Medical report information extraction

1 Introduction

The main project goal is to extract structured information out of medical reports to answer questions such as: what is the main diagnosis in the given report? Is the diagnosis confirmed or as of now only suspected? Could other diagnoses be excluded? When was the diagnosis made? To be able to answer such questions is of practical relevance for hospitals, for instance to assemble patient cohorts for retrospective studies.

To achieve this, we could extract the diagnosis by using an appropriate dictionary with associated keywords for all conditions. Next the classification task should be solved: how certain is this diagnosis in the given report? Is it excluded, confirmed or suspected? This is less trivial, since the context and the language around the keyword need to be analysed, for example:

- ”[condition] can be excluded”
- ”the symptoms suggest [condition]”
- ”evidence for [condition], based on . . . can be excluded”

In other words, we have a multi-class classification task.

Currently, this task is solved using a pipeline based on regular expressions and predefined keywords for one condition (Multiple Sclerosis). However, this approach only addresses a very limited amount of language patterns to assess the reliability of a diagnosis. Extending the amount of patterns would be a lot of manual work. Instead, machine learning based approaches should be explored. The proposed solution ideally should be done in an unsupervised or semi-supervised way, to reduce the annotation costs.

2 Proposed solution

The multiclass classification task could be formalized as unsupervised or semi-supervised sentiment analysis. The information about the reliability of a diagnosis is contained in the sentence and could be extracted with the help of semantic approaches and knowledge about sentence structure. To obtain the sentence structure, different parsers [Mutalikdesai et al., 2020, Chen and Manning, 2014] could be used to extract the nested structure of the subphrases and to identify the

subject and object. To compute the score for this subphrases or for the whole sentence different word and sentence embedding model could be used: BERT [Devlin et al., 2018], fasstext [Mikolov et al., 2018], word2vec [Mikolov et al., 2013]. These model should be pretrained before on the medical reports to catch all required lexicon. Score could be the cosine similarity between the given subphrases and small validation set of sentences (subphrases). Also to extend the validation set of the given sentences (subphrases) Nearest Neighbour search could be used — we could find the nearest neighbours in terms of some distance function to the given subphrases in the embedding space.

Another approach for this task is using Variational Autoencoders [Zeng et al., 2020, Wu et al., 2019] and different Neural Network architectures[Ruder and Plank, 2018, Han et al., 2020].

Major point: all medical reports are written in German.

For evaluation and validation a subset of the medical reports should be labelled (i.e what diagnosis is, what conditions are).

3 Proposed steps

1. Study the related literature;

2. Test the existing pipeline;

3. Propose a solution based on the studied literature and on the section (2);

4. Train or finetune all necessary models for the word or sentence embeddings;

5. Create the final algorithm and test the quality - an important point of the work will be the discussion about trade-off in using different kinds of models: proposed approach and existing pipeline.

4 General Information

Supervisor: Prof. Dr. Gunnar Rätsch D-INFK

Advisors: Marc Zimmermann, Rita Kuznetsova D-INFK

References

D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 740–750, 2014.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Y. Han, Y. Liu, and Z. Jin. Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. Neural Computing and Applications, 32(9):5117–5129, 2020.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems , 26:3111–3119, 2013.

T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

M. Mutalikdesai, M. Anagha, S. Srivastava, and L. Durgani. Unsupervised sentiment analysis for multiple subjects of interest using dependency parsing and linguistic heuristics, May 21 2020. US Patent App. 16/198,660.

S. Ruder and B. Plank. Strong baselines for neural semi-supervised learning under domain shift. arXiv preprint arXiv:1804.09530, 2018.

C. Wu, F. Wu, S. Wu, Z. Yuan, J. Liu, and Y. Huang. Semi-supervised dimensional sentiment analysis with variational autoencoder. Knowledge-Based Systems, 165:30–39, 2019.

Z. Zeng, W. Zhou, X. Liu, Z. Lin, Y. Song, M. D. Kuo, and W. H. K. Chiu. A variational approach to unsupervised sentiment analysis. arXiv preprint arXiv:2008.09394, 2020.