.47%	374	89	97
per:title	4424	3.70%	2733	1065	626
Total	119474	100.00%	75050	25764	18660

<TACRED 예시: Zhang et al., 2017>

2. RE Issue

Model Heuristics

- 모델이 **일종의 편법(어림짐작)**으로 문제를 해결.
- 문제를 이해하지 못했지만 score는 높을 수 있음

*"... **decision rules** that are used by ML models to score high on a test set, but **which are too simplistic to solve the underlying problem***

...

*We show that state of the art models trained on TACRED are often "**right for the wrong reasons**" (McCoy et al., 2019): instead of learning to perform the intended task, they **rely on shallow heuristics which are effective for solving many dataset instances**, but which may **fail on more challenging examples.**" (*Rosenman et al., 2020*)*

2. RE Issue

Model Heuristics In Relation Extraction

- Event Heuristic: Sentence 내용이 relation과 관련있기 때문에 해당 relation을 prediction. 즉 entity간의 관계 고려하지 않음.

e.g.

Edward[PERSON] was **born in** York in **1561[YEAR]** , the son of John, and his wife Mary.
-> per:birth_date(O)

e.g.

Edward was **born in** York in **1561[YEAR]** , the son of John, and his wife **Mary[PERSON]**.
-> per:birth_date(X)

- Type Heuristic: Entity1, 2 Type이 relation과 관련있기 때문에 해당 relation을 prediction. 즉 sentence 무시.

e.g.

He[PERSON] *escaped the worst taint of the Watergate scandal because his image as a* ***Southern Baptist[RELIGION]*** *prevented other Nixon aides from fully trusting him .*
-> per:religion(O)

e.g.

He escaped the worst taint of the Watergate scandal because his image as a ***Southern Baptist[RELIGION]*** *prevented other* ***Nixon aides[PERSON]*** *from fully trusting him .*
-> per:religion(X)

3. Proposed Method

Challenge Relation Dataset

- Heuristic 방법을 사용하면 틀리도록 Revised Dataset 생성
- 10,844 instance(3,000 distinct sentences)
- e.g. person:birth_date

[e1 Ed] was born in **[e2 1561]**, the son of John, a carpenter, and his wife Mary.

Ed was born in **[e2 1561]**, the son of **[e1 John]**, a carpenter, and his wife Mary.
Ed was born in York in **[e2 1561]**, the son of John, a carpenter, and his wife **[e1 Mary]**.

3. Proposed Method

QA model

- 지금까지 Relation Extraction은 classfication task로 학습
  - (sentence, e1, e2) -> relation

e.g.

sentence: **Sam** was born in **1991**
(sentence, **Sam**, **1991**) -> person:birth_date

- QA task 학습방식으로 모델링하면 Heuristic 해결
  - relation마다 query template 작성
  - (sentence, query=(e1, relation)) -> e2
  - (sentence, query=(e2, relation)) -> e1

e.g.

person:birth_date
QA1. (sentence, When was **Sam** born?) -> **1991**
QA2. (sentence, Who was born in **1991**?) -> **Sam**

4. Experiments

Metric

- Acc(+): CRE dataset 전체에서 True Postive 비율. 즉 (True) relation_r -> (Prediction) relation_r
- **Acc(-)**: CRE dataset 전체에서 True Negative 비율. 즉 (True) no_relation -> (Prediction) no_relation
- e.g.
  - TruePositive
    - (True)per:birth_date -> (Prediction)per:birth_date
    - **[e1 Ed]** was born in **[e2 1561]**, the son of John, a carpenter, and his wife Mary.
  - TrueNegative
    - (True)no_relation -> (Prediction)no_relation
    - Ed was born in **[e2 1561]**, the son of **[e1 John]**, a carpenter, and his wife Mary.
    - Ed was born in York in **[e2 1561]**, the son of John, a carpenter, and his wife **[e1 Mary]**.

- Heuristic을 사용하는 모델은 Acc(-)가 낮음. 즉, (True)no_relation -> (Prediction)per:birth_date (X)

4. Experiments

Model	Acc	Acc_+	Acc_-
RC-SpanBERT	63.5	89.7	42.5
RC-BERT	67.1	70.0	64.8
RC-KnowBERT	72.4	84.2	62.9
RC-RoBERTa	73.1	82.9	65.3
QA-SpanBERT	75.5	71.5	78.7
QA-BERT	67.4	62.9	70.9
QA-ALBERT	75.3	71.5	78.8

Table 2: CRE accuracy for the RE and QA models. Acc_+ refers to accuracy on positive instances. Acc_- refers to accuracy on negative instances.

Model	Acc	Acc_+	Acc_-
(a) Trained on TACRED			
RC-SpanBERT	62.8	89.5	41.6
RC-BERT	65.8	68.4	63.7
RC-KnowBERT	71.6	83.0	62.5
RC-RoBERTa	75.5	85.4	68.0
(b) Trained on TACRED + half CRE			
RC-SpanBERT	84.4	85.7	83.4
RC-BERT	78.7	86.1	72.7
RC-KnowBERT	82.4	81.9	82.7
RC-RoBERTa	83.0	83.4	82.6

Table 4: Acc/Acc_+/Acc_- scores on half of the CRE dataset, models trained on TACRED training set (a), models trained on TACRED training set with examples from the second half of the CRE dataset (b).

<Experiments score: Rosenman et al., 2020>

5. Conclusion

Issue
- Relation Extraction에서 Model이 어림짐작Heuristics으로 testset에 대해 high score 달성

Proposed solution
- QA task방식으로 모델 학습
- 데이터셋에 Challenging examples추가

6. References

- Rosenman, S., Jacovi, A., & Goldberg, Y. (2020, November). Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

- Wu, S., & He, Y. (2019, November). Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2361-2364).

- Zhang, Y., Zhong, V., Chen, D., Angeli, G., & Manning, C. D. (2017, September). Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 35-45).