Instance-based learning for ICD10 categorization

Julien Gobeill[1-2] and Patrick Ruch[1-2]

[1] HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland
[2] SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland
julien.gobeill@hesge.ch

Abstract. In the framework of the CLEF 2018 eHealth campaign, we investigated an instance-based approach for extracting ICD10 codes from death certificates. The 360,000 annotated sentences contained in the training data were indexed with a standard search engine. Then, the k-Nearest Neighbors (k-NN) generated out of an input sentence were exploited in order to infer potential codes, thanks to majority voting. Compared to a standard dictionary-based approach, this simple and robust k-Nearest Neighbors algorithms achieved remarkable good performances (F-Measure 0.79, +13% compared to our dictionary-based approach, +70% compared to the official baseline). This purely statistical approach uses no linguistic knowledge, and could a priori be applied to any language with similar performance levels. The combination of the k-NN with a dictionary-based approach is also a simple way to improve the categorization effectiveness of the system. The reported results are consistent with inter-rater agreements (79-80%) for diagnosis encoding as achieved by trained professional staff. Any significant improvement should therefore be questioned.

Keywords: Information Extraction, Instance-based learning, International Classification of Diseases.

1 Introduction

The SIB Text Mining group [1], at the Swiss Institute of Bioinformatics in Geneva, has a long history of participation in TREC and CLEF campaigns, including TREC Genomics [2], TREC Medical Records [3], TREC Clinical Decision Support [4], or imageCLEF [5] and CLEF eHealth [6] tracks. In parallel, the group is currently involved in several translational medicine research projects, including the MyHealthMyData project (EU H2020 Programme), and SVIP-O (Swiss Variant Interpretation Platform for Oncology, funded by the Swiss Personalized Health Network Initiative or SPHN), and SPOP (Swiss Personalized Oncology and Pathology project, also funded by SPHN), three projects, which aims at helping clinicians to retrieve similar cases within clinical health records, including narratives, and genome-associated data modalities (e.g. gene variants). The group also leaded several local projects at the University and Hospitals of Geneva [7].

One of these projects, in 2016, dealt with the automatic categorization of clinical records into descriptors from the International Classification of Diseases (ICD-10). In

Swiss hospitals, ICD-10 codes are a posteriori assigned by trained curators to every episode of care, for medico economic purposes. In this local project, the available dataset contained 5 years of clinical records (between 40,000 and 50,000 per year), along with their assigned ICD-10 codes. The goal was to learn from the training data how to automatically reproduce the human ICD-10 encoding. We investigated an approach based on instance-based learning: the k-Nearest Neighbors algorithm (kNN). In such approaches, training data are used as a Knowledge Base (KB). For any unseen record, the most similar records contained in the KB are retrieved, then their encoding is used in order to infer potential encoding to the input record. The system obtained performances competitive with human curators (~80%), consistent with [7].

We also used this instance-based learning approach in the past with another categorization task related to biological curation: in this task, the goal was to reproduce the human curation of protein functions from scientific articles, using Gene Ontology (GO) concepts [8]. Such as ICD-10 encoding, GO curation involves thousands of annotatable concepts, and large already curated instances can be exploited in a Knowledge Base. We demonstrated that instance-based learning outperformed standard dictionary-based approach, in which annotatable concepts are mapped in the input text. Moreover, the continual growth of the available training data made the effectiveness of the instance-based learning approach improving across the time: the more the KB was populated, the more accurate was the system. Such approach achieved top-performing results in the BioCreative challenge in 2016 [9].

We capitalized on this experience in order to participate in the CLEF eHealth 2018 campaign, Task 1: Multilingual Information Extraction – ICD10 coding [10,11]. We had a limited amount of time and effort to invest in this campaign, thus the simplicity and robustness of a kNN was seen as an asset. We limited our participation to the French aligned dataset. Yet, our approach is applicable to potentially all languages, without prior linguistic knowledge.

2 Methods

Data. The gathered French training data contained 360,000 sentences from death certificates, annotated with 500,000 ICD10 codes. As training instances were short (4.06 words on average), the initial plan to exploit our local Knowledge Base containing full clinical records was quickly discarded. The test set contained 70,633 sentences to encode.

Similar search engine. The first step in the kNN algorithm for treating an input instance is to retrieve the *k* most similar instances in the Knowledge Base (the so-called *k* nearest neighbors). For such a task, we deployed a standard search engine and indexed all training sentences as if they were individual documents. For Information Retrieval, we used the Terrier platform [12]. We used no stemming nor stop words, and an Okapi BM25 weighting scheme [13].

Score computation. For a given input, once the *k* most similar instances from the KB are retrieved, our system simply exploits assigned ICD10 codes and uses majority voting: codes that are assigned more than *n* times are finally submitted. The hypothesis is that similar instances are more likely to share similar codes with the input text.

Additional dictionary-based module. In parallel, we exploited the manually curated ICD10 dictionary provided with the training data in order to map ICD10 concepts directly in the input sentence. A manually list of 40 stop words (such as cancer, or maladie) was designed in order to discard too general terms. A default score of *m* (between 1 and *k*) was assigned to mapped concepts in order to be combined with the kNN module.

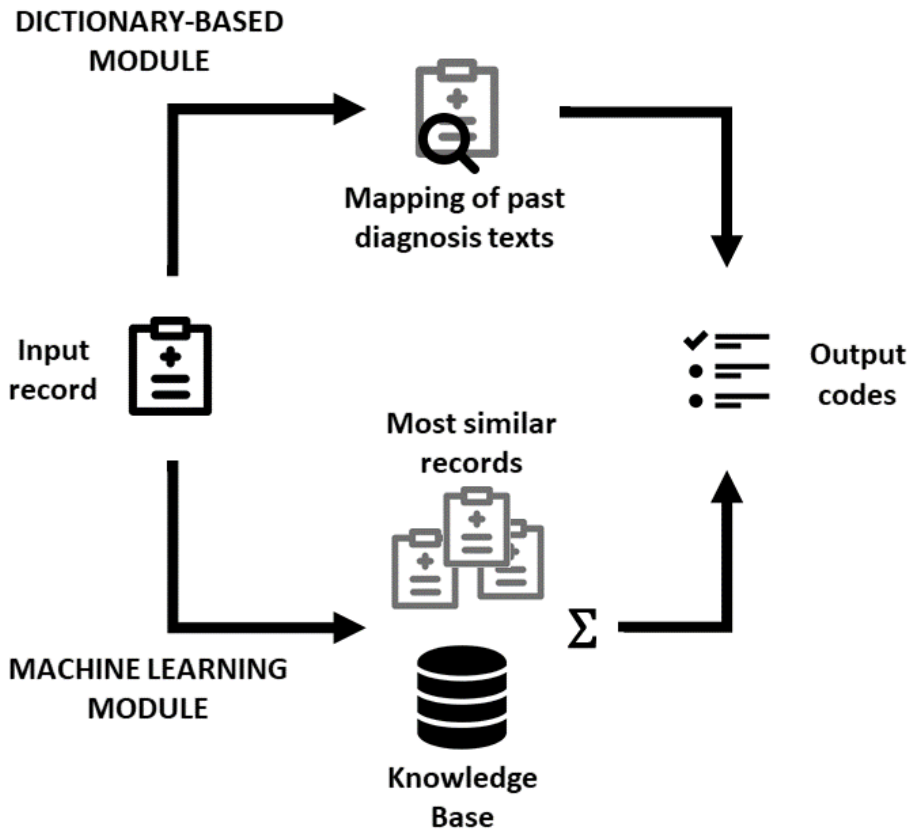

Fig. 1. Global architecture of the system. In the Machine learning module, *k* most similar records to the input are retrieved from the Knowledge Base, which contains training data ; then, IPC codes assigned to these records are aggregated and selected if they are present at least *n* times. In the dictionary-based module, past diagnosis texts are searched in the input and obtain a score of *m* when they are mapped. The list of output codes is combined from both modules.

3 Results

Setting of parameters. We discarded from the training data a set of 3,600 sentences for setting the *k* and *n* parameters. Macro Precision, Recall and Fmeasure were computed in order to compare settings. Results are presented in Tables 1 to 3.

Table 1. Macro Precision with different *k* and *n*. Maximum observed values are in bold.

	n=2	*n*=4	*n*=6	*n*=8	*n*=10
k = 5	0.90	0.96			
k = 10	0.80	0.89	0.94	**0.96**	
k = 15	0.74	0.84	0.89	0.92	0.94
k = 20	0.68	0.79	0.85	0.89	0.91
k = 25	0.64	0.75	0.81	0.86	0.88

Table 2. Macro Recall with different *k* and *n*. Maximum observed values are in bold.

	n=2	*n*=4	*n*=6	*n*=8	*n*=10
k = 5	0.69	0.53			
k = 10	0.76	0.68	0.61	0.53	
k = 15	0.78	0.72	0.67	0.62	0.58
k = 20	0.80	0.75	0.70	0.66	0.63
k = 25	**0.82**	0.77	0.72	0.69	0.66

Table 3. Macro F Measure with different *k* and *n*. Maximum observed values are in bold.

	n=2	*n*=4	*n*=6	*n*=8	*n*=10
k = 5	**0.78**	0.68			
k = 10	**0.78**	0.77	0.74	0.68	
k = 15	0.76	**0.78**	0.77	0.74	0.72
k = 20	0.74	0.77	0.77	0.76	0.75
k = 25	0.72	0.76	0.77	0.76	0.76

Combination with the dictionary-based module. Taken alone, the dictionary-based approach achieves on the same tuning set performances of P 0.71, R 0.68 and FM 0.69. We combined both modules with different values of *m*, with *k*=10 and *n*=2 (P 0.80, R 0.76 and FM 0.78), and we finally achieved with *m*=2 performances of P 0.79, R 0.79

and FM 0.79. These final setting were used in order to compute runs with the official test set.

Official results. The SIB Text Mining group submission achieves performances of P 0.76, R 0.76 and FM 0.76. Our official FM performance represent an improvement of +70% above the baseline, +20% above the participants mean, and +19% above the participants median

4 Discussion

The data used for these experiments are somehow relatively cleaner than standard EHR reports as they are significantly shorter. Basically, a realistic diagnosis encoding task would involve longer documents (surgery of anatomo-pathology report, discharge letter, etc). In the same spirit, potentially more than one report is generated by clinicians per episode of care, which is traditionally the time unit where encoding is performed. Further, it is important to question the stability of the data, which were provided and thus the stability of the resulting models. In particular, if we look at historical data used by [7], it is estimated that temporal drifts, which are intrinsically associated with diagnosis encoding (e.g. revision of billing/encoding guidelines, annual updates of ICD-10 by WHO and national authorities, etc), significantly reduce the validity of any generative models to a few months.

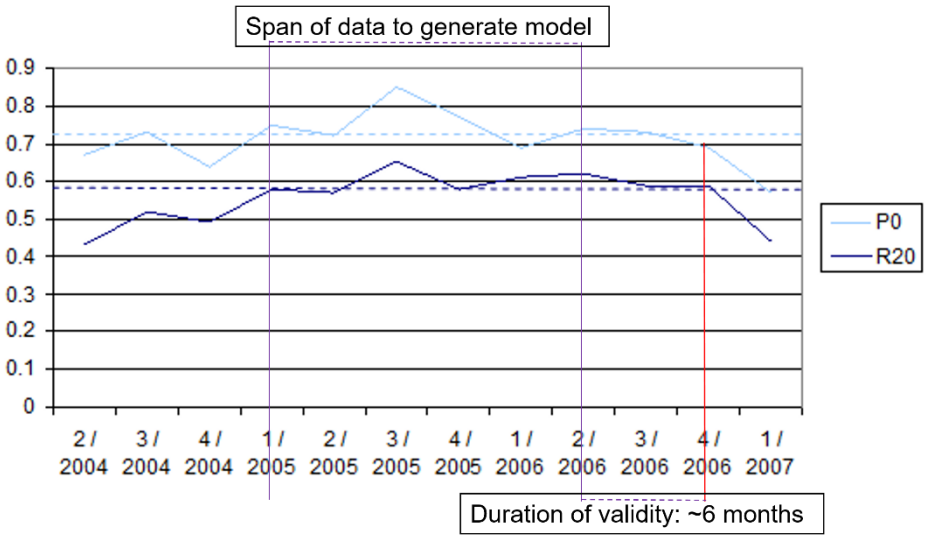

Fig. 2. Duration of the validity of data-driven categorization models. We see that the model generated with data acquired between quarter 1 in 2005 and quarter 2 in 2006 is performing well on posterior cases from quarter 2 of 2006 until quarter 4 of the same year. Beyond that time stand the results drops significantly.

Furthermore, the inter-encoder agreement achieved by trained professional in hospitals is in the range of 79-83%, see e.g. [7]. This score is the theoretical upper bound threshold achievable by automatic systems for such tasks. Any score higher is therefore likely to be caused by data biases or over-fitting phenomena.

5 Conclusion

Our simple and robust approach, mostly based on instance-based learning, but also combined with dictionary-based mapping, achieves remarkable performances for extracting ICD10 codes from death certificates sentences. Best observed F-Measure is 0.79, but different settings achieve high level of Precision (P 0.93 and R 0.53 with k=10 and n=8), or high level of Recall (R 0.82 and P 94 with k=25 and n=2). This purely statistical approach uses no linguistic knowledge, and could a priori be applied to any language with similar performances.

Acknowledgements

Results reported in this article have been partially supported by the HUG (University Hospitals of Geneva), which is a previous affiliation of the authors. They would not have been possible without the contribution of several HUG team members, including Drs. Robert Baud, Phedon Tahintzi, Claudine Bréant, Francois Borst, Rodolphe Meyer and Prof. Antoine Geissbühler.

References

1. http://bitem.hesge.ch/
2. Gobeill, J., Tbahriti, I., Ehrler, F., & Ruch, P. (2007). Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics. In TREC.
3. Gobeill, J., Gaudinat, A., Ruch, P., Pasche, E., Teodoro, D., & Vishnyakova, D. (2011). Bitem group report for TREC medical records track 2011. In TREC.
4. Gobeill, J., Gaudinat, A., & Ruch, P. (2015). Exploiting incoming and outgoing citations for improving Information Retrieval in the TREC 2015 Clinical Decision Support Track. In TREC.
5. Gobeill, J., Ruch, P., & Zhou, X. (2008, September). Query and document expansion with medical subject headings terms at medical imageclef 2008. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 736-743). Springer, Berlin, Heidelberg.
6. Mottin, L., Gobeill, J., Mottaz, A., Pasche, E., Gaudinat, A., & Ruch, P. (2016). BiTeM at CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction. In CLEF (Working Notes) (pp. 94-102).
7. Ruch, P., Gobeill, J., Tbahriti, I., & Geissbühler, A. (2008). From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. In AMIA Annual Symposium Proceedings (Vol. 2008, p. 636). American Medical Informatics Association.

8. Gobeill, J., Pasche, E., Vishnyakova, D., & Ruch, P. (2013). Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases. Database, 2013.
9. Mao, Y., Van Auken, K., Li, D., Arighi, C. N., McQuilton, P., Hayman, G. T., ... & Gobeill, J. (2014). Overview of the gene ontology task at BioCreative IV. Database, 2014, bau086.
10. Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Névéol, A., Ramadier, L., Robert, A., Palotti, J. & Zuccon, G. (2018). Overview of the CLEF eHealth Evaluation Lab 2018. CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September 2018.
11. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikán, L., Ramadier, L., Rey, G. & Zweigenbaum, P. (2018). CLEF eHealth 2018 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September, 2018.
12. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006, August). Terrier: A high performance and scalable information retrieval platform. In Proceedings of the OSIR Workshop (pp. 18-25).
13. Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models for information retrieval based on divergence from randomness.