TNT-KID: Transformer-based Neural Tagger for Keyword Identification

Matej Martinc **Blaž Škrlj** **Senja Pollak**

Jozef Stefan Institute, Ljubljana, Slovenia

EMBEDDIA presentation Paris - November 8th 2021

Introduction and Methodology

Supervised approach to keyword detection

Keyword identification deals with automatic extraction of words that represent crucial semantic aspects of the text.

Supervised approach allows the model to adapt to a specific language, domain and keyword assignment regime of a specific text (e.g., a variance in a number of keywords).

The problem is the availability of manually labeled training sets. ⇒ To reduce the needed amount of manually labeled data, we employ a two step training approach:

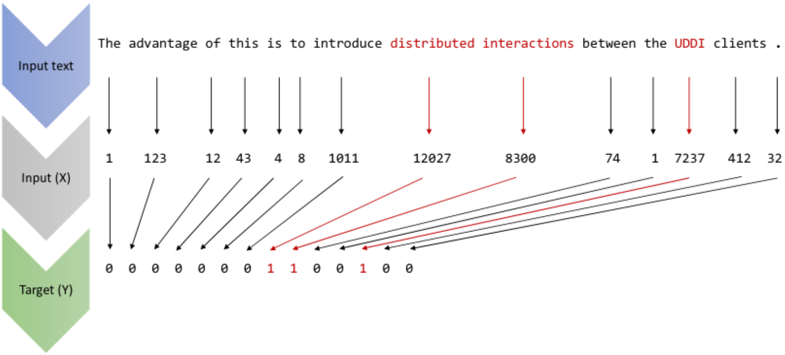
- Language model pretraining on a **small** domain specific corpus ⇒ two training objectives are tested, **autoregressive** and **masked** language modelling!
- Fine-tuning on a manually labeled dataset

Fine-tuning

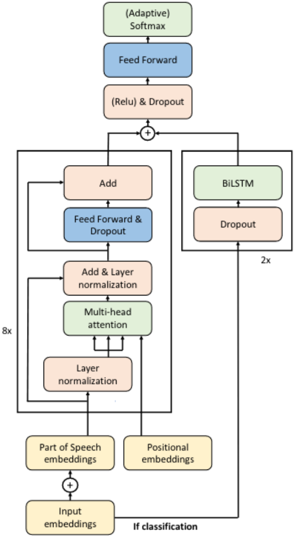
Sequence labelling approach towards keyword identification

Adaptation of the transformer architecture [Vaswani et al., 2017]:

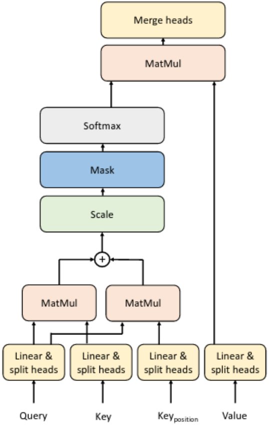
- Re-parametrization of the attention mechanism
- Additional BiLSTM encoder
- Custom loss function ⇒ **due to imbalance between positive and negative classes**



Architecture



(a) Model architecture.



(b) The attention mechanism.

Experiments

Experimental design - English

The approach was tested on 8 datasets from two domains, computer science and news articles

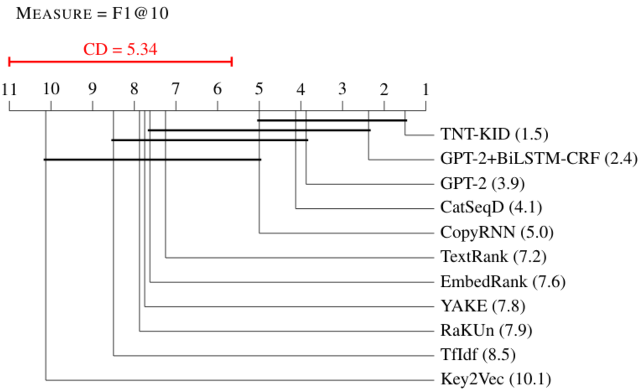
The approach was compared to a set of SOTA **supervised approaches**:

- keyword extraction as a **sequence-to-sequence generation** task: CopyRNN [Meng et al., 2019] and CatSeqD [Yuan et al., 2020]
- keyword extraction as **sequence labelling**: vanilla GPT-2 [Radford et al., 2019] and GPT-2 with a BiLSTM-CRF classification head [Sahrawat et al., 2020]

And to a set of SOTA **unsupervised approaches**:

- **graph-based** approaches: TextRank [Mihalcea and Tarau, 2004], RaKUn [Škrlj et al., 2019]
- **statistical** approaches: YAKE [Campos et al., 2020]
- **embedding-based** approaches: Key2Vec [Mahata et al., 2018], EmbedRank [Bennani-Smires et al., 2018]

Results - English



Experimental design - less resourced languages

- Additional testing of the TNT-KID model on four news datasets covering less resourced languages (Croatian, Estonian, Latvian, Russian)
- Additional experiments:
  - Problem: Datasets contain very few keywords per document
  - Solution: Improving the recall of the system by combining TNT-KID and TF-IDF based keyword extractor that can only return keywords that were manually tagged as keywords in the past. The new system returns constant k=10 keywords.

Results - less resourced languages

Model	P@5	R@5	F1@5	P@10	R@10	F1@10
Croatian						
TF-IDF(tm)	0.2226	0.4543	0.2988	0.1466	0.5888	0.2347
TNT-KID	0.3296	0.5135	0.4015	0.3167	0.5359	0.3981
BERT + BiLSTM-CRF	**0.4607**	0.4672	**0.4640**	**0.4599**	0.4708	**0.4654**
TNT-KID & TF-IDF(tm)	0.2659	**0.5670**	0.3621	0.1688	**0.6944**	0.2716
Estonian						
TF-IDF(tm)	0.0716	0.1488	0.0966	0.0496	0.1950	0.0790
TNT-KID	**0.5194**	0.5676	**0.5424**	**0.5098**	0.5942	**0.5942**
BERT + BiLSTM-CRF	0.5118	0.4617	0.4855	0.5078	0.4775	0.4922
TNT-KID & TF-IDF(tm)	0.3463	**0.5997**	0.4391	0.1978	**0.6541**	0.3037
Russian						
TF-IDF(tm)	0.1764	0.2314	0.2002	0.1663	0.3350	0.2223
TNT-KID	**0.7108**	0.6007	**0.6512**	**0.7038**	0.6250	**0.6621**
BERT + BiLSTM-CRF	0.6901	0.5467	0.5467	0.6849	0.5643	0.6187
TNT-KID & TF-IDF(tm)	0.4519	**0.6293**	0.5261	0.2981	**0.6946**	0.4172
Latvian						
TF-IDF(tm)	0.2258	0.5035	0.3118	0.1708	0.5965	0.2655
TNT-KID	0.6089	0.6887	**0.6464**	0.6054	0.6960	**0.6476**
BERT + BiLSTM-CRF	**0.6215**	0.6214	0.6214	**0.6204**	0.6243	0.6223
TNT-KID & TF-IDF(tm)	0.3402	**0.7934**	0.4762	0.2253	**0.8653**	0.3575

Table: Results on the EMBEDDIA media partner datasets.

Examples

Abe's 15-month reversal budget fudges cost of swapping people and butter for concrete and guns. The government of Shinzo Abe has just unveiled its budget for fiscal 2013 starting in April. Abe's stated intention was to radically reset spending priorities. He is indeed a man of his word. For this is a budget that is truly awesome for its radical step backward into the past a past where every public spending project would do wonders to boost economic growth. It is also a past where a cheaper yen would bring unmitigated benefits to Japan's exporting industries. None of it is really true anymore. Public works do indeed do wonders in boosting growth when there is nothing there to begin with. But in a mature and well-developed economy like ours, which is already so well equipped with all the necessities of modern life, they can at best have only a one-off effect in creating jobs and demand. And in this globalized day and age, an exporting industry imports almost as much as it exports. No longer do we live in a world where a carmaker makes everything within the borderlines of its nationality. Abe's radical reset has just as much to do with philosophy as with timelines. Three phrases come to mind as I try to put this budget in a nutshell. They are: from people to concrete,from the regions to the center and from butter to guns. The previous government led by the Democratic Party of Japan declared that it would put people before concrete. No more building of ever-empty concert halls and useless multiple amenity centers where nothing ever happens. More money would be spent on helping people escape their economic difficulties. They would give more power to the regions so they could decide for themselves what was really good and worked for the local community. Guns would most certainly not take precedence over butter. Or rather over the low-fat butter alternatives popular in these more health-conscious times. All of this has been completely reversed in Abe's fiscal 2013 budget. Public works spending is scheduled to go up by more than 15 percent while subsistence payments for people on welfare will be thrashed to the tune of more than 7 percent. If implemented, this will be the largest cut ever in welfare assistance. The previous government set aside a lump sum to be transferred from the central government's coffers to regional municipalities to be spent at their own discretion on local projects. This sum will now be clawed back into the central government's own public works program.

Predicted keywords: shinzo abe, japan, economy
True keywords: shinzo abe, budget

Quantum market games. We propose a quantum-like description of markets and economics. The approach has roots in the recently developed quantum game theory.

Predicted keywords: quantum market games, economics, quantum like description
True keywords: economics, quantum market games, quantum game theory

Conclusion

Conclusion

Main takeaways

- Supervised approaches perform much better than unsupervised approaches.
- Sequence labelling approaches perform better that sequence to sequence generation approaches.
- Autoregressive language modelling outperforms masked language modelling in low-resource scenarios.

Availability

Documentation and code available under the MIT license at:
`https://gitlab.com/matej.martinc/tnt_kid`
Trained news models available as dockers:

- Estonian and Russian: `https://gitlab.com/matej.martinc/tnt_kid_app`
- Latvian: `https://gitlab.com/boshko.koloski/tnt_kid_app_lv`
- Croatian: `https://gitlab.com/boshko.koloski/tnt_kid_app_hr`

References

References I

Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., and Jaggi, M. (2018).
Simple unsupervised keyphrase extraction using sentence embeddings.
In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 221–229, Brussels, Belgium. Association for Computational Linguistics.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020).
Yake! keyword extraction from single documents using multiple local features.
Information Sciences, 509:257–289.

Mahata, D., Kuriakose, J., Shah, R. R., and Zimmermann, R. (2018).
Key2Vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 634–639, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Meng, R., Yuan, X., Wang, T., Brusilovsky, P., Trischler, A., and He, D. (2019).
Does order matter? an empirical study on generating multiple keyphrases as a sequence.
arXiv preprint arXiv:1909.03590.

Mihalcea, R. and Tarau, P. (2004).
TextRank: Bringing order into text.
In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

References II

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019).
Language models are unsupervised multitask learners.
Technical report, OpenAi.

Sahrawat, D., Mahata, D., Kulkarni, M., Zhang, H., Gosangi, R., Stent, A., Sharma, A., Kumar, Y., Shah, R. R., and Zimmermann, R. (2020).
Keyphrase extraction from scholarly articles as sequence labeling using contextualized embeddings.
In *Proceedings of European Conference on Information Retrieval (ECIR 2020)*, pages 328–335, Lisbon, Portugal. Springer.

Škrlj, B., Repar, A., and Pollak, S. (2019).
RaKUn: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation.
In *International Conference on Statistical Language and Speech Processing*, pages 311–323, Ljubljana, Slovenia. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
In *Advances in neural information processing systems*, pages 5998–6008, Vancouver, Canada. Curran Associates, Inc.

Yuan, X., Wang, T., Meng, R., Thaker, K., Brusilovsky, P., He, D., and Trischler, A. (2020).
One size does not fit all: Generating and evaluating variable number of keyphrases.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7961–7975, Online. Association for Computational Linguistics.