Report from Summer NLP Conferences

SIGIR 2021 Doctoral Consortium and RANLP 2021

Vítek Novotný
witiko@mail.muni.cz

Faculty of Informatics, Masaryk University

September 23, 2021

Introduction

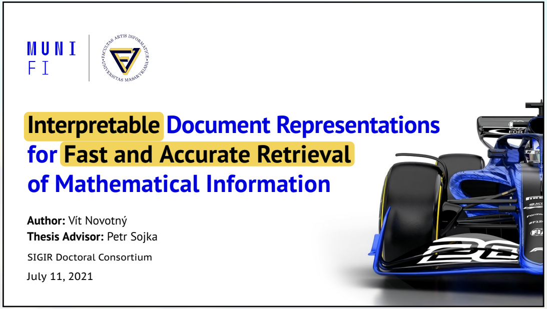
- This summer, I attended two online conferences:
  - SIGIR 2021 The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval Doctoral Consortium (July 11)
  - RANLP 2021 Recent Advances in Natural Language Processing (September 1–3)
- At both conferences, I presented a research paper:
  - SIGIR 2021 Interpretable Document Representations for Fast and Accurate Retrieval of Mathematical Information [3]
  - RANLP 2021 One Size Does Not Fit All: Finding the Optimal Subword Sizes for FastText Models across Languages [4]
- In this talk, I will report on both conferences and list useful study materials:
  - SIGIR 2021 Prerecorded tutorials about Natural Language Processing
  - RANLP 2021 New Czech language models by our colleagues from ZČU (Plzeň)

SIGIR 2021 Doctoral Consortium (July 11)

My Talk and Feedback from Mentors

- At the SIGIR 2021 Doctoral Consortium, I presented my doctoral research [3].
- After the talk, I discussed my research with two mentors:

Yongfeng Zhang	Assistant prof. at Rutgers University
Fiana Raiber	Researcher at Yahoo Research



Interpretable Document Representations for Fast and Accurate Retrieval of Mathematical Information

Author: Vít Novotný
Thesis Advisor: Petr Sojka
SIGIR Doctoral Consortium
July 11, 2021

- Here is the feedback from the mentors:
  - We have exploited the *interpretability* of our models in a visual browser of math documents [5]. We can further exploit it by producing counterfactual explanations.
  - We use shallow models to *jointly* model both natural language and math. We can model math using deep learning on graphs [8] and/or use code generation models to translate.

SIGIR 2021 Doctoral Consortium (July 11)

Tutorials

Besides talks, a number of prerecorded tutorials were released:

1. Addressing Bias and Fairness in Search Systems: Part 1
2. Addressing Bias and Fairness in Search Systems: Part 2
3. Beyond Probability Ranking Principle: Modeling the Dependencies among Documents
4. Deep Learning on Graphs for Natural Language Processing
5. Interactive Information Retrieval: Models, Algorithms, and Evaluation
6. Tutorial on Fairness of Machine Learning in Recommender Systems
7. Pretrained Transformers for Text Ranking: BERT and Beyond
8. Interactive Information Retrieval with Bandit Feedback
9. Stance Detection: Concepts, Approaches, Resources, and Outstanding Issues
10. Reinforcement Learning for Information Retrieval (tutorial website)

RANLP 2021 (September 1–3)

My Talk

- At RANLP 2021, I presented our research of hyperparameters in shallow fastText language models [4].
- A number of questions were asked:

1. To what degree do you think the optimal subword sizes are specific to fastText and to what degree are they universal? *They are even task-specific.*
2. How did you calculate inter-language distance using the suggested subword sizes? *Language l_i with s.s.s. j_i–k_i is represented by pt. $[j_i, k_i]$ in the Euclidean 2-space* $(\mathbb{R}^2, \ell_2)$.
3. Why do you think different languages have different optimal subword sizes? *They have different morpheme sizes, e.g. synthetic (Czech: 1–4) $<$ analytic (English: 4–5).*

One Size Does Not Fit All

Finding the Optimal Subword Sizes for FastText Models across Languages

Vít Novotný, Ayetiran E. F., Bačovský D., Lupták D., Štefánik M., and Sojka P.
{witiko,ayetiran,456662,dluptak,stefanik.m}@mail.muni.cz, sojka@fi.muni.cz

Faculty of Informatics, Masaryk University

September 3, 2021

By iconicbestiary / Freepik

RANLP 2021 (September 1–3)

Other Talks from Czech Scientists

Besides my talk and other programme, Czech scientists from ZČU (Plzeň) presented:

- Czert – Czech BERT-like Model for Language Representation (Sido, Pražák, Přibáň, Pašek, Seják, and Konopík) [9]
- Are the Multilingual Models Better? Improving Czech Sentiment with Transformers (Přibáň and Steinberger) [7]
- Multilingual Coreference Resolution with Harmonized Annotations (Pražák, Konopík, and Sido) [6]
- Evaluation Datasets for Cross-lingual Semantic Textual Similarity (Hercig and Král) [1]
- Transfer Learning for Czech Historical Named Entity Recognition (Hubková and Král) [2]

Conclusion

- This summer, I presented at two online conferences: SIGIR 2021 and RANLP 2021.
- At the SIGIR 2021 Doctoral Consortium, I presented my doctoral research [3] and received useful feedback by mentors from academia and industry.
- Besides talks, a number of useful prerecorded tutorials were released at SIGIR 2021.
- At RANLP 2021, I presented our research of fastText [4] and received a number of insightful questions from the audience.
- Besides my talk, our colleagues from ZČU (Plzeň) gave several talks about Czech NLP:
  - Czert – Czech BERT-like Model for Language Representation [9]
  - Are the Multilingual Models Better? Improving Czech Sentiment with Transformers [7]
  - Multilingual Coreference Resolution with Harmonized Annotations [6]
  - Evaluation Datasets for Cross-lingual Semantic Textual Similarity [1]
  - Transfer Learning for Czech Historical Named Entity Recognition [2]

Thank You for Your Attention!

Bibliography I

[1] Tomáš Hercig and Pavel Král. “Evaluation Datasets for Cross-lingual Semantic Textual Similarity”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by Ruslan Mitkov and Galia Angelova. Varna, Bulgaria: INCOMA Ltd., 2021, pp. 524–529. DOI: `10.26615/978-954-452-072-4_059`. URL: `https://ranlp.org/ranlp2021/proceedings-20Sep.pdf#page=548`.

[2] Helena Hubková and Pavel Král. “Transfer learning for Czech Historical Named Entity Recognition”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by Ruslan Mitkov and Galia Angelova. Varna, Bulgaria: INCOMA Ltd., 2021, pp. 576–582. DOI: `10.26615/978-954-452-072-4_065`. URL: `https://ranlp.org/ranlp2021/proceedings-20Sep.pdf#page=600`.

Bibliography II

[3] Vít Novotný. “Interpretable Document Representations for Fast and Accurate Retrieval of Mathematical Information”. In: New York, NY, USA: Association for Computing Machinery, 2021, p. 2705. ISBN: 9781450380379. DOI: 10.1145/3404835.3463269.

[4] Vít Novotný et al. “One Size Does Not Fit All. Finding the Optimal Subword Sizes for FastText Models across Languages”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by Ruslan Mitkov and Galia Angelova. Varna, Bulgaria: INCOMA Ltd., 2021, pp. 1068–1074. DOI: 10.26615/978-954-452-072-4_120. URL: https://ranlp.org/ranlp2021/proceedings-20Sep.pdf#page=1092.

Bibliography III

[5] Michal Petr. “Document Maps. visualization tool for semantic document representations”. Bachelor’s Thesis. Brno: Masaryk University, Faculty of Informatics, 2021. URL: `https://is.muni.cz/th/x86jd/`.

[6] Ondřej Pražák et al. “Multilingual Coreference Resolution with Harmonized Annotations”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by Ruslan Mitkov and Galia Angelova. Varna, Bulgaria: INCOMA Ltd., 2021, pp. 1119–1123. DOI: `10.26615/978-954-452-072-4_125`. URL: `https://ranlp.org/ranlp2021/proceedings-20Sep.pdf#page=1143`.

Bibliography IV

[7] Pavel Přibáň and Josef Steinberger. "Are the Multilingual Models Better? Improving Czech Sentiment with Transformers". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by Ruslan Mitkov and Galia Angelova. Varna, Bulgaria: INCOMA Ltd., 2021, pp. 1138–1149. DOI: 10.26615/978-954-452-072-4_128. URL: https://ranlp.org/ranlp2021/proceedings-20Sep.pdf#page=1162.

[8] Shaoyun Shi et al. "Neural Logic Reasoning". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 1365–1374. ISBN: 9781450368599. DOI: 10.1145/3340531.3411949.

Bibliography V

[9] Jakub Sido et al. "Czert. Czech BERT-like Model for Language Representation". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by Ruslan Mitkov and Galia Angelova. Varna, Bulgaria: INCOMA Ltd., 2021, pp. 1326–1338. DOI: `10.26615/978-954-452-072-4_149`. URL: `https://ranlp.org/ranlp2021/proceedings-20Sep.pdf#page=1350`.

MUNI

FACULTY

OF INFORMATICS