De-identification of free-text descriptions of suspected **adverse drug reactions** using deep learning

Eva-Lisa Meldau, Sara Hedfors Vidlin, Lucie Gattepaille, Henric Taavola, Lovisa Sandberg, Yasunori Aoki, G. Niklas Norén **Uppsala Monitoring Centre, Uppsala, Sweden**

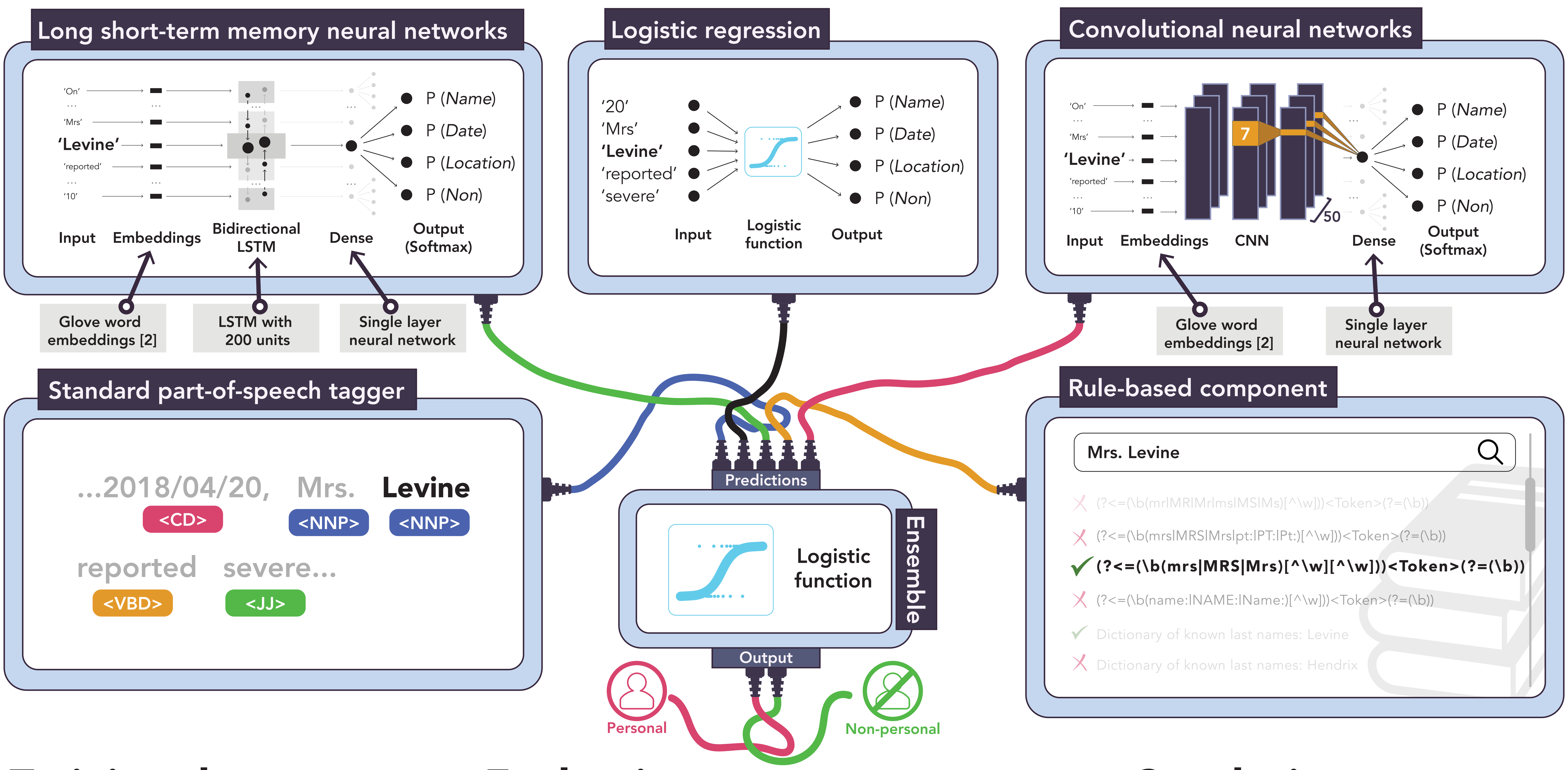
Free-text descriptions of suspected adverse drug reactions can be crucial in the analysis of the effects of medicines [1]...

...but might include personal identifiers that could compromise patient confidentiality.

Removing personal identifiers can allow sharing information between organizations working to protect patient safety.

For this study, we considered names, dates, and locations as personal identifiers, and all other tokens as not personal identifiers ('non-personal').

On [redacted], Mrs. Levi reported severe itching, soreness, and reddening of the genital area and an inability to sit while on therapy with dapagliflozin (started on 2018/04/10).



Training data:

Individual algorithms: 521 medical records from training set of the 2014 i2b2 de-identification challenge [3].

Ensemble: 3/4 of the 269 records in the validation set from the same challenge.

Evaluation:

i2b2

95% recall of personal identifiers

90% precision

On held-out 1/4 of the i2b2 2014 validation data set.

VigiBase

90% recall of personal identifiers

45% precision

On 300 narratives from VigiBase, the World Health Organization's global database of individual case safety reports [4].

Conclusions:

The proposed method can remove personal identifiers in free-text narrative descriptions of suspected adverse drug reactions with high recall and intermediate precision even though it was trained only on medical records. Fine-tuning, and possibly re-training parts of the method directly on reports of suspected adverse drug reactions can be expected to further improve performance.

Acknowledgments

The authors are indebted to the national centres who make up the WHO Programme for International Drug Monitoring and contribute reports to VigiBase. However, the opinions and conclusions of this study are not necessarily those of the various centres nor of WHO.

References

[1] Karimi, G, et al. Clinical stories are necessary for drug safety. Clin Med, 14(3):326-327, 2014.
[2] Pennington, J, et al. GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on EMNLP, 2014.
[3] Stubbs, A, Uzuner, O. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform, 58(Suppl): S20-29, 2015.
[4] Lindquist, M. VigiBase, the WHO Global ICSR Database System: Basic Facts. Drug Inf J, 42(5):409-19, 2008.
[5] Toutanova, K, et al. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. Proceedings of HLT-NAACL, 2003.

Uppsala Monitoring Centre

Uppsala Monitoring Centre (UMC)
Box 1051, SE-751 40 Uppsala, Sweden
+46 18 65 60 60 www.who-umc.org