Evaluation of a Sequence Tagging Tool for Biomedical Texts

Julien Tourille[1,2], Matthieu Doutreligne[1,3], Olivier Ferret[4]
Nicolas Paris[1,3,5], Aurélie Névéol[1], Xavier Tannier[5,6]

[1]LIMSI, CNRS, Université Paris-Saclay – [2]Univ. Paris-Sud
[3]WIND-DSI, AP-HP – [4]CEA, LIST – [5]Sorbonne Université – [6]INSERM, LIMICS

LOUHI - October 31, 2018

Motivation

- **Sequence tagging** e.g. NER $\rightarrow$ initial step which allows to perform more complex analysis in the biomedical domain (e.g. relation extraction or text classification)
- Several SOTA neural **sequence tagging models** have been published
  - LSTM-CRF (Lample et al. 2016)
  - CNN-CRF (Ma and Hovy 2016)
  - LSTM-CRF + Char. Att. (Rei et al. 2016)
  - LSTM + LM Objective (Rei 2017)
- **Lack of performance analysis** of such models in the biomedical domain
- **Lack of an off-the-shelf efficient implementation**

Motivation

- **Sequence tagging** e.g. NER $\rightarrow$ initial step which allows to perform more complex analysis in the biomedical domain (e.g. relation extraction or text classification)
- Several SOTA neural **sequence tagging models** have been published
  - LSTM-CRF (Lample et al. 2016)
  - CNN-CRF (Ma and Hovy 2016)
  - LSTM-CRF + Char. Att. (Rei et al. 2016)
  - LSTM + LM Objective (Rei 2017)
- **Lack of performance analysis** of such models in the biomedical domain
- **Lack of an off-the-shelf efficient implementation**

Motivation

- **Sequence tagging** e.g. NER $\rightarrow$ initial step which allows to perform more complex analysis in the biomedical domain (e.g. relation extraction or text classification)
- Several SOTA neural **sequence tagging models** have been published
  - LSTM-CRF (Lample et al. 2016)
  - CNN-CRF (Ma and Hovy 2016)
  - LSTM-CRF + Char. Att. (Rei et al. 2016)
  - LSTM + LM Objective (Rei 2017)
- **Lack of performance analysis** of such models in the biomedical domain
- **Lack of an off-the-shelf efficient implementation**

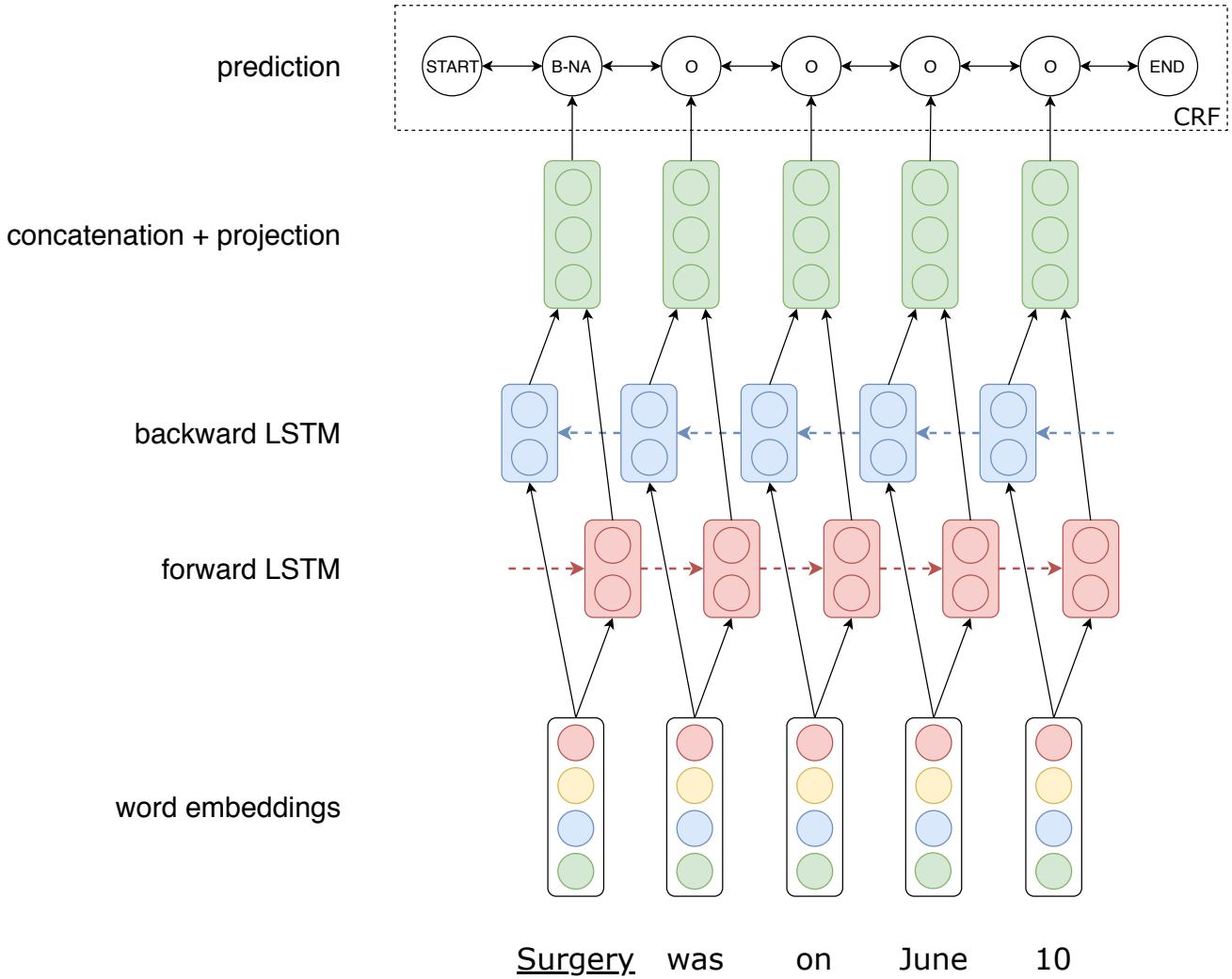
Contributions

- **Yet Another Sequence Tagger** (YASET): fast and accurate implementation of the LSTM-CRF model (Lample et al. 2016)
  - supports the use of **handcrafted features**
  - **easy-to-use** interface
- **Evaluation** on various biomedical corpora and the CoNLL 2003 corpus

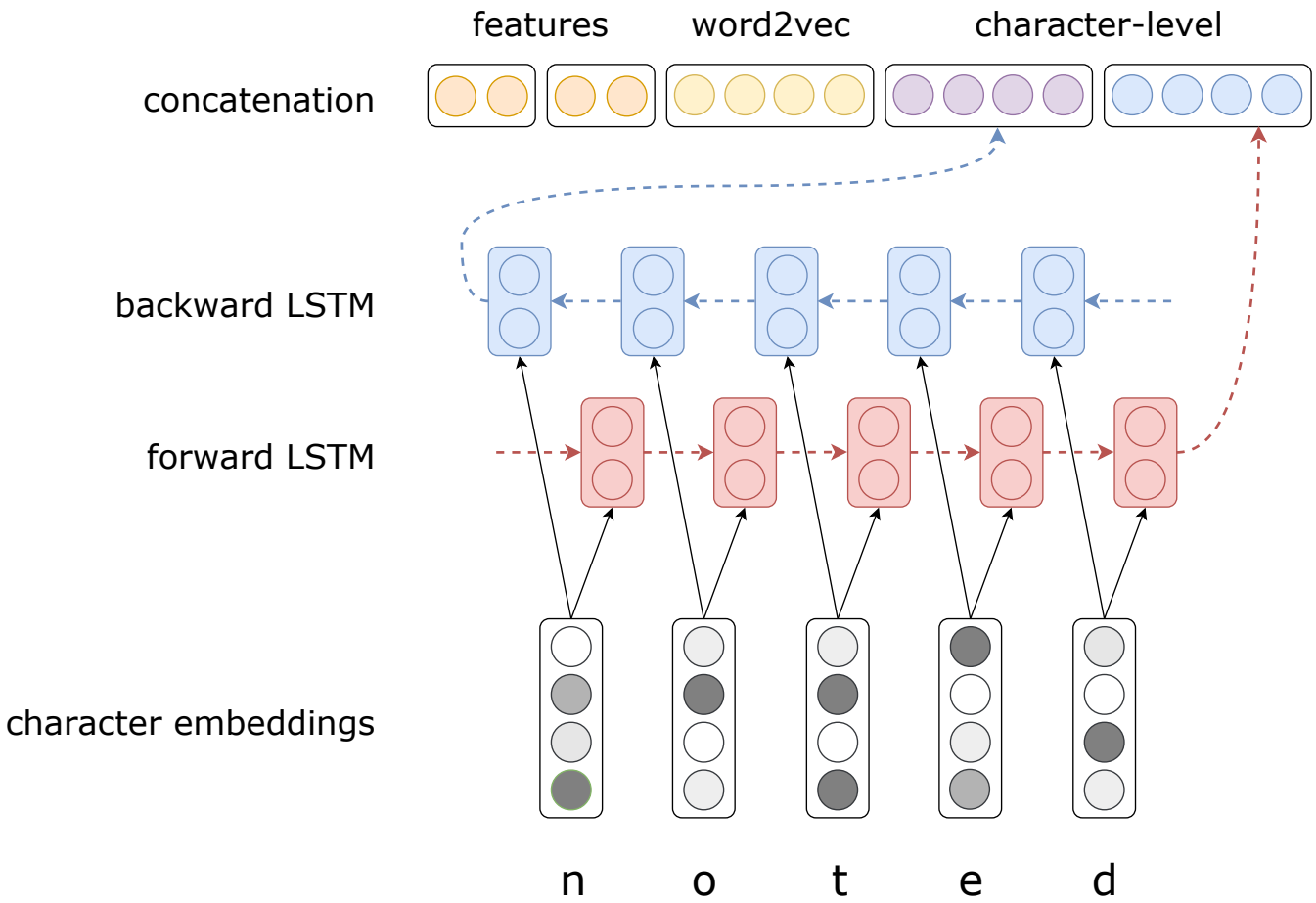
Contributions

- **Yet Another Sequence Tagger** (YASET): fast and accurate implementation of the LSTM-CRF model (Lample et al. 2016)
  - supports the use of **handcrafted features**
  - **easy-to-use** interface
- **Evaluation** on various biomedical corpora and the CoNLL 2003 corpus

YASET - Main Component



YASET - Input Embeddings



YASET - Training

- **Input data**: CoNLL format
- **Out of vocabulary tokens** $\rightarrow$ training of an *unknown* embedding by replacing singletons in the training corpus (according to probability p)
- Classical hyperparameters for neural models (early stopping, evaluation metric, hidden layer sizes, ...)
  $\rightarrow$ Located in one configuration file

Experiments

Two experiment sets

1. Performance evaluation on various corpora
   - MedPost (POS) (Smith et al. 2004)
   - NCBI disease (NER) (Doğan et al. 2014)
   - MERLoT (NER) (Campillos et al. 2018)
   - CoNLL (NER) (Sang and Meulder 2003)
2. Influence of training data size on performance (annotation is time-consuming and expensive): MERLoT corpus

Experiments - Corpora

Lang.	Corpus	# sent.	# tok.	# ann.		# en.
EN	MedPost (POS)[a]	6,701	182,319	182,319	(100%)	51
EN	NCBI disease (NER)[a]	7,279	151,005	11,350	(7,5%)	4
FR	MERLoT medical (NER)[b]	5,137	123,942	56,680	(46%)	19
FR	MERLoT PHI[d] (NER)[b]	25,599	177,158	31,723	(18%)	11
EN	CoNLL 2003 (NER)[c]	18,451	256,145	42,646	(17%)	4

[a] MEDLINE abstracts
[b] Medical reports
[c] News articles
[d] Protected Health Information

Experiments - Hyperparameters

- Pre-trained word embeddings
  - CoNLL 2003: Google News
  - NCBI disease and MedPost: PubMed corpus
  - MERLoT: 138,000 clinical reports in French
- Other hyperparameters selected with hyperopt (Bergstra et al. 2013) on the MERLoT de-identification dataset
  $\rightarrow$ Reused for other corpora
- Robustness to the random seed $\rightarrow$ 30 runs (Reimers and Gurevych 2017)

Experiments - Corpus Specific Results

Dataset	Model	F1
NCBI	Dang et al. (2018)	**84.41**
	Islamaj Dogan and Lu (2012)	81.80
	This paper	81.33
MERLoT med.*	Campillos et al. (2018)	81.40
	This paper	**82.87**
CoNLL 2003	Lample et al. (2016)	90.94
	Ma and Hovy (2016)	91.21
	Yang and Zhang (2018)	91.35
	Peters et al. (2018)	**92.22**
	This paper	87.31
MedPost*	Smith et al. (2004)	97.43
	This paper	**97.83**
MERLoT PHI	Grouin and Névéol (2014)	94.00
	This paper	**99.40**

* Corpora where the experimental set-up differed between our experiments and that of prior work. For MEDPOST, we used 51 categories instead of 63; for MERLoT med. we removed nested entities.

Experiments - Score Variability Across Runs

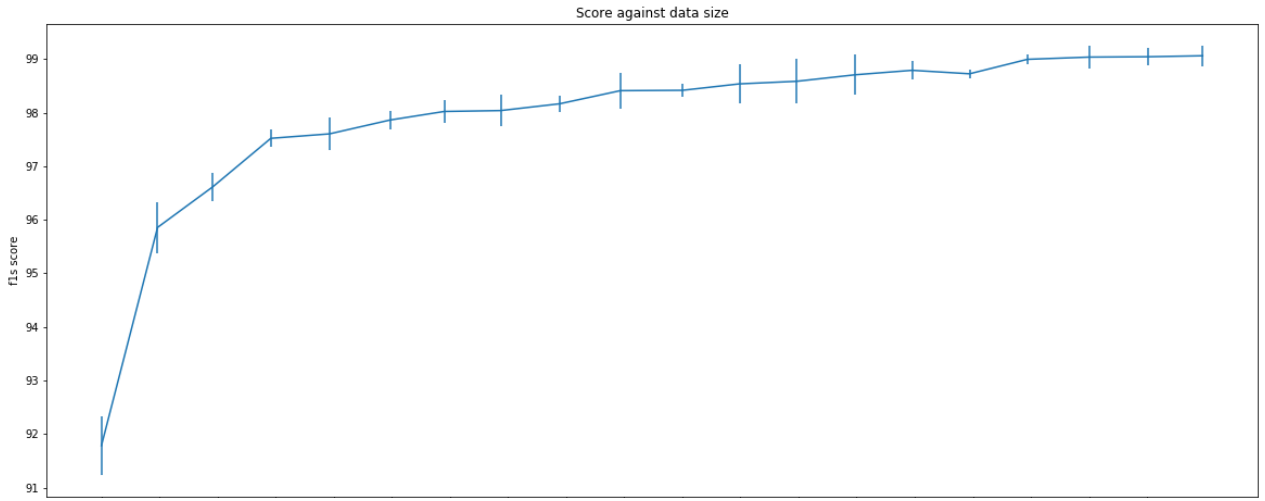
Task	Dataset	Mean F1	σ
NER	NCBI disease	81.33	0.83
NER	MERLoT medical	82.87	0.36
NER	CoNLL 2003	87.31	0.41
POS	MedPost	97.83	0.39
NER	MERLoT PHI	99.40	0.14

Performance of YASET on the 4 datasets.

Experiments - Training Data Size

Dataset	5%	10%	25%	50%	100%
MERLoT PHI	91.79 (0.55)	95.86 (0.47)	97.60 (0.30)	98.41 (0.33)	99.06 (0.19)

Performance (F1) against train set size (as percentage of the total training set indicated on the first row). Standard errors appear between parentheses.



These scores are computed over 6 training iterations per chunk. Vertical bars show the standard deviation σ.

Conclusion & Future Work

This paper

- Yet Another Sequence Tagger (YASET): accurate implementation of the LSTM-CRF model
- Hyperparameters optimized on one corpus can be transferred to others
- Performance improvement is logarithmic with the training data size

Future directions

- **Nested entity model** implementation (Campillos et al. 2018; Ju et al. 2018; Katiyar and Cardie 2018)
- **Other model** implementation (Ma and Hovy 2016; Rei and Yannakoudakis 2016)
- **Elmo** embeddings (Peters et al. 2018)

Evaluation of a Sequence Tagging Tool for Biomedical Texts

Julien Tourille[1,2], Matthieu Doutreligne[1,3], Olivier Ferret[4]
Nicolas Paris[1,3,5], Aurélie Névéol[1] and Xavier Tannier[5,6]

[1] LIMSI, CNRS, Université Paris-Saclay – [2] Univ. Paris-Sud
[3] WIND-DSI, AP-HP – [4] CEA, LIST – [5] Sorbonne Université – [6] INSERM

julien.tourille@limsi.fr

Bergstra, James, Daniel Yamins, and David Cox (June 2013). "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures". In: *Proceedings of the 30th International Conference on Machine Learning* (Atlanta, Georgia, US, June 16–21, 2013), pp. 115–123 (cit. on p. 12).

Campillos, Leonardo, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol (2018). "A French Clinical Corpus with Comprehensive Semantic Annotations: Development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)". In: *Language Resources and Evaluation* 52 (2), pp. 571–601 (cit. on pp. 10, 13, 16).

Dang, Thanh Hai, Hoang-Quynh Le, Trang M. Nguyen, and Sinh T. Vu (2018). "D3NER: Biomedical Named Entity Recognition Using CRF-biLSTM Improved with Fine-Tuned Embeddings of Various Linguistic Information". In: *Bioinformatics* 34.20, pp. 3539–3546 (cit. on p. 13).

Doğan, Rezarta Islamaj, Robert Leaman, and Zhiyong Lu (2014). "NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization". In: *Journal of Biomedical Informatics* 47, pp. 1–10 (cit. on p. 10).

Grouin, Cyril and Aurélie Névéol (2014). "De-Identification of Clinical Notes in French: Towards a Protocol for Reference Corpus Developpement". In: *Journal of Biomedical Informatics* 50, pp. 151–61 (cit. on p. 13).

Islamaj Dogan, Rezarta and Zhiyong Lu (2012). "An Improved Corpus of Disease Mentions in PubMed Citations". In: *Proceedings of the 2012 BioNLP Workshop* (Montreal, Canada, June 8, 2012). Association for Computational Linguistics, pp. 91–99 (cit. on p. 13).

Ju, Meizhi, Makoto Miwa, and Sophia Ananiadou (June 2018). "A Neural Layered Model for Nested Named Entity Recognition". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (New Orleans, Louisiana, USA, June 1–6, 2018). Association for Computational Linguistics, pp. 1446–1559 (cit. on p. 16).

Katiyar, Arzoo and Claire Cardie (June 2018). "Nested Named Entity Recognition Revisited". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (New Orleans, Louisiana, USA, June 1–6, 2018). Association for Computational Linguistics, pp. 861–871 (cit. on p. 16).

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (June 2016). "Neural Architectures for Named Entity Recognition". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California, USA, June 12–17, 2016). Association for Computational Linguistics, pp. 260–270 (cit. on pp. 2–6, 13).

Ma, Xuezhe and Eduard Hovy (Aug. 2016). "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Berlin, Germany, Aug. 7–12, 2016). Association for Computational Linguistics, pp. 1064–1074 (cit. on pp. 2–4, 13, 16).

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (New Orleans, Louisiana, USA, June 1–6, 2018). Association for Computational Linguistics, pp. 2227–2237 (cit. on pp. 13, 16).

Rei, Marek (July 2017). "Semi-supervised Multitask Learning for Sequence Labeling". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vancouver, Canada, July 30–Aug. 4, 2017). Association for Computational Linguistics, pp. 2121–2130 (cit. on pp. 2–4).

Rei, Marek, Gamal Crichton, and Sampo Pyysalo (Dec. 2016). "Attending to Characters in Neural Sequence Labeling Models". In: *Proceedings of the 26th International Conference on Computational Linguistics* (Osaka, Japan, Dec. 11–16, 2016), pp. 309–318 (cit. on pp. 2–4).

Rei, Marek and Helen Yannakoudakis (Aug. 2016). "Compositional Sequence Labeling Models for Error Detection in Learner Writing". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Berlin, Germany, Aug. 7–12, 2016). Association for Computational Linguistics, pp. 1181–1191 (cit. on p. 16).

Reimers, Nils and Iryna Gurevych (Sept. 2017). "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark, Sept. 7–11, 2017). Association for Computational Linguistics, pp. 338–348 (cit. on p. 12).

Sang, Erik F. Tjong Kim and Fien De Meulder (May 2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the 7th Conference on Natural Language Learning* (Edmonton, Canada, May 31–June 1, 2013). Association for Computational Linguistics (cit. on p. 10).

Smith, Larry, Tom Rindflesch, and W. John Wilbur (2004). "MedPost: A Part-of-Speech Tagger for BioMedical Text". In: *Bioinformatics* 20.14, pp. 2320–2321 (cit. on pp. 10, 13).

Yang, Jie and Yue Zhang (July 2018). "NCRF++: An Open-source Neural Sequence Labeling Toolkit". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Melbourne, Australia, July 15–20, 2018). Association for Computational Linguistics, pp. 74–79 (cit. on p. 13).