Text2TCS

universität wien

Transforming Term Extraction

Transformer-Based Approaches to Multilingual Term Extraction Across Domains

Authors: Christian Lang*, Lennart Wachowiak*, Barbara Heinisch, Dagmar Gromann (*equal contribution)

Contact: christian.lang@univie.ac.at, lennart.wachowiak@univie.ac.at

The Task: Automatic Term Extraction (ATE)

"We meta-analyzed mortality using random-effect models."

The Task: Automatic Term Extraction (ATE)

- Prior approaches:
  - Linguistic (e.g. POS patterns, phrase chunking…)
  - Statistical (e.g. TF-IDF, C-Value…)
  - Hybrid
- Traditional approaches generally operate on corpus or document level
- Recent approaches:
  - Machine Learning:
    - Topic modelling (e.g. Šajatović et al., 2019; Bolshakova et al., 2013)
    - Search Engine Queries , Wikipedia Lookups (Link Probability, Key Concept Relatedness) (e.g. Qasemizadeh and Handschuh 2014)
  - Deep Learning:
    - Word Embeddings (Amjadian et al., 2016), Neural Networks (e.g. Kucza et al., 2018, Gao and Yuan 2019)

The Task: Automatic Term Extraction (ATE)

- Machine Learning ATE strongly depends on provided features
  - ➢ Still requires linguistic pre-processing
  - ➢ Model is specific to a languages’ / domain’s feature set

- Deep learning addresses the issue of language / domain-dependence
  - ➢ Enables “featureless” end-to-end models for ATE (Gao and Yuan 2019)

Evaluating ATE performance

- Traditionally precision, recall and F1
- Total number of terms in texts often unknown, therefore only precision is reported
  - Hybrid methods improve precision, not recall
  - Recall is mostly dependent on manually set cut-off point
- For comparability, we chose F1 and report both precision and recall

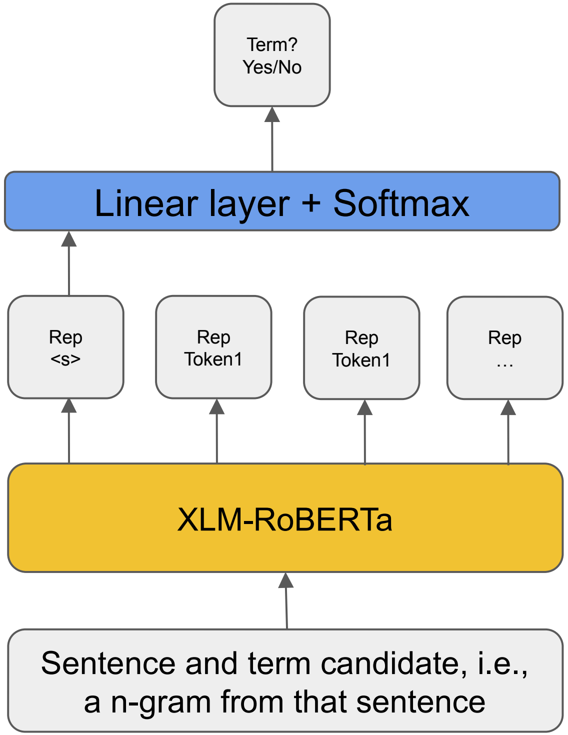
Other related work (excerpt)

- Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 3(2): 259–289.
- Nikita Astrakhantsev. 2018. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in scala. Language Resources and Evaluation,52(3): 853–872.
- Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Beatrice Daille. 2020. TermEval 2020: TALN-LS2N system for automatic term extraction. In Proceedings of the 6th International Workshop on Computational Terminology, pages 95–100, Marseille, France. European Language Resources Association.
- Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In Proceedings of the 6th International Workshop on Computational Terminology, pages 85–94, Marseille, France. European Language Resources Association.

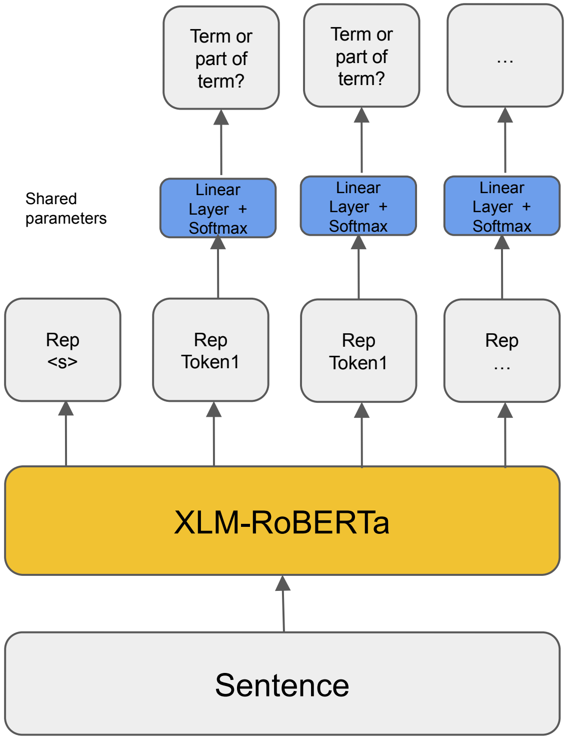
Methods: Comparing 3 Models

1. Sequence Classifier – XLM-RoBERTa (Conneau et al., 2020)
2. Token Classifier – XLM-RoBERTa
3. Neural Machine Translation (NMT) – mBART (Liu et al., 2020)

Methods: Sequence Classifier and Token Classifier

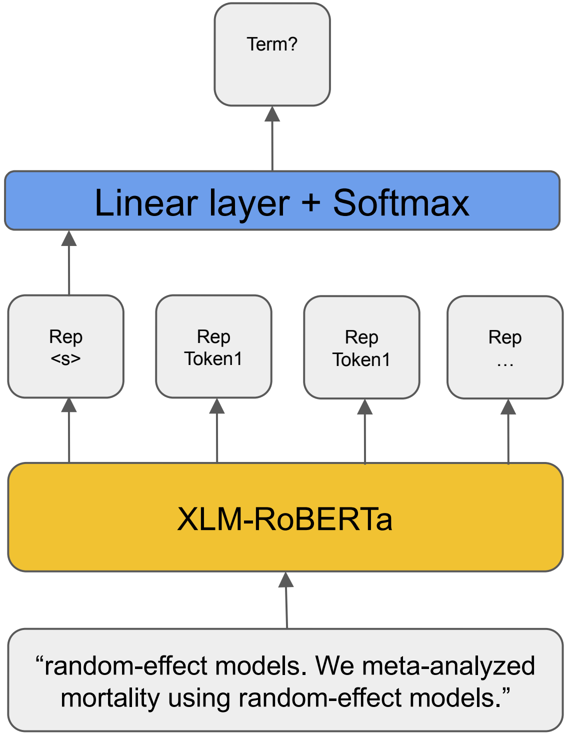


XLM-R Sequence Classifier



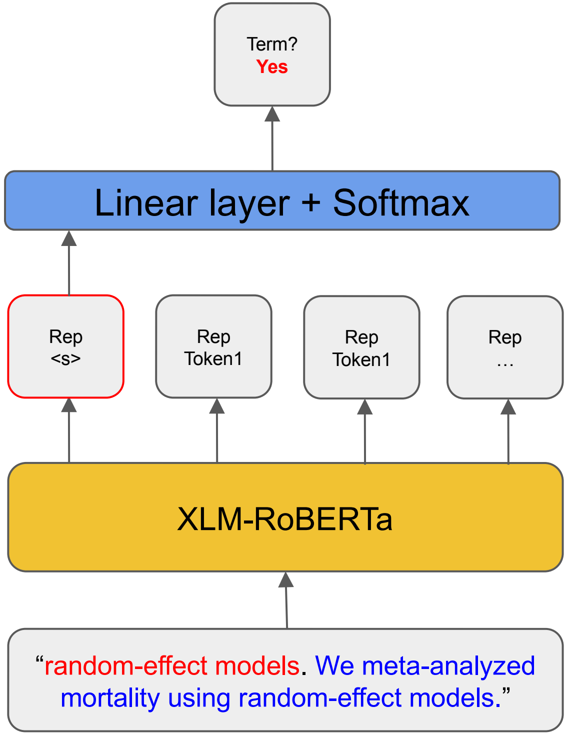
XLM-R Token Classifier

Sequence Classifier



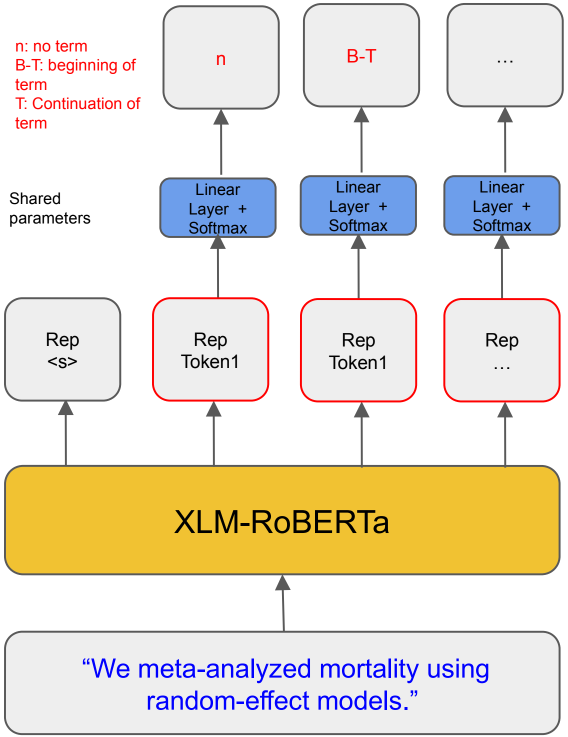
XLM-R Sequence Classifier

Sequence Classifier



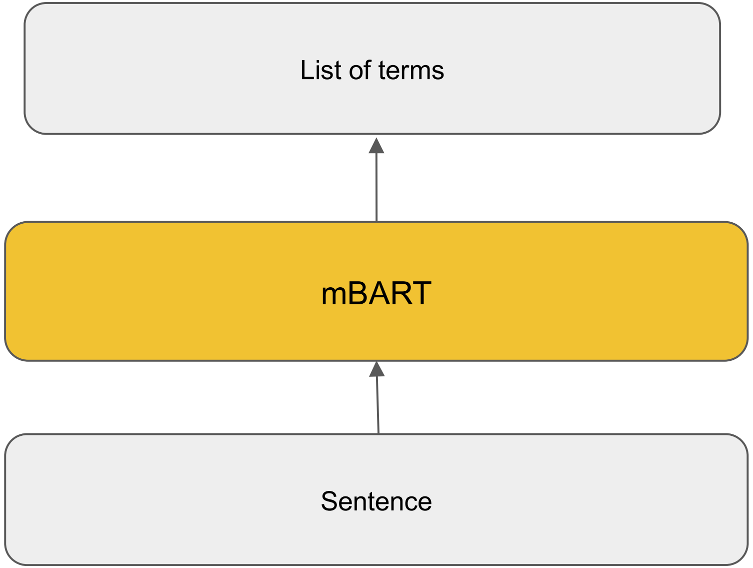
XLM-R Sequence Classifier

TokenClassifier



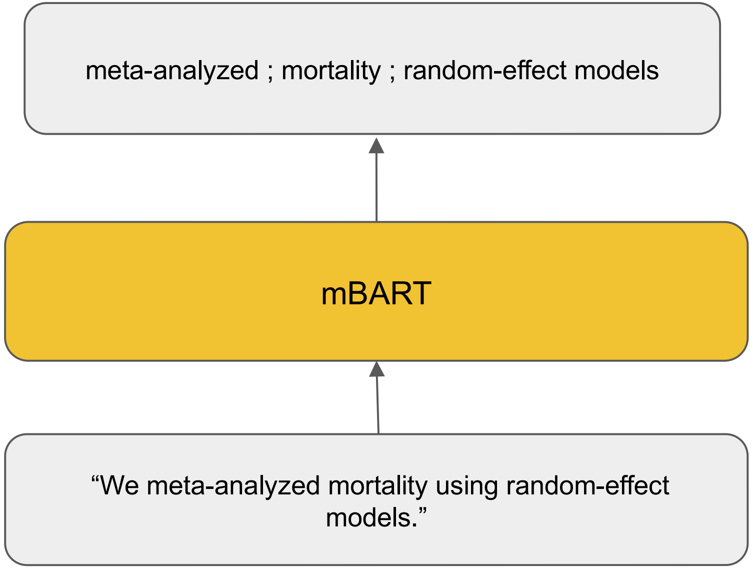
XLM-R Token Classifier

Methods: Neural Machine Translation



mBART NMT

Methods: Neural Machine Translation



mBART NMT

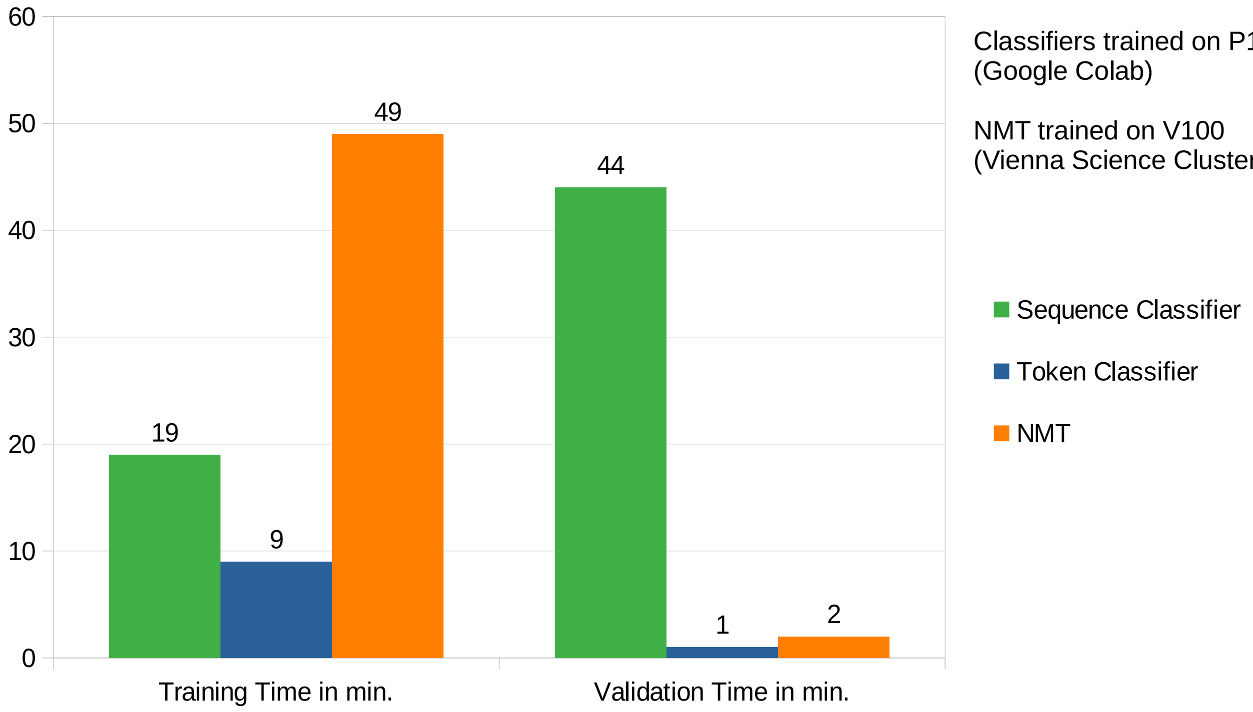
Datasets

- ACTER (Annotated Corpora for Term Extraction Research) (Rigouts et. al 2019)
  - Available at https://github.com/AylaRT/ACTER
  - 3 languages (EN, FR, NL), 4 domains
  - Approx. 200,000 words per language (50/25/25 Train, Val, Test split)
  - Terms as structured list of unique terms (no inline annotations)
  - Clear baseline thanks to TermEval 2020 Shared Task
  - Single annotator, experienced in ATE (not domain specialist)

Datasets

- ACL RD-TEC 2.0 (Quasemi Zadeh and Schumann, 2016)
  - Available at https://github.com/languagerecipes/acl-rd-tec-2.0
  - English, computational linguistics domain
  - Approx. 50,000 words (471 abstracts)
  - No suggested Train, Val, Test split
  - No readily available baselines
  - Offers inline annotations
  - Annotated by two domain experts

Average training time on ACTER



Classifiers trained on P100 (Google Colab)

NMT trained on V100 (Vienna Science Cluster)

Results: ACTER (Heart-Failure domain)

	Sequence Clf.			Token Clf.			NMT		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Rec.	Prec.	F1
EN	30.9	84.0	46.0	55.3	61.8	*58.3*	50.2	61.6	55.2
FR	34.6	79.0	48.1	65.4	51.4	*57.6*	55.0	60.4	*57.6*
NL	**40.4**	**91.5**	**58.0**	**67.9**	**71.7**	***69.8***	**60.6**	**70.0**	**64.9**

Results: ACL RD-TEC 2.0 (Evaluate on 20% of data)

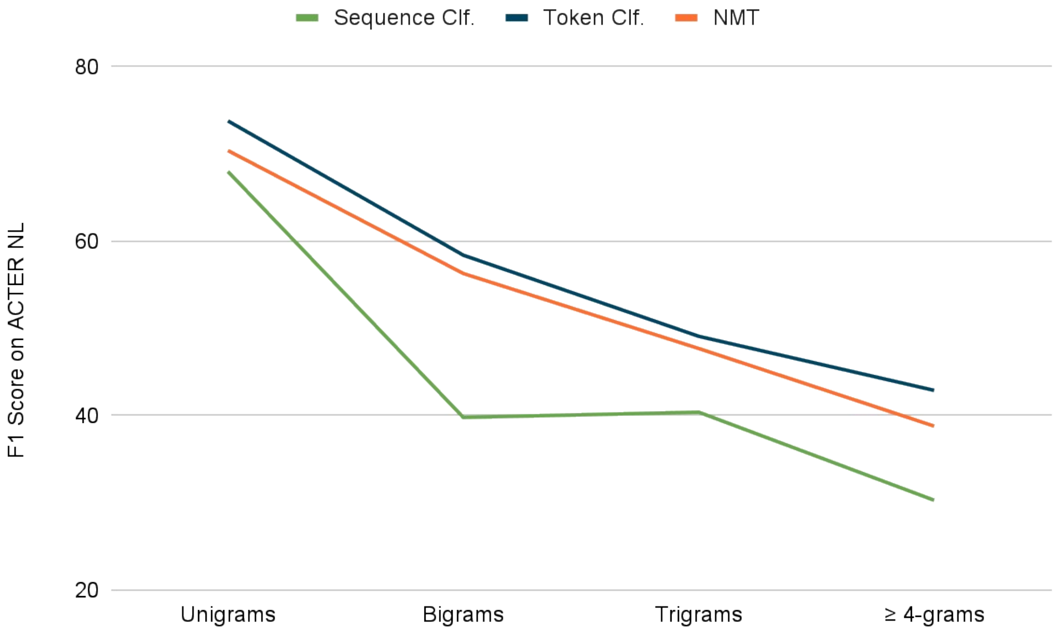
	Token Clf.			**NMT**		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Annotator 1	74.4	77.2	75.8	73.2	77.2	75.2
Annotator 2	80.1	79.3	80.0	79.4	80.7	80.0

Term-based analysis of false positives/negatives

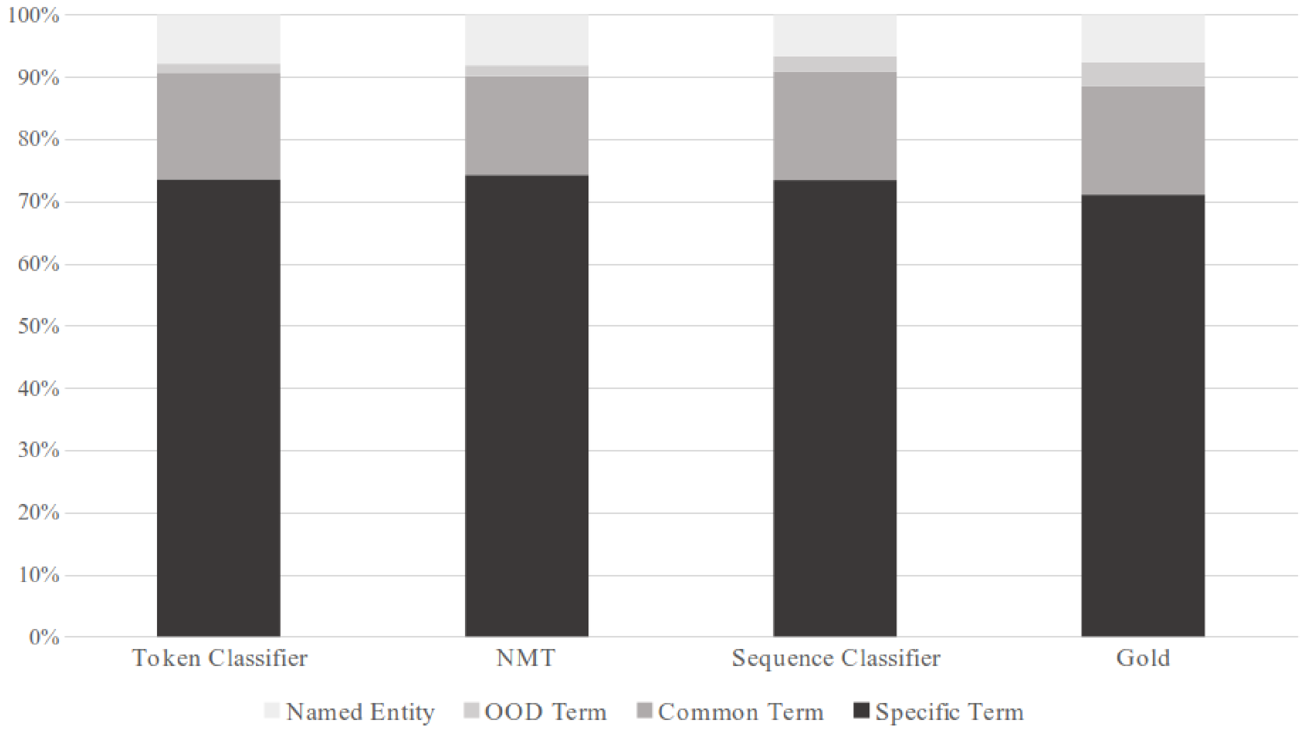
- Acronyms generally handled well by all models
  - Exception: e.g. “LV strain rate”, only parts extracted → False negatives
- Very long terms often not extracted
  - e.g. “resynchronization reverses remodeling in systolic left ventricular dysfunction”
  - Tendency of Token Classifier to split longer terms
- False positives are often nested terms (single-word terms in multi-word entities)
- NMT model might translate (parts) of terms (e.g. “toxicité cardiaque” extracted as “toxicity cardiaque”)

Results: Term Length

Single-word terms are easier to extract than multi-word terms



Term-based analysis (Terms by term-type on ACTER Dataset)



Results: Scores

- **Compared to previous SOTA baselines (ACTER / Termeval2020)**
    - +11.6 for EN
    - +9.5 for FR
    - +51.1 for NL

- **Strong Zero-Shot performance**
- **Higher recall than traditional methods**

Key Takeaways

- Comparison of 3 models working across languages and domains
- Token classifier and NMT strongly outperform previous approaches
- Models work with the limited context of single sentences
- Multi-word terms are more challenging than single-word terms

Discussion

- Training data quality matters more than quantity for pretrained models
  - Mixing datasets not trivial, as term definition varies between projects
- Discontinuous Entities, nested terms
- Single- or Multilingual Training
- Future Work:
  - Consider extra-sentential context
  - Publish own dataset
  - Improve multi-word term handling

References

Bolshakova, E., Loukachevitch, N., and Nokel, M. 2013. Topic Models Can Improve Domain Term Extraction. In: David Hutchison, et al., editors, Advances in Information Retrieval, volume 7814, pages 684–687. Springer Berlin Heidelberg, Berlin, Heidelberg.

Qasemi Zadeh, B. and Handschuh, S. 2014. Investigating Context Parameters in Technology Term Recognition. In: Proceedings of SADAATL 2014, pages 1–10, Dublin, Ireland.

Amjadian, E., Inkpen, D., Paribakht, T., and Faez, F. 2016. Local-Global Vectors to Improve Unigram Terminology Extraction. In Proceedings of the 5th International Workshop on Computational Terminology, pages 2–11, Osaka, Japan.

Qasemi Zadeh, B. and Anne-Kathrin Schumann .2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., and Stüker, S. 2018. Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In Interspeech 2018, pages 2072–2076, Hyderabad, India, September. ISCA.

Gao, Y. and Yuan, Y. 2019. Feature-Less End-to-End Nested Term Extraction. arXiv:1908.05426 [cs, stat], August. arXiv: 1908.05426.

Šajatović, A., Buljan, M., Šnajder, J., and Bašić, B. D. 2019. Evaluating Automatic Term Extraction Methods on Individual Documents. In Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 149–154, Florence, Italy. ACL.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2019. In: No uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. Language Resources and Evaluation, 54:385–418.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal,Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.Transactions of the Association for Computational Linguistics, 8:726–742.