Classification of Syncope Cases in Norwegian Medical Records

Ildikó Pilán[*,†], Pål H. Brekke[‡], Fredrik A. Dahl[‡‡], Tore Gundersen[**], Haldor Husby[**], Øystein Nytrø[††], Lilja Øvrelid[*]

[*]Dept. of Informatics, University of Oslo, [†]Norwegian Computing Center,
[‡]Dept. of Cardiology, Oslo University Hospital Rikshospitalet,
[‡‡]Dept. of Health Services Research, Akershus University Hospital,
[**]Analysis Dept., Akershus University Hospital,
[††]Dept. of Computer Science, Norwegian University of Science and Technology
pilan@nr.no, paul.brekke@gmail.com
{Fredrik.A.Dahl,Tore.Gundersen,haldor.husby}@ahus.no,
nytroe@ntnu.no, liljao@ifi.uio.no

Abstract

Loss of consciousness, so-called *syncope*, is a commonly occurring symptom associated with worse prognosis for a number of heart-related diseases. We present a comparison of methods for a diagnosis classification task in Norwegian clinical notes, targeting syncope, i.e. fainting cases. We find that an often neglected baseline with keyword matching constitutes a rather strong basis, but more advanced methods do offer some improvement in classification performance, especially a convolutional neural network model. The developed pipeline is planned to be used for quantifying unregistered syncope cases in Norway.

1 Introduction

Neural methods have revolutionized the field of NLP, including the clinical domain in recent years. The amount of performance gain, however, may not always be proportional to the increased complexity and decreased transparency that their use might entail, especially in data-sparse domains and target languages. The limited availability of data and its linguistic characteristics, i.e. a high density of terminology, repetitions, abbreviations and misspellings (Allvin et al., 2011), are aspects that influence greatly the efficiency of the NLP methods applied. These have been compared to some extent in previous work (Baumel et al., 2018; Mascio et al., 2020; Karimi et al., 2017), however, they are often evaluated on the same (and often limited) openly available datasets (Pestian et al., 2007; Johnson et al., 2016). The real-world utility of various approaches in clinical text processing, especially for languages other than English, however, remains still to be investigated (Ching et al., 2018). Moreover, comparison to a simple rule-based baseline is often missing, leaving some uncertainty around the advantage of more advanced methods.

Starting from a close collaboration with Akershus University Hospital, we re-examine the question of the optimal methodological choice in the context of diagnosis coding in Norwegian clinical notes. Diagnosis codes are standard alpha-numeric codes representing a disease, a widely adopted scheme being ICD-10 (World Health Organization et al., 2004). ICD-10 codes are used for a variety of purposes, including hospital billing and reimbursement, population health statistics, and clinical research. Additionally, the re-use of structured health data in clinical decision support and risk assessment has also been suggested. ICD-10 coding is used as a most relevant classification of the reason for contact, underlying conditions or procedures related to the stay. A host of signs, events and observations are not coded. Syncope, or similar signs, may be regarded as secondary or irrelevant for a certain patient, and thus only mentioned but not coded. Clearly, accurate coding is important, but as a human process prone to error and biases, the quality of ICD-10 codes has been questioned. This is the case for *syncope* - a transient loss of conciousness typically due to insufficient blood flow to the brain - which was chosen as the use case for our study. A large study of Danish medical records (Ruwald et al., 2012) found that around a third of actual syncope records did not have the appropriate ICD-10 code. Since syncope can be an important sign of heart disease and a marker of elevated risk of death in certain conditions such as hypertrophic cardiomyopathy (Elliott et al., 2015), being able to retrieve information about patient's syncope events even when an ICD-10 code is not present, is crucial for better risk assessment. Also, this work constitutes a first step in the direction of an automatic

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 79–84

diagnosis coding system for Norwegian, which is currently not available. The research questions we investigate in this context are: (i) How do linear and neural models compare to a simple keyword matching baseline for binary automatic diagnosis code classification?; and (ii) How useful are pre-trained embeddings for this task? In what follows, we first describe our health record data and our pre-processing steps. We then compare three types of methods for syncope classification: a rule-based one relying on keyword matching, linear machine learning models and neural models. Besides estimating the amount of unregistered syncope cases in Norway, our processing and classification pipeline can also easily be re-used to train more generic diagnosis code classifiers.

2 Background

Since medical language is rather terminology-heavy, rule-based methods can often go a long way in clinical NLP tasks and are, therefore, still rather wide-spread (Koleck et al., 2019). Statistical approaches handle better linguistic phenomena such as synonyms, code-switching and negation, however, they are computationally more expensive, require resources and, in particular neural ones, are often less interpretable (Linzen et al., 2019). Moreover, neural methods substantially alleviate the burden of feature-engineering, but are considerably more challenging in terms of hyper-parameter tuning. Incorporating such models into clinical data processing pipelines is thus an advantage only if they can demonstrate a clear advantage over their simpler counterparts.

Dipaola et al. (2019) developed linear classifiers with manually and automatically selected n-grams as features for classifying syncope in Italian medical records. A frequent target of investigations has been the 2007 Computational Medicine Challenge (CMC) dataset, focusing on automatic ICD coding in radiology reports. Both rule-based (Farkas and Szarvas, 2008) and statistical methods (Crammer et al., 2007) including neural ones (Karimi et al., 2017), have been tested and sometimes compared on this data. Karimi et al. (2017) reported that the performance of a Support Vector Machine (SVM) with term frequency–inverse document frequency (TF-IDF) bag-of-words (BOW) features remained considerably below the results of a Convolutional Neural Network (CNN) with dynamic in-domain pre-trained Word2Vec embeddings with F1 scores of .65 and .81 respectively. A direct comparison across these works, however, is difficult given differences in the evaluation and data subset used.

More recently, using another dataset, MIMIC-III (Johnson et al., 2016), experiments presented by Baumel et al. (2018) indicated that neural methods outperform linear models for the same type of multi-class classification of ICD codes, although not always by a large margin. Mascio et al. (2020) also described a comparison between linear and neural models, but for different clinical binary classification tasks (e.g. status and negation prediction) and showed that recurrent neural networks tuned for their task performed on par with the more recent, transformer models (Devlin et al., 2019). Rule-based baselines were often not included in these recent studies (Karimi et al., 2017; Baumel et al., 2018; Mascio et al., 2020), the practical advantage of different approaches therefore remains somewhat unclear compared to methods based on heuristics.

3 Dataset

Our data consisted of de-identified discharge summaries from Akershus University Hospital Hospital. Half of the notes were diagnosed syncope cases (SYN), the other half were notes with a variety of diagnosis codes for patients with no recorded and coded history of syncope (NONS). The documents were authored between 2005–2016.[1] While patients in SYN were from a variety of departments, all NONS patients were from the Cardiology Department. Moreover, only patients who were ≥ 18 years old at the time of discharge were included. Table 1 provides an overview of the number of documents and their average length in number of tokens used in our dataset.

	SYN	NONS	ALL
# texts	501	500	1,001
Avg # tokens	667.51	546.52	607.02

Table 1: Overview of the dataset.

The notes contained free text where some structuring is present in the form of titled sections with information about e.g. diagnosis, family history and current status. There were, however, inconsistencies in the section titles as well as in the presence

[1] Data from the years 2017-2018 were reserved for evaluating the proportion of syncope cases with no diagnosis code.

and order of these sections. A previous study (Røst et al., 2020) using EHRs from Akershus University Hospital in a text classification task has also identified a need for improving interoperability when exporting such unstructured data.

4 Experimental Setup

The first pre-processing step consisted of tokenization with UDPipe (Straka et al., 2016). Diagnosis information reflecting the labels used for classification (SYN vs. NONS) was then removed from the documents using: (i) lexical matching for section title identification; and (ii) UDPipe paragraph information for determining section boundaries. We divided our data into three stratified splits: 70% of it reserved for training, 15% used as validation data for hyper-parameter tuning and the remaining 15% was set aside for testing. We compared a keyword matching baseline to two linear classifiers, a Logistic Regression (LR) classifier and an SVM, and to neural models, namely CNNs. These learning algorithms have been commonly and successfully used in previous NLP studies, including the clinical domain (Dipaola et al., 2019; Karimi et al., 2017).

Baseline with lexical matching (LEXM) We computed a baseline consisting of a simple lexical matching applied to the pre-processed documents using the term *synkope* 'syncope', which would find both its baseform and other derived forms without additional lemmatization. Whenever a document contained this term at least once, it was classified as belonging to the SYN class, and otherwise as NONS.

Linear models For training the linear models, we use scikit-learn (Pedregosa et al., 2011), and we employ Keras with Tensorflow (Abadi et al., 2016) as backend for the neural models. For both SVM and LR, we use BOW features extracted with a TF-IDF vectorizer. We perform a grid search for finding the optimal hyper-parameters on the validation data.

Neural models For the CNN, Word2Vec (Mikolov et al., 2013) embeddings were used as input representation to capture contextual similarity between words. We adopted a common CNN architecture (Kim, 2014) consisting of an input layer of 100 dimensions, a convolutional layer concatenating 100 filters of sizes 3 to 5, with rectified linear units, max pooling and a dropout of 0.5, followed by a fully connected softmax layer. We used binary cross-entropy loss, the Adam Optimizer, a learning rate of 0.001 and a batch size of 32. We trained for 10 epochs with early stopping based on validation accuracy and a patience of 2 epochs.

We experimented with different embedding initializations, inspired by Kim (2014): a randomly initialized one (W2V-R) and two where weights were based on pre-trained embeddings. In one case, weights were not trainable during the learning process (*static*) and in the other, we continued training these weights (*dynamic*). This type of transfer learning consisting of fine-tuning pre-trained embeddings for a specific task is often beneficial when the size of the available training data is small (Kim, 2014).

Pre-trained embeddings In the absence of pre-trained clinical embeddings for Norwegian, we compared two other types of pre-trained embeddings, both 100 dimensional Word2Vec skip-gram models trained with Gensim (Řehůřek and Sojka, 2010): (i) general language embeddings W2V-G trained on OCR-ed books, news and web corpora, namely model nr. 100 from the NLPL repository[2] (Fares et al., 2017); and (ii) domain-related embeddings W2V-M, which we trained on data from the *Norsk legemiddelhåndbok*[3] 'Norwegian drug manual'. The medical vocabulary of these disease and drug descriptions was closely connected to the clinical domain. We used default parameters for training W2V-M, but lowered minimum word count to 1 given the small data size.

5 Model Comparison and Error Analysis

In Table 2, we present the classification results for the approaches tested, where R-SENS represents *sensitivity*, i.e. recall for the positive class, SYN, and R-AVG is average recall for both classes. For the CNN models, an average of three runs (and standard deviation) is reported.

Lexical matching provided a rather high baseline, namely .80 accuracy, which suggests that similar terminology matching methods are worth testing and comparing to in terminology-rich domains such as the clinical one. Although we started from a strong baseline, we found that, with increasing computational complexity, performance improved somewhat. LR proved to be the best linear model (.86 accuracy) with L1 penalty, $C = 10$ with a

[2] http://vectors.nlpl.eu/repository/
[3] https://www.legemiddelhandboka.no/

Method	Features	Init	ACC	F1	PREC	R-SENS	R-AVG
LEXM	N/A	N/A	.80	.80	.80	.79	.80
SVM	BOW	N/A	.85	.85	.85	.87	.85
LR	BOW	N/A	**.86**	**.86**	**.86**	**.89**	**.86**
CNN	W2V-R	Random	**.92** (±.01)	**.92** (±.01)	**.91** (±.01)	**.93** (±.01)	**.92** (±.01)
	W2V-G	Static	.87 (±.01)	.87 (±.01)	.87 (±.01)	.85 (±.03)	.87 (±.01)
		Dynamic	.89 (±.00)	.89 (±.01)	.89 (±.01)	.88 (±.01)	.89 (±.01)
	W2V-M	Static	.68 (±.04)	.67 (±.05)	.70 (±.02)	.80 (±.07)	.68 (±.04)
		Dynamic	.77 (±.02)	.77(±.02)	.78 (±.02)	.78 (±.03)	.77 (±.02)

Table 2: Binary classification results on the test set.

liblinear solver as optimal hyper-parameters based on our grid search. The best neural model W2V-R, achieved .92 accuracy and a sensitivity of .93. To put these results into perspective, Dipaola et al. (2019) reported a sensitivity of .92 for an SVM-based syncope classification model for Italian.

For neural models, initializing embeddings randomly worked best. The number of in-embedding words was rather low in fact for both W2V-G and W2V-M, namely 51% and 27.5% respectively. In addition, W2V-G results might be influenced by a difference in domains. Models with W2V-M produced not only lower scores, but also more instability as standard deviation shows, likely due to the small vocabulary size (50K) and few in-embedding words. Dynamic embeddings showed improvements over static ones, especially for W2V-M, in line with previous findings (Kim, 2014).

We compared our methods also with McNemar's test (McNemar, 1962)[4] and found statistically significant difference in the misclassifications at $\alpha = 0.05$ only between the baseline and CNN-W2V-R ($p = 0.003$), but not between the other two model pairs, namely LR vs. baseline ($p = 0.163$) and LR vs. CNN-W2V-R ($p = 0.077$). Figure 1 shows the receiver-operating characteristic (ROC) curve for LR and CNN-W2V-R on the test set, which also shows a rather similar performance.

To gain a better understanding into what the best performing linear and neural models, LR and CNN-W2V-R respectively have learned, we inspected the 30 words which received the highest weights after training. These included for both models near-synonyms such as *svimmel* 'dizziness' and *bevissthetstap* 'unconsciousness' and even the English

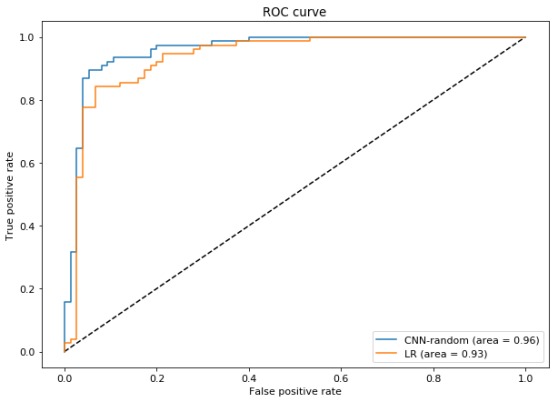

Figure 1: ROC curve for LR and CNN-W2V-R.

translation of the term (*syncope*). Yet another group of informative features described typical circumstances of syncope (e.g. *gulvet* 'floor'). LR also captured inflectional variants like *synkopert*, 'syncopated'. Both models' decisions relied thus on factors relevant to the target medical phenomenon.

Our error analysis revealed that around half of the NONS instances missclassified by both LR and CNN-W2V-R as SYN did contain mentions of 'syncope', but sometimes either as part of a patient's previous history of illnesses or with negation (*aldri synkopert* 'never syncoped'). Slightly more (60%) of misclassifications occurred for NONS texts, however, 23% (LR) and 38% (CNN-W2V-R) of these appeared to be unregistered syncope cases. Manually re-diagnosed data might therefore improve performance.

6 Conclusions

We described a set of experiments using keyword-matching as well as machine learning methods for the classification of syncope cases in Norwe-

[4]With binomial distribution given the small sample size.

gian clinical notes. Our results indicate that neural methods provide some advantage over a keyword baseline, but the latter performs surprisingly well, which indicates that terminological cues can be easily leveraged for such binary clinical text classification tasks in the absence of access to training data. This type of baseline constitutes thus a valuable starting and reference point for comparison to more advanced methods.

Future work includes hyper-parameter tuning of the neural models and comparing the generalizability of our models to new data, including different note types. We plan to use the developed models for quantifying the amount of unregistered syncope cases in Norway and to extend them to classify a variety of diagnostic codes. Embeddings trained on large Norwegian clinical data would be valuable to boost performance for both this and other tasks.

This work showcases a fruitful collaboration between an NLP research environment and a hospital. Aligning clinical data processing interests and needs is particularly important for smaller languages without publicly available data for both moving the clinical NLP research front forward and to bring findings closer to the clinical practice.

Ethics

According to Norwegian law, the project has been approved by the hospital's internal Privacy Ombudsman, ref. 2019_15.

Acknowledgments

This work was funded by the Norwegian Research Council and more specifically by BigMed, an IKT-PLUSS Lighthouse project.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, page 265–283, USA. USENIX Association.

Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravičius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgrén-Laine, Gunnar H Nilsson, Øystein Nytrø, et al. 2011. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*, 2:1–11.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. 2018. Multi-label classification of patient notes: case study on ICD code assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 409–416.

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387.

Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and Steven Carroll. 2007. Automatic code assignment to medical text. In *Biological, translational, and clinical language processing*, pages 129–136, Prague, Czech Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Franca Dipaola, Mauro Gatti, Veronica Pacetti, Anna Giulia Bottaccioli, Dana Shiffer, Maura Minonzio, Roberto Menè, Alessandro Giaj Levra, Monica Solbiati, Giorgio Costantino, et al. 2019. Artificial intelligence algorithms and natural language processing for the recognition of syncope patients on emergency department medical records. *Journal of Clinical Medicine*, 8(10):1677.

Perry M Elliott, Aris Anastasakis, Michael A Borger, Martin Borggrefe, Franco Cecchi, Philippe Charron, Albert Alain Hagege, Antoine Lafont, Giuseppe Limongelli, Heiko Mahrholdt, et al. 2015. 2014 ESC guidelines on diagnosis and management of hypertrophic cardiomyopathy. *Revista espanola de cardiologia*, 68(1):63.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-base ICD-9-CM coding systems. *BMC Bioinformatics*, 9(S3):S10.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Sarvnaz Karimi, Xiang Dai, Hamed Hassanzadeh, and Anthony Nguyen. 2017. Automatic diagnosis coding of radiology reports: A comparison of deep learning and conventional classification methods. In *BioNLP 2017*, pages 328–332, Vancouver, Canada,. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken. 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):364–379.

Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes. 2019. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 86–94, Online. Association for Computational Linguistics.

Quinn McNemar. 1962. *Psychological statistics*, volume 3. Wiley New York.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA.

Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby, Ingrid Andås Berg, and Øystein Nytrø. 2020. Identifying catheter-related events through sentence classification. *International Journal of Data Mining and Bioinformatics*, 23(3):213–233.

Martin Huth Ruwald, Morten Lock Hansen, Morten Lamberts, Søren Lund Kristensen, Mads Wissenberg, Anne-Marie Schjerning Olsen, Stefan Bisgaard Christensen, Michael Vinther, Lars Køber, Christian Torp-Pedersen, et al. 2012. Accuracy of the ICD-10 discharge diagnosis for syncope. *Europace*, 15(4):595–600.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

World Health Organization et al. 2004. *ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision*.