Machine learning to detect ICD10 codes in causes of death

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, M. Teresa Martín-Valdivia, and L. Alfonso Ureña-López

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{plubeda,mcdiaz,maite,laurena}@ujaen.es

Abstract. In this paper we present our first participation as SINAI research group from the Universidad de Jaén at Task 1 ”Multilingual Information Extraction - ICD10 coding”. Our main goal is make a system based on Natural Language Processing (NLP) techniques to detect International Classification Diseases (ICD10) codes using different machine learning algorithms. First, we find all the possibles ICD10 codes mentioned in the text. Next, we calculate several measures of quality of these codes. With these metrics we trained different machine learning algorithms and we choose the best model to use in our system. Most of the techniques used are independent of the language, therefore our system are easily adaptable to other languages.

Keywords: Natural Language Processing, Named Entity Recognition, Text Classification, ICD10, French, Biomedical Text

1 Introduction

The goal of the task Multilingual Information Extraction [11] is to automatically assign International Classification Diseases (ICD10) codes to the text content of death certificates. This task can be treated as a named entity recognition and normalization task, but also as a text classification task. This builds upon the 2016 [13] and 2017 [12] tasks which already addressed the analysis of French biomedical text.

This task focuses on the Named Entity Recognition (NER), one of the basic problems of text mining [2]. Therefore, information extraction and Natural Language Processing (NLP) techniques will be developed to manage the information contained in a medical record [8]. There are many investigations related to the detection of medical entities [7,10,16]. Tools such as Unified Modeling Language MetaMap Transfer (UMLS MMTx) [14] is a configurable tool commonly used by system developers in biomedicine and Text Analysis and Knowledge Extraction System (cTAKES)[1] [15]. However, both tools are implemented to be used in English language collection. There are few automatic tools for identifying concepts in languages such as French, Hungarian or Italian.

[1] ctakes.apache.org

This group has a large experience participating in several tasks of other editions of CLEF task related to the medical domain [3,6,4,5] and by participating in this task we try to expand our horizons.

We have focused on a part of the task in French and have worked with the aligned data and dictionaries provided by the organization.

This paper is organized as follows: In the next section, we introduce the material provided by the organizers. Our method is described in section 3. In the section 4 showing the experiments performed and we are done with the conclusions and future work.

2 French aligned dataset

The CLEF e-Health 2018 Task 1 CepiDC Gold Standard Training data [9] composed of the following death certificates:

- Aligned causes 2006-2012: 65.843 death certificates and associated gold standard ICD10 codes
- Aligned causes 2013: 27.850 death certificates
- Aligned causes 2014: 31.690 death certificates

The files provided are in CSV format and contain one line for each death certificate. Each line of the file is structured as shown in the table 1 below.

Table 1. Fields in each row of the dataset

Field	**Description**
DocID	Death certificate ID
Year coded	Year the death certificate
Gender	Gender of the deceased
Age	Age at the time of death
Location of death	Location of death, according to the following categories: 1 - Home; 2 - Hospital; 3 - Private Clinic; 4 - Hopice, Retirement home; 5 - Public place; 6 - Other Location
Line ID	Line number within the death certificate
Raw text	Raw text entered in the death certificate
Int type	Type of time interval the patient had been suffering from coded cause, according to the following categories: 1 - minutes; 2 - hours; 3 - days; 4 - mouths; 5 - years
Int value	Length of time the patient had been suffering from coded cause
Cause rank	ICD10 code assigned
Standard text	Dictionary entry or excerpt of the raw text that supports the selection of an ICD10 code
ICD10	Gold standard ICD10 code

3 Methodology

In this section, we present the different strategies that we have followed in our participation in CLEF eHealth 2018 Task 1: Multilingual Information Extraction ICD10 coding.

We have created a tool to recognize ICD10 codes stored in dictionaries. This tool performs a Natural Language Processing (NLP) on the input text and tries to match with the dictionary terms.

3.1 System description

For both the input text and the dictionary, the system performs the following steps:

1. Normalize: to unify the form of terms we used unicode canonical decomposition (NFD)[1].
2. Tokenize: the tokenizer used for this task is WhitespaceTokenizer from the Python Natural Language Toolkit (NLTK)[2] Library. The whitespace tokenizer breaks text into terms whenever it encounters a whitespace character.
3. Stemmer: the algorithm used to make the stemmer in French is Snowball. Snowball[3] is a small string processing language designed for creating stemming algorithms for use in Information Retrieval.

The system returns a list of relevant ICD10 codes with a value between 0 and 1 corresponding with the percentage of success between the term detected within the text entered, the figure 1 shows an example of a system output.

Fig. 1. System output for cause of death with ID 100000

```
100000;Démence type Alzheimer à un stade sévère;F03; démence sévère;1
100000;Démence type Alzheimer à un stade sévère;G309;Alzheimer;1
100000;Démence type Alzheimer à un stade sévère;G309;démence type Alzheimer;1
100000;Démence type Alzheimer à un stade sévère;G309;démence type maladie Alzheimer sévère;0.8
100000;Démence type Alzheimer à un stade sévère;G309;démence type Alzheimer débutante;0.75
...
```

For the first three recognized concepts, the system returns 1 because the percentage of success of that term appears in the input text, for the case of "*démence type maladie Alzheimer sévère*" the system returns 0.8 because of the fact that of the 5 words that the recognized term matches, 4 appear in the text.

[2] https://www.nltk.org/ (last visited: May 30, 2018)

[3] https://snowballstem.org/

3.2 Measures for recognized terms

After developing the system proposed above, we created several measures to treat the words of the recognized terms:

- **Exact words**. This measurement is set to 1 if the term appears exactly the same in the cause of death text and assigns 0 otherwise.
- **Number of words between the first and last word detected**. This measure takes into account the total word number of the recognized term divided by the number of words between the first word of the term and the last word. With this measure, we are able to give a higher score to those terms with a higher number of words.
- **Weight according to position in the detected term**. Assigns a weight to each position of the recognized term:

  $$\sum 1/i$$

  where:
  i = word text position.
- **Weight by category ICD10**. The last measure used is related to the categories of ICD10 (the first letter in the code: A,B, ..., Z). We think that if the same category is detected many times, it will be more important and we should give it more weight. For this reason, we divide the number of occurrences of the detected category by a number of codes chosen β, in our case, $\beta = 10$ (that is, the system is configured to return the top 10 relevant codes).

4 Experiments

4.1 Linear regression for classification codes ICD10

Our goal was to use predictive analysis to improve the accuracy of the results. We use linear regression because is one of the most well known algorithms in machine learning.

In order to choose the best classifier we tested several machine learning techniques. The table 2 shows the results obtained. We trained with the 2013 and 2014 documents and tested with 2006-2012.

We used all the measures described above to build a training model on the data provided, for every cause of death and code detected in the collections: 2006-2012, 2013 and 2014. Then we predicted the results of the new ICD10 codes for the causes of death in 2015.

4.2 Adding measurements

We found that the sum of all measures calculated for each recognized term also worked well. To perform the sum, the value of the measures were normalized.

The results obtained with the 2013 and 2014 training collections are shown in the table 3.

Table 2. Results obtained for each classifier

Classifier	Precision	Recall	F1 score
Logistic Regression	0.4802	0.3369	0.396
Linear Regression	**0.5412**	**0.3863**	**0.4508**
SVC	0.4980	0.3381	0.4028
Linear Discriminant Analysis	0.4577	0.3477	0.3952
Decision Tree Classifier	0.4262	0.3272	0.3702
Gaussian NB	0.4499	0.3003	0.3602

Table 3. Results obtained using the sum of all measurements

Collection	Precision	Recall	F1 score
Causes of death 2013	0.7167	0.5324	0.6109
Causes of death 2014	0.7140	0.5088	0.5942

5 Conclusion and future works

In this work, the SINAI group has participated for the first time in the task of recognition of medical entities and classification in ICD10 codes in French. We have developed an automatic system that identifies medical terms using the dictionaries provided and almost independent of the language. We have created some measures for the treatment of the recognized concepts, with these measures, we have tested our system using different machine learning techniques to predict the final results. The technique with the best result obtained was Linear Regression.

For the next tasks of the CLEF, among others, we will want to improve our results train our systems by providing new information, will explore word embeddings models and we will incorporate new sources of knowledge such as ontologies specialized in the medical domain in French. In addition, other types of measurements that use bigrams and trigrams could be used. It will also be studied how much each measure used contributes to the final result.

Acknowledgments

This work has been partially supported by a grant from Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

1. Atkin, S.E.: Meta normalization for text (Apr 19 2005), uS Patent 6,883,007
2. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Briefings in bioinformatics 6(1), 57–71 (2005)

3. Díaz-Galiano, M.C., García-Cumbreras, M., Martín-Valdivia, M., Urena-López, L., Montejo-Ráez, A.: SINAI at ImageCLEFmed 2008. In: Peters, C., Ferro, N. (eds.) CLEF (Working Notes). CEUR Workshop Proceedings, vol. 1174. CEUR-WS.org (2008)
4. Díaz-Galiano, M.C., García-Cumbreras, M.Á., Martín-Valdivia, M.T., Montejo-Ráez, A.: Knowledge integration using textual information for improving imageclef collections. In: ImageCLEF, pp. 295–313. Springer (2010)
5. Dıaz-Galiano, M.C., Martın-Valdivia, M.T., Marıa, S., Jiménez-Zafra, A.A., López, L.A.U.: Sinai at clef ehealth 2017 task 3. In: CLEF 2017 Working Notes. CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017)
6. Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.: Query expansion with a medical ontology to improve a multimodal information retrieval system. Computers in biology and medicine 39(4), 396–403 (2009)
7. Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., Sinclair, G.: Exploiting context for biomedical entity recognition: From syntax to the web. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. pp. 88–91. Association for Computational Linguistics (2004)
8. Hahn, U., Romacker, M., Schulz, S.: How knowledge drives understanding—matching medical ontologies with the needs of medical language processing. Artificial Intelligence in Medicine 15(1), 25–51 (1999)
9. Lavergne, T., Névéol, A., Robert, A., Grouin, C., Rey, G., Zweigenbaum, P.: A dataset for icd-10 coding of death certificates: Creation and usage. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016). pp. 60–69 (2016)
10. Leser, U., Hakenberg, J.: What makes a gene name? named entity recognition in the biomedical literature. Briefings in bioinformatics 6(4), 357–369 (2005)
11. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikán, L., Ramadier, L., Rey, G., Zweigenbaum, P.: Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2018)
12. Névéol, 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: CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS. p. 17 (2017)
13. Névéol, A., Cohen, K.B., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., et al.: Clinical information extraction at the clef ehealth evaluation lab 2016. In: CEUR workshop proceedings. vol. 1609, p. 28. NIH Public Access (2016)
14. Osborne, J.D., Lin, S., Zhu, L.J., Kibbe, W.A.: Mining biomedical data using metamap transfer (mmtx) and the unified medical language system (umls). Gene Function Analysis pp. 153–169 (2007)
15. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17(5), 507–513 (2010)
16. Zhou, G., Zhang, J., Su, J., Shen, D., Tan, C.: Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 20(7), 1178–1190 (2004)