Joint task learning for relation extraction and named entity recognition

Authors

Institution

In this work we present our system for RuREBus challenge held together with Dialog 2020 conference. The task consisted of 3 tracks: named entity recognition, relation extraction with provided named entity tags and end-to-end relation extraction. Our system took the first place in the named entity recognition track and the second place in the third track. For the second task we failed to submit the solution till the deadline but it was among the best systems. The systems for all tasks are based on Transformer models.

Key words: relation extraction, named entity recognition, transformer, bert

Совместное обучение моделей для извлечения отношений и именованных сущностей

Авторы

Организация

В данной работе мы представляем нашу систему для соревнования RuREBus, проводящегося совместно с конференцией Dialog 2020. Задача состояла из 3 дорожек: распознавание именованных сущностей, классификация отношений между заранее аннотированными именованными сущностями и извлечение отношений из неаннотированного текста. Наша система заняла первое место в задаче распознавания именованных сущностей и второе место на третьей дорожке. Для второй задачи мы не успели своевременно представить решение, но оно оказалось в числе лучших систем. Системы для всех задач основаны на моделях Transformer.

Ключевые слова: извлечение отношений, распознавание именованных сущностей, transformer, bert

1 Introduction

There are many ways to extract information from text. One of the most popular approaches is to extract named entities and classify relations between them. One of the most popular datasets for this task is TACRED [16] where semantic relations are understood as relations between two pairs of entities.

<TACRED METHODS>

Unfortunately, such annotated datasets are scarce for most languages besides English. Some researchers have tried to solve this problem for the Russian language. They have used unsupervised approaches based on knowledge databases such as Wikidata and online encyclopedias such as Wikipedia. Models trained this way tend to be not specialized because the original database does not contain relations from the required domain. They also tend to work only for the most popular relation types such as geographical or professional ones which are common to Wikipedia.

There are few annotated datasets for the Russian language. Among similar tasks to relation extraction there was held FactRuEval 2016 within the conference Dialog 2016. Within the competition contestants had to extract facts from news articles and to fill special slots in these facts (e.g. one of the fact types was 'Occupation' and its fields were 'POSITION', 'WHO', 'WHERE' and 'PHASE').

RuREBus competition was devoted to the problem of relation extraction and named entities recognition in a specialized business domain.

2 Shared task overview

The organizers of the competition have provided 188 annotated texts as the training dataset and 544 texts as the test dataset for the first and thirds tracks and <N> <N> for the second track respectively. All texts were provided by the Ministry of Economic Development of the Russian Federation. The corpus consists of various regional and strategic plan reports. There are in total 8 named entity classes and 11 semantic relation classes. The organizers have also provided a large unannotated dataset for language model fine-tuning. However, we did not use it.

TEXT EXAMPLE

Named entity groups could contain rather broad types of entities, for example "SOC" entities contained social groups as well as various social attributes - phrases like 'blue collar workers' and 'housing accessibility' corresponded to this group.

2.1 Dataset

Type	Description
MET	Some quantitative metric
ECO	An economy entity or facility
BIN	A binary attribute
CMP	Comparative attribute
QUA	Qualitative attribute
ACT	Activity, actions, implemented policies
INST	Institutions and organizations
SOC	Social groups and characteristics

Table 1: Named entity types

Group	Type	Description
Current state of affairs	NNG	now negative
Current state of affairs	NNT	now neutral
Current state of affairs	NPS	now positive
Results	PNG	past negative
Results	PNT	past neutral
Results	PNS	past positive
Forecasts	FNG	future negative
Forecasts	FNT	future neutral
Forecasts	FNS	future positive
Goals	GOL	some abstract goals
Tasks	TSK	tasks and performed actions to achieve goals

Table 2: Semantic relation types

3 Our solution

3.1 Named entities recognition

3.2 Stand-alone relation extraction

3.3 Relation classification with provided named entity tags

4 Results

5 Conclusion

References

[1] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1247–1249, 2008.

[2] Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. DeepPavlov: Open-Source library for dialogue systems. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*, pages 122–127, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. oct 2018.

[4] Ralph Weischedel et Al. OntoNotes Release 5.0 LDC2013T19. *Linguistic Data Consortium*, 2013.

[5] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving Pre-training by Representing and Predicting Spans. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy, 2019. Association for Computational Linguistics.

[6] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*, 2019.

[7] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, Omer Levy, and † Allen. SpanBERT: Improving Pre-training by Representing and Predicting Spans. Technical report.

[8] T A Le, M A Petrov, Y. M. Kurato, and M S Burtsev. Sentence Level Representation and Language Models in The Task of Coreference Resolution for Russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2019)*, pages 341–350, 2019.

[9] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, Atlanta, Georgia, 2013. Association for Computational Linguistics.

[10] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. pages 1003–1011, 2009.

[11] James Pustejovsky and Amber Stubbs. *Natural language annotation for machine learning.* 2013.

[12] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6323 LNAI, pages 148–163, 2010.

[13] A. S. Starostin, V. V. Bocharov, S. V. Alexeeva, A. A. Bodrova, A. S. Chuchunkov, S. S. Dzhumaev, I. V. Efimenko, D. V. Granovsky, V. F. Khoroshevsky, I. V. Krylova, M. A. Nikolaeva, I. M. Smurov, and S. Y. Toldova. FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 702–720, 2016.

[14] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, 2012. Association for Computational Linguistics.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 5999–6009, 2017.

[16] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 35–45, 2017.

[17] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017.