KFU at CLEF eHealth 2017 Task 1: ICD-10 coding of English death certificates with recurrent neural networks

Miftahutdinov Z., Tutubalina E.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

This paper describes the participation of the KFU team in the CLEF eHealth 2017 challenge. Specifically, we participated in Task 1, namely "Multilingual Information Extraction - ICD-10 coding" for which we implemented recurrent neural networks to automatically assign ICD-10 codes to fragments of death certificates written in English. Our system uses Long Short-Term Memory (LSTM) to map the input sequence into a vector representation, and then another LSTM to decode the target sequence from the vector. We initialize the input representations with word embeddings trained on user posts in social media. The encoderdecoder model obtained F-measure of 85.01% on a full test set with significant improvement as compared to the average score of 62.2% for all participants' approaches. We also obtained significant improvement from 26.1% to 44.33% on an external test set as compared to the average score of the submitted runs.

Keywords

Cepidc, CLEF eHealth, Death certificates, Deep learning, Encoder-decoder model, Healthcare, Icd-10 codes, ICD-10 coding, Machine learning, Medical concept coding, Recurrent neural network, Sequence to sequence, Sequence-to-sequence architecture

References

[1] Miftahutdinov, Z., Tutubalina, E., Tropsha, A.: Identifying Disease-related Expressions in Reviews using Conditional Random Fields. In: Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog. Volume 1. (2017) 155-167

[2] Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In: EMNLP. (2015) 1753-1762

[3] Solovyev, V., Ivanov, V.: Knowledge-driven event extraction in Russian: corpusbased linguistic resources. Computational intelligence and neuroscience 2016 (2016) 16

[4] Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)

[5] Cho, K., Van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)

[6] Sutskever, I., Vinyals, O., Le, Q. V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104-3112

[7] Dos Santos, C. N., Gatti, M.: Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In: COLING. (2014) 69-78

[8] Liu, P., Joty, S. R., Meng, H. M.: Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. In: EMNLP. (2015) 1433-1443

[9] Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., De Luca, V., Jaggi, M.: SwissCheese at SemEval-2016 Task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. Proceedings of SemEval (2016) 1124-1128

[10] Tutubalina, E., Nikolenko, S.: Automated Prediction of Demographic Information from Medical User Reviews. In: International Conference on Mining Intelligence and Knowledge Exploration, Springer, Cham (2016) 174-184

[11] Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data, EACL (2017)

[12] Elman, J. L.: Finding structure in time. Cognitive science 14 (2) (1990) 179-211

[13] Schuster, M., Paliwal, K. K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45 (11) (1997) 2673-2681

[14] Graves, A., Fernandez, S., Schmidhuber, J.: Bidirectional LSTM networks for improved phoneme classification and recognition. Artificial Neural Networks: Formal Models and Their Applications-ICANN 2005 (2005) 753-753

[15] Neveol, A., Goeuriot, L., Kelly, L., Cohen, K., Grouin, C, Hamon, T., Lavergne, T., Rey, G., Robert, A., Tannier, X., et al.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (September 2016). (2016)

[16] Lavergne, T., Neveol, A., Robert, A., Grouin, C, Rey, G., Zweigenbaum, P.: A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage. BioTxtM 2016 (2016) 60

[17] Goeuriot, L., Kelly, L., Suominen, H., Nvol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth Evaluation Lab Overview. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2017)

[18] Nvol, A., Anderson, R. N., Cohen, K. B., Grouin, C, Lavergne, T., Rey, G., Robert, A., Rondet, C, Zweigenbaum, P.: CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings. (2017)

[19] Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5 (2) (1994) 157-166

[20] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., Schmidhuber, J.: LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems (2016)

[21] Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM networks. In: Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. Volume 4., IEEE (2005) 2047-2052

[22] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013) 3111-3119

[23] Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)

[24] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15 (1) (2014) 1929-1958

[25] Kinga, D., Adam, J. B.: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). (2015)

[26] Moen, S., Ananiadou, T. S. S.: Distributional semantics resources for biomedical text processing (2013)

[27] Van Mulligen, E., Afzal, Z., Akhondi, S. A., Vo, D., Kors, J. A.: Erasmus MC at CLEF eHealth 2016: Concept recognition and coding in French texts, CLEF (2016)

[28] Cabot, C, Soualmia, L. F., Dahamna, B., Darmoni, S. J.: SIBM at CLEF eHealth Evaluation Lab 2016: Extracting Concepts in French Medical Texts with ECMT and CIMIND, CLEF (2016)

[29] Mottin, L., Gobeill, J., Mottaz, A., Pasche, E., Gaudinat, A., Ruch, P.: BiTeM at CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction

[30] Dermouche, M., Looten, V., Flicoteaux, R., Chevret, S., Velcin, J., Taright, N.: ECSTRA-INSERM@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates, CLEF (2016)

[31] Zweigenbaum, P., Lavergne, T.: LIMSI ICD10 coding experiments on CepiDC death certificate statements, CLEF (2016)

[32] Zweigenbaum, P., Lavergne, T.: Hybrid methods for icd-10 coding of death certificates. EMNLP 2016 (2016) 96

[33] Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 32 (suppl1) (2004) D267-D270

[34] Spackman, K. A., Campbell, K. E., Cote, R. A.: SNOMED RT: a reference terminology for health care. In: Proceedings of the AMIA annual fall symposium, American Medical Informatics Association (1997) 640

[35] Aronson, A. R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium, American Medical Informatics Association (2001) 17

[36] Leaman, R., Islamaj Dogan, R., Lu, Z.: DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 29 (22) (2013) 2909-2917

[37] Limsopatham, N., Collier, N.: Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In: ACL. (2016)

[38] Grnarova, P., Schmidt, F., Hyland, S. L., Eickhoff, C: Neural Document Embeddings for Intensive Care Patient Mortality Prediction. arXiv preprint arXiv:1612.00467 (2016)

[39] Fries, J. A., Center, M.: Brundlefly at SemEval-2016 Task 12: Recurrent Neural Networks vs. Joint Inference for Clinical Temporal Information Extraction. Proceedings of SemEval (2016) 1274-1279

[40] Dernoncourt, F., Lee, J. Y., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association 24 (3) (2017) 596-606