Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates

Pierre Zweigenbaum[1] and Thomas Lavergne[2]

[1] LIMSI, CNRS, Université Paris-Saclay
[2] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay
first.last@limsi.fr,
WWW home page: https://perso.limsi.fr/{pz,lavergne}/
F-91405 Orsay Cedex, France

Abstract. We present concept detection and normalization experiments on the French and English CLEF eHealth 2017 death certificate datasets. For this purpose, we start from our last published system, which relied upon dictionary projection and supervised multi-class, mono-label text classification using simple features. We extend this system in several dimensions with multi-label classification and new features, including an additional combination of dictionary and classifier. Because it only relies on the material provided by the task organizers, we could apply the same system to both the French and English datasets. Its results, registered as unofficial runs, equal or exceed those of the best submitted systems.

Keywords: concept normalization, death certificates, ICD-10,

1 Introduction

Most uses of medical information require its representation in a standardized, normalized form: this form is abstracted away from the natural language utterances that are generally used by health care professionals to record their observations. For instance, diagnoses for diseases in hospitals or for causes of deaths in death certificates are represented according to the International Classification of Diseases (ICD-10), a large classification maintained by the World Health Organization. Actually, causes of death in death certificates are initially described and recorded in natural language. National and international statistics, however, require their normalization according to ICD-10 classes, following specific rules established by the WHO. Efforts at automating this normalization process (also called *coding*) have been made to lighten the burden of human coders [6]. However, there is much room for improvement to obtain high-quality automated ICD-10 coding of death certificates.

ICD-10 coding of death certificates is an instance of concept detection and normalization in very short texts. It was proposed as a shared task in CLEF eHealth 2016 [5]. The organizers provided gold standard ICD codes for each input line of a death certificate. Systems were expected to produce the correct codes for each input line. CLEF eHealth 2016 participants addressed it both as an entity detection and normalization task based on dictionaries [7] (P=88.6, R=81.3, F=84.8%) and as a text classification task based on training examples [1] (P=88.2, R=65.5, F=75.2%). After the 2016 shared task we explored hybrid methods that combine dictionary and supervised machine learning [8] that rivalled with the best CLEF eHealth 2016 results.

The CLEF eHealth 2017 Shared Task [2] on ICD-10 coding of death certificates [4] brings two novelties with respect to the 2016 edition:

- Whereas the 2016 data only included French sources, a US English dataset was added in the 2017 data. Like the French dataset, it included a dictionary that maps terms to one or more ICD-10 codes.
- Whereas the 2016 data provided a line-level gold standard [3], line-level alignment was not performed on the English 2017 data. US certificates are provided in their original form, in which an easy mapping of input to ICD-10 codes is only available at the level of a full certificate. Therefore, for the US dataset, a specific sub-task was defined and evaluated as the production of ICD-10 codes from the global text of a death certificate. For consistency, the same sub-task was also defined for the French data, namely, producing ICD-10 codes from a full death certificate instead of individually producing codes for each line of a certificate.

Additionally, the 2016 task was continued for the French data: a line-level gold standard was provided for the French training data (which was extended with respect to the 2016 dataset), and line-level ICD codes were expected to be produced in that sub-task.

We built on these datasets to investigate the following points:

1. To examine the differences between the line-level sub-task and the certificate-level sub-task. A natural hypothesis is that training a classifier on certificate-level annotations will lead to a loss in classification quality. The question is how much. Besides, for longer (certificate-size) texts, the optimal combination of features might be different from that found for shorter (line-size) texts.
2. To examine the differences between the French certificate-level sub-task and the English certificate-level sub-task. Various reasons might make one more accurate than the other, including differences in the size of training data or dictionary, differences in intrinsic properties of each language such as inflection, differences in coding conventions, and differences in the distribution of ICD codes in the provided datasets.
3. In our previous work, as in [1], a mono-label classifier was used, which led to a low recall. Here we use a multi-label classifier, in the purpose of increasing recall while keeping a good precision.
4. We also wanted to continue exploring the combination of dictionary-based and supervised learning methods. Here we test an additional configuration not yet explored in our previous experiments, in which dictionary projection contributes features to the supervised classifier.

5. We tested a different representation of the age at the time of death, which better takes into account its ordered nature.

Being members of the organizing team, we prepared a system in parallel with the participants and submitted unofficial runs that we report in the results section.

2 Datasets

2.1 French data: lines and certificates

French data exists under two forms:

- Lines (aligned data): training data, test data and evaluation are performed at the level of each line of a certificate.
- Certificates (unaligned data): training data, test data and evaluation are performed at the level of each full certificate.

For each of these two forms, we work with three datasets:

- Training: 2016 training data: certificates of 2006–2012
- Development: 2016 test data: certificates of 2013
- Test: 2017 test data: certificates of 2014

We tuned our methods on the training data. Then we retrained them using the tuned parameters on the training+development data and applied them to the test data.

2.2 English data: certificates

English data exists under one form:

- Certificates (unaligned data): training data, test data and evaluation are performed at the level of each full certificate.

We split the provided training set into two parts:

- Training: training data except the last 666 certificates. This number was designed to obtain a similar ratio of examples in the training vs. development splits as in the French dataset.
- Development set: the last 666 certificates of the training data.

We used the training and development sets as for the French data to tune and train the system.

3 Methods

3.1 Background: our published methods

Our main methods were presented in [8]. We summarize them here to make the present paper self-standing. In all of these methods the input expressions first undergo a normalization step: case folding, diacritic removal, stop word removal, and stemming. Then we perform:

Dictionary projection: This relies on a left-to-right scan of the text with longest span exact match, without overlap. If multiple dictionary entries exist, with distinct ICD codes, for the same string, all these ICD codes are proposed when this string is matched in a text.

Dictionary calibration: Because dictionary entries are sometimes ambiguous, projection sometimes leads to false positives. We train a classifier to determine whether an ICD code assigned by the dictionary is likely to be correct or incorrect. Then we apply this trained classifier to the ICD codes produced for the test split: if it yields a negative answer for a code, we discard this code from the output of the dictionary. We call this a *calibrated* dictionary. The features given to the classifier are the produced ICD code and the bag of words obtained for the input text (line or certificate, depending on the dataset).

Supervised classification (linear SVM), with bag-of-word features.

The list of ICD codes and coding rules has evolved over the years; it is therefore useful to take into account the year coding was performed (coding year), which we encode as a set of interval features. For instance, a certificate with coding year 2011 receives features >2007, >2008, >2009, >2010, >2011, <2012, <2013, <2014, <2015, <2016. We observed that coding year is not relevant in the English dataset of CLEF eHealth 2017, because all its certificates belong to the same year (2015). It is therefore not included in the features for that dataset.

In our previous work [8] we used a mono-label classifier: because it produced at most one label per input, it had much lower recall than precision. Therefore we were interested in features that increase recall, possibly at the expense of precision, such as character n-grams: we used character trigrams, which made the classifier less sensitive to morphological variants and misspellings, together with token unigrams and coding year.

In the present work, we use a multi-label classifier, whose precision and recall are more balanced. In this context, character n-grams decrease precision more than they increase recall, and are thus a less useful feature to optimize F1-score. Instead we used word bigrams, with the aim of increasing precision. We thus start from the following features: bag of word unigrams (noted u), bag of word bigrams (b), coding year (y).

Union (and intersection) of the labels predicted by (calibrated) dictionary projection and classification. This is a crude way of combining these two methods, but it proved effective in our former experiments [8].

In [8] we applied these methods to the French-language, line-oriented dataset of CLEF eHealth 2016.

3.2 Additional variants

In addition to [8], we explore the following variants.

1. We provide the dictionary calibration classifier with an additional feature: to decide whether to keep a code detected by a dictionary entry, we provide the entry string itself (or, equivalently, the n-gram of the input text matched by this dictionary entry) on top of the associated ICD code and bag of unigram tokens.
2. We use multi-label classification instead of mono-label classification. We do this by training one classifier per label (ICD code) then selecting all codes that obtain a score better than a tuned threshold, as implemented in the OneVsRestClassifier meta-classifier of scikit-learn.
3. Beyond features for tokens unigrams (u) and bigrams (b), we add features for the age at death (which is provided rounded to the inferior 5 years). As for coding year, we encode age with intervals such as *before 30* or *after 30*. For instance, an age of 25 is encoded with the following set of features: $>$0, $>$5, $>$10, $>$15, $>$20, $>$25, $<$30, $<$35, ..., $<$100 (feature noted a).
4. We test the inclusion of dictionary projection results as features for the classifier (feature noted *fst*, after the finite-state transducer we use to store dictionaries and to match their entries to an input string).

Multi-label classification, with unigrams and bigrams, plus coding year for the French datasets ($u\ b\ y$, shortened as uby), was used to submit unofficial runs LIMSI 1 for each dataset. Its union with dictionary projection was submitted as unofficial runs LIMSI 2 for each dataset.

4 Results

Here we test the above methods on the CLEF eHealth 2017 training datasets to choose the best performing ones:

- French, lines (in CLEF terms, *aligned*)
- French, certificates (in CLEF terms, *raw*)
- English, certificates (in CLEF terms, *raw*)

As mentioned above, we first submitted for each dataset one basic run with a multi-label, supervised classifier (run1) and one with the union of this classifier and calibrated dictionary projection output (run2). We also submitted runs with variant features.

4.1 Lines: French

We compared various configurations on the training data. In this purpose, we trained the system on the training set (the CLEF eHealth 2016 training set) and tested it on the development set (the CLEF eHealth 2016 test set). The presented results are thus comparable to those published by CLEF eHealth 2016 participants (who indeed had less time to prepare their systems). For want of time, we did not use cross-validation on the training set, which would be more appropriate. Evaluation was performed internally using scikit-learn functions, which produce approximately the same results as the official CLEF eHealth evaluation program. The results are presented in Table 1, upper pane (FR, line).

- Dictionary with calibration (column *Cal Dict*) obtains 78.1% F-score: this is a basis for comparison. Note that this is still much below the best CLEF eHealth 2016 system (F=84.8), which was dictionary-based, but also used post-processing
- Multi-label classifier (column *Sup*): with the baseline features (*uby*), the supervised classifier is ten points above the calibrated dictionary and above the best CLEF eHealth 2016 system and our previously published results (F=85.9) which used the union of the mono-label classifier and of the calibrated dictionary [8].
- Union of dictionary and classifier (multi-label) modifies the F-score between +0.4pt (unigrams, bigrams, year, age: *ubya*) and -0.1pt (unigrams, bigrams, year, dictionary features: *ubyfst*). It may be that the more features are provided to the classifier, the less it can still improve with the simple addition of dictionary output as predictions.
- Age contributes 0.5pt F-score.
- Introducing dictionary projection results as features contributes about 0.6pt F-score to the classifier alone. They contribute 0.2pt F-score in union configurations, where dictionary projection results are also added as predicted labels directly.

Not displayed in the tables, we noticed that quite a few dictionary entries have more specific ICD-10 codes with an additional digit, whereas the training datasets never had such codes. We examined the impact of removing this extra digit from the dictionary. Indeed, this drastically improves dictionary projection by 6pt F-score. However, this has nearly no impact on F-score in supervised and union configurations, where it marginally decreases precision and increases recall; this shows that on the one hand, the supervised classifier independently learns to recognize the involved expressions (compensates for recall) and filters out the longer codes, because they are not seen in the training set (precision).

On the development set, the best configurations are close to each other. Nevertheless, we kept the one with the fixed dictionary and all features for running on the test set: unigrams, bigrams, coding year, age, dictionary projection results as features (*ubyafst*, run 3), and union with dictionary projection results (run 4).

Dataset	Features	Cal Dict	Sup	Union
FR, line	uby	78.14	88.23	88.46
	uby a	78.14	88.63	89.02
	uby fst	78.14	88.79	88.72
	uby a fst	78.14	**89.20**	**89.21**
FR. certif.	uby	78.72	87.62	88.63
	uby a	78.72	87.93	89.33
	uby fst	78.72	88.84	89.12
	uby a fst	78.72	89.26	**89.78**
EN, certif.	ub	74.03	89.85	90.79
	ub a	74.03	88.83	90.22
	ub fst	74.03	**90.97**	**91.36**
	ub a fst	74.03	90.29	90.92

Table 1. Experiments on the training and development sets: F1-score (%). Cal Dict = calibrated dictionary; Sup = supervised multi-label classifier

4.2 Certificates: French

We compared the same configurations as for the line-oriented dataset, using the same training and development sets, in their line versions. The results are shown in Table 1, middle pane (FR, certif.).

- Dictionary with calibration obtains an F1-score comparable to that on the line-level dataset.
- Surprisingly, the multi-label classifier looses very little (or even gains a little) in F1-score when applied to the certificate-level dataset. A closer inspection shows that its precision increases whereas its recall decreases. We assume its increase in precision is related to its better handling of *contextual codes*. We return to this point below.
- Union of dictionary and classifier (multi-label) increases the F-score by between 0.3pt (*ubfst*) and 1.5pt (*uba*). As can be expected, it is less useful when dictionary projection results are already provided as features to the classifier.
- Age contributes 0.7pt F-score.
- Introducing dictionary projection results as features contributes about 1.2–1.3pt F-score to the classifier alone. They contribute 0.5pt F-score in union configurations, where dictionary projection results are also added as predicted labels directly.

Fixing overlong codes in the dictionary leads to the same observations as in the line-level dataset.

The same overall observations hold as on the line-oriented dataset, therefore we selected the same configuration to run on the test set (*ubyafst*).

Comparison with line-level dataset Training with gold labels at the level of full certificates is a more difficult condition than training with gold labels at the level of individual lines because the identification of specific features in longer texts is made more difficult. However, training with gold labels at the level of full certificates is likely to help identify labels that depend on a larger context than a single line. For instance, this is the case for diagnoses that are coded differently depending on whether or not they are caused by a trauma, such as those in Chapter XIX (S00–T989, *Injury, poisoning and certain other consequences of external causes*) of ICD-10. Specifically, error analysis revealed that statements mentioning a *hemorrhagic shock* (*choc hémorragique*) should be coded T794 (*traumatic shock*) if a trauma is mentioned elsewhere in the death certificate, but are often confused with R571 (*hypovolemic shock*), which applies in the absence of a trauma. Certificate-level analysis can thus be beneficial for such codes: the best line-level classifier (*ubyafst*) over-predicts R571 and under-predicts T794, whereas the certificate-level classifier (*ubyafst*) is much closer to the true distribution of these two codes.

Besides, evaluating with gold labels at the level of full certificates is a more lenient condition than evaluating with gold labels at the level of individual lines: a label may be incorrectly attributed to a given line (false positive for line-level evaluation) but be present elsewhere in the same certificate (true positive for certificate-level evaluation). Additionally, since the line-level alignments were performed automatically [3], they contain a small percentage of errors: this may cause correctly predicted codes to be evaluated as false positives in the line-level evaluation, whereas certificate-level evaluation will count them as correct. These are the most likely explanations for the improved results of the dictionary on certificates compared to lines (+0.6pt).

4.3 Certificates: English

The only differences from French when applying the system to English data are the handling of apostrophes in word segmentation, the choices of stop word lists (based on those in NLTK) and stemmers (FrenchStemmer and EnglishStemmer from nltk.stem.snowball).

We compared various configurations on the training set. In this purpose, we split it into a test split (the last 666 certificates, ordered by DocID number) and a training split (the other certificates). For want of time, we did not use cross-validation on the training set, which would be more appropriate.

We recall that coding year is not relevant in the English dataset and is therefore not included in the features for this dataset.

- Dictionary with calibration obtains 4pt F1-score less than for French. This is not directly linked to their relative sizes, which is larger for English (170,282 lines) than for French (147,342 lines).
- The multi-label classifier obtains better results than on the French dataset: we return to this point below.
- Union of dictionary and classifier (multi-label) increases the F-score by between 0.4pt (ubfst) and 1.4pt (uba). As can be expected, it is less useful when dictionary projection results are already provided as features to the classifier.

- Age decreases F-score by 0.5pt. The reason why this is so in the English dataset remains to be investigated.
- Introducing dictionary projection results as features contributes about 1.1–1.5pt F-score to the classifier alone. They contribute 0.6–0.7pt F-score in union configurations, where dictionary projection results are also added as predicted labels directly.

Not displayed in the tables, fixing overlong codes in the dictionary improves dictionary projection by only 1pt F-score. However, it has nearly no impact on F-score in supervised and union configurations.

On the development set, the best configuration is that with the fixed dictionary, unigrams, bigrams, dictionary projection results as features (but not the age feature), and union with dictionary projection results. We retain it for running on the test set (ubfst, runs 3 and 4).

Comparison with the French certificate-level dataset Although the English dataset is smaller than the French certificate-level dataset, the results on the development set are better on the English dataset. This may be explained by the smaller number of codes and their different distribution in the English dataset.

4.4 Results on the test set

We show the results obtained for our four unofficial runs in Table 2.

Dataset	Run	Config	Precision	Recall	F-score
FR, line	SIBM-run1		83.46	**77.51**	**80.38**
	WBI-run1		77.98	75.06	76.49
	TUC-MI-run2		**87.44**	61.06	71.91
	LIMSI-run1	uby	86.51	86.47	86.49
	LIMSI-run2	U(uby,D)	85.37	**88.14**	86.74
	LIMSI-run3	ubyafst	**88.82**	85.60	87.18
	LIMSI-run4	U(ubyafst,D)	87.33	87.21	**87.27**
FR, certif.	SIBM-run1		**85.68**	**68.86**	**76.36**
	LIMSI-run1	uby	88.26	76.04	81.70
	LIMSI-run2	U(uby,D)	87.19	**78.36**	82.54
	LIMSI-run3	ubyafst	**90.41**	75.42	82.24
	LIMSI-run4	U(ubyafst,D)	89.08	77.25	**82.74**
EN, certificate	KFU-run1		89.30	81.12	**85.01**
	KFU-run2		89.11	**81.24**	85.00
	TUC-MI-run1		**94.02**	72.51	81.87
	LIMSI-run1	ub	90.86	76.53	83.08
	LIMSI-run2	U(ub,D)	89.90	80.13	84.73
	LIMSI-run3	ubfst	90.11	80.64	**85.11**
	LIMSI-run4	U(ubfst,D)	90.06	80.59	85.06

Table 2. Results on the test sets (%): best participants and 4 LIMSI runs

For the French datasets, the line-level classifier loses about 2pt F1-score from the development set to the test set in each of the four tested configurations. This shows that it did not overfit the training set.

The certificate-level classifier loses about 7pt F1-score, which is a much higher loss. This comes from a loss of 10pt in recall, whereas precision is maintained overall or even increased. This shows that this classifier overfits the training set.

Therefore, compared to the line-level classifier, the certificate-level classifier loses 10pt in recall and 5pt in F1-score. A similar loss was observed in the results of the best CLEF eHealth 2017 participant (SIBM: –10pt recall, –4pt F1-score).

On the English test set, the certificate-level classifier loses about 5–6pt in precision, recall and F1-score compared to the development set. Cross-validation tests on the training set should now be performed to check whether this is a general property of the training set.

Table 2 reproduces the best results obtained by CLEF eHealth 2017 participants. It shows that the methods presented here obtain better results on the French datasets and comparable results on the English dataset.

5 Conclusion and perspectives

We presented dictionary-based and supervised classification methods for ICD-10 coding of French and English death certificates. These methods use various combinations of dictionary and other features and obtain state-of-the-art results on the CLEF eHealth 2017 datasets.

We saw that certificate-level training and evaluation obtained similar results as line-level training and evaluation on the development set, and even improved for some context-dependent codes. However, on the test set, the certificate-level classifier proved less robust than the line-level classifier. This is an encouragement to study methods that can align the gold-standard codes with the input lines of the certificates in the training data, as was done in the CLEF eHealth 2016 dataset.

The results obtained on the English dataset are higher than those for the French dataset. This is likely due to its smaller set of codes (about one third). The addition of dictionary output as features further increased the performance of the classifier, while reducing the contribution of the union with dictionary output. The addition of the age of death helped in the French dataset, but not in the English dataset.

Perspectives for further work include, among others, the exploration of word embeddings and other neural methods, the introduction of other dictionary sources, better combination of dictionary output and supervised classifier beyond simple union, context-dependent coding of those ICD codes that require it, and automatic line-level alignment of input codes in certificates before training. The differences between the French and English datasets remain to be investigated further, as well as the potential for their joint usage.

References

1. Dermouche, M., Looten, V., Flicoteaux, R., Chevret, S., Velcin, J., Taright, N.: ECSTRA-INSERM @ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates. In: CLEF 2016 Online Working Notes. CEUR-WS (2016)
2. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth evaluation lab overview. In: CLEF 2017 - 8th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS), Springer (Sep 2017)
3. 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. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), http://aclweb.org/anthology/W16-5107
4. 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 (Sep 2017)
5. Névéol, A., Cohen, K.B., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., Zweigenbaum, P.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: CLEF eHealth Evaluation Lab. pp. 28–42. CEUR-WS (2016)
6. Pavillon, G., Coilland, P., Jougla, E.: Mise en place de la certification électronique des causes médicales de décès en France : premier bilan et perspectives [Implementation of the electronic certification of medical causes of death in France: first results and propects]. Bulletin épidémiologique hebdomadaire 35-36, 306–308 (Sep 18 2007)
7. 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. In: CLEF 2016 Online Working Notes. CEUR-WS (2016)
8. Zweigenbaum, P., Lavergne, T.: Hybrid methods for ICD-10 coding of death certificates. In: Seventh International Workshop on Health Text Mining and Information Analysis. pp. 96–105. EMNLP 2016, Austin, Texas, USA (Nov 2016)