Fusion Methods for ICD10 Code Classification of Death Certificates in Multilingual Corpora

Mike Ebersbach, Robert Herms, and Maximilian Eibl

Chair Media Informatics,
Chemnitz University of Technology, 09107 Chemnitz, Germany
{robert.herms,maximilian.eibl}@cs.tu-chemnitz.de

Abstract. In this working notes paper, we present our methodology and the results for Task 1 of the CLEF eHealth Evaluation Lab 2017. This benchmark addresses information extraction in written text with focus on unexplored languages corpora, specifically English and French. The goal is to automatically assign codes (ICD10) to text content of death certificates. Our approach is focused on fusion methods in conjunction with support vector machines for ICD10 code classification. First, we composed a large scale feature set comprising more than 40k features based on bag of words, bag of 2-grams, bag of 3-grams, latent Dirichlet allocation, and the ontologies of WordNet and UMLS. In the development phase, we evaluated three different methods: each feature type separately (no fusion), early feature-level fusion, and late fusion including the rules majority vote, maximum, and average. For the English test set, the best F-measure was 0.8187 using early fusion. For the two French test sets, we achieved 0.6692 and 0.7216 using late fusion in connection with the rule average for bag of words and bag of 2-grams.

Keywords: Natural language processing, Clinical texts, ICD10 coding, Death certificates, Machine learning, Fusion

1 Introduction

The amount of digital medical documents expands over the years, which is a major challenge regarding data processing and management in clinical institutions. However, state-of-the-art technologies can assist workflows including verbal handover supplemented with written material. For instance, the work of [1] applied automatic speech recognition to transform verbal clinical information into written free-text records. These records can then be structured by automatically identifying relevant text-snippets (e.g., [2–4]). A further aspect in hospitals and clinical institutions involves the assignment of ICD codes to reports of diseases, disorders, injuries and other related health conditions. ICD – the *International Classification of Diseases* system – is published by the World Health Organisation (WHO). Some previous work has been done for the processing of medical text corpora in conjunction with ICD codes (e.g., [5–9]). In this context, the CLEF eHealth Evaluation Lab 2017 [10] aims to ease patients and nurses in understanding and accessing eHealth information. Task 1 (Multilingual Informa-

tion Extraction - ICD10 coding) [11] of this benchmark addresses information extraction in written text with focus on unexplored languages corpora, specifically English and French. The goal of this task is to automatically assign codes (ICD10) to text content of death certificates. This challenge can be regarded as a classification task.

In this working notes paper, we present our methodology and the results for Task 1 of the CLEF eHealth Evaluation Lab 2017. Our approach is focused on the investigation of fusion methods for multilingual text classification regarding ICD10 codes. Hence, we implemented different fusion techniques to evaluate which method leads to the best result in conjunction with support vector machines. First, we composed a large scale feature set comprising more than 40k features based on the types bag of words, bag of 2-grams, bag of 3-grams, latent Dirichlet allocation [12], and the ontologies of WordNet [13] and UMLS [14]. In the development phase, we evaluated three different methods: each feature type separately (no fusion), early feature-level fusion, and late fusion including the rules majority vote, maximum, and average.

This paper is organized as follows: In the next section, we introduce the dataset. Our approach including feature extraction and fusion methods are proposed in Section 3. In Section 4, the experimental setup and the evaluation results are described. Finally, we conclude the paper in Section 5 and give some future directions.

2 Dataset

The used dataset is divided into two parts regarding the language: the CépiDc corpus (French) and the CDC corpus (English). The documents comprise free-text descriptions of causes of death as reported by physicians in standardized forms. Each document was manually labeled with one or more ICD10 codes. Two different formats are considered, the so called raw and aligned format. For English, only the raw format is included whereas the French version consists of the raw and the aligned format. Altogether, we used all three different data subsets for the evaluation. The data is partitioned into training sets (English raw with 1,073 classes and French aligned with 3,232 classes), development sets (English raw with 663 classes and French aligned with 2,363 classes), and test sets (English raw, French raw, and French aligned).

3 Methods

In this work, ICD10 code assignment to text content of death certificates is regarded as a classification task. Machine learning is performed using support vector machine (SVM). Moreover, each language is treated separately, i.e., training, development, and testing is performed on the basis of the same language. The following subsections introduce the features and the applied fusion methods.

3.1 Feature Extraction

In the preprocessing phase, all terms were stemmed and transformed to lower case and all special characters like punctuation or brackets were removed. For the French dataset, we transformed typical suffixes to their English counterpart. Furthermore, infrequent terms were removed to reduce the number of features. Subsequently, the following features were extracted:

- **Bag of n-grams:** The tf-idf (term frequency - inverse document frequency) of the terms from all documents is calculated. Feature vectors were created for bag of words (about 9k features for the French corpus and about 2k for the English corpus), bag of 2-grams, and bag of 3-grams (both with about 14k features for the French and about 3k features for the English corpus).
- **Latent Dirichlet allocation (LDA) features:** Similarities between the documents were determined by categorizing them to a preset number of topics. The confidence values of the topic assignments were used as features. For our experiments we used a number of 20 topics.
- **WordNet features:** Related terms of words in the documents were extracted to enrich the feature set with semantic information. In more detail, the first synonym and hypernym of a word (noun, verb, adjective, and adverb) ranked by WordNet was added to the feature set. The search was repeated concerning hypernyms to find more general hypernyms which were also added to the feature set. In summary, 2,784 features were extracted for the French dataset and 1,704 features for the English dataset.
- **UMLS features:** Semantic types of health vocabulary were extracted from the Unified Medical Language System (UMLS) using MetaMap [15]. There are 133 semantic types described in the UMLS. As not all types appear in the dataset, we considered a subset of 107 types (features). A feature vector was then created where each feature represents the number of search results for a particular semantic type.

3.2 Fusion

We implemented an analysis framework to investigate two fusion methods: early fusion to combine features before classification and late fusion to combine the outputs after classification. These fusion methods are illustrated in Fig. 1.

Early fusion is performed on feature-level. In this case, the feature vectors from different sources are concatenated into one large feature vector which will then be used for classification. As this vector consists of many features, training and classification time will increase. However, a large scale feature vector in conjunction with suitable learning methods can lead to much better performance in the end. Furthermore, only one learning phase is needed.

Late Fusion (or decision-level fusion) indicates combining the outputs after classification. This process predicts the final output by considering the individual labels (hard level) or scores (soft level) of the involved classifiers [16]. The following decision rules were used: majority vote (most represented class label), maximum (class label with the highest confidence), and average (class label with the highest averaged confidence).

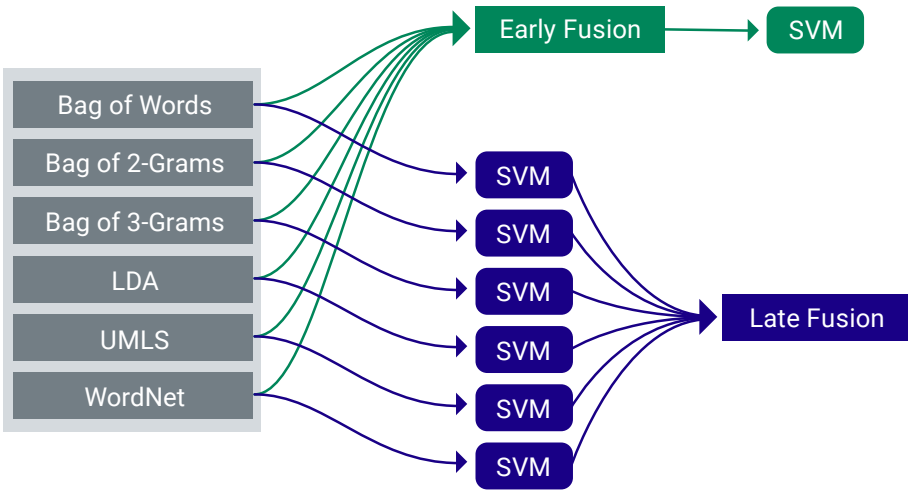

Fig. 1. *Overview of fusion methods: early feature-level fusion and late fusion.*

4 Experiments and Results

In this section, we describe the setup for the experiments. Afterwards, we report the results obtained using the six feature types and the fusion methods early feature-level fusion and late fusion.

4.1 Experimental Setup

The system performance is assessed by precision, recall, and F-measure (F1) for ICD10 code assignment. For development, we used only the F1 score as a reference for the best methods.

Classification was performed using SVM; the LIBLINEAR library [17] is used for model training. In the development phase we optimized the complexity parameter C of the SVM classifier only for early fusion. The goal is to observe the generalization performance of the classifier. We used six different values of C (1, 0.1, 0.01, 0.001, 0.0001, 0.00001). The evaluation of other methods was performed using complexity $C = 1$.

The two best performing methods (no fusion, early fusion, or late fusion) of each dataset version (language) in the development phase are applied on the corresponding test set.

Table 1. Summary of the development set (devel) results for the English and French version. Three methods are considered: 1) each feature type separately (no fusion); 2) early feature-level fusion in conjunction with complexity parameter C optimization for SVM; 3) late fusion including the rules majority vote (maj), maximum (max), and average (avg). The best F1 score of each method and language is highlighted in bold.

	English devel			French devel		
Method	Prec.	Recall	F1	Prec.	Recall	F1
BoW	0.8431	0.6562	0.7380	0.8742	0.6796	0.7647
Bo2G	0.8815	0.6826	**0.7694**	0.8882	0.6744	**0.7667**
Bo3G	0.8662	0.6372	0.7342	0.8472	0.6199	0.7159
LDA	0.3436	0.2237	0.2710	0.1648	0.1278	0.1440
UMLS	0.1360	0.1146	0.1244	0.1493	0.0660	0.0916
WordNet	0.6794	0.4983	0.5749	0.2564	0.1880	0.2170
Early Fusion (All) and SVM complexity C optimization						
$C = 1$	0.8971	0.6973	**0.7847**	0.8825	0.6596	**0.7549**
$C = 0.1$	0.8640	0.6663	0.7524	0.8572	0.6280	0.7249
$C = 0.01$	0.7952	0.5874	0.6757	0.7467	0.5529	0.6354
$C = 0.001$	0.7314	0.5035	0.5964	0.5147	0.3924	0.4453
$C = 0.0001$	0.7120	0.4638	0.5617	0.3801	0.2946	0.3320
$C = 0.00001$	0.7073	0.4551	0.5538	0.3080	0.2399	0.2697
Late Fusion						
BoW+Bo2G (avg)	0.8733	0.6782	0.7635	0.8901	0.6903	**0.7775**
BoW+Bo2G (max)	0.8701	0.6773	0.7617	0.8914	0.6874	0.7762
BoW+Bo2G+Bo3G (avg)	0.8807	0.6805	0.7678	0.8897	0.6896	0.7770
BoW+Bo2G+Bo3G (max)	0.8710	0.6780	0.7625	0.8931	0.6838	0.7745
BoW+Bo2G+Bo3G (maj)	0.8805	0.6817	**0.7684**	0.8892	0.6819	0.7719
All (avg)	0.8505	0.6441	0.7330	0.8425	0.6564	0.7379
All (max)	0.7936	0.5944	0.6797	0.8757	0.6695	0.7588
All (maj)	0.8522	0.6536	0.7398	0.8538	0.6501	0.7381

4.2 Results

A series of experiments was carried out for the automatic classification of ICD10 codes in medical text corpora. Table 1 summarizes the development set results for the English and French dataset. Although our criterion for the selection of the best two methods of each language is the F1 score, the results of precision and recall are shown for comparison purposes.

In the feature type experiments without fusion, the best F1 results were obtained by bag of 2-grams (Bo2G) for both languages; 0.7694 for English and 0.7667 for French. In contrast, the highest recall measure for French (0.6796) was achieved with bag of words (BoW).

For early fusion, the best F1, precision, and recall measures were obtained using SVM complexity $C = 1$ concerning both languages (F1 is 0.7847 for English and 0.7549 for French). With $C < 1$, the values are too small which results in over-generalization, i.e., underfitting of the SVM model.

Table 2. Summary of the results for the English and French test set. Additionally, the average scores of all participants were computed excluding non-official runs (Task average). Two evaluation types are considered: main evaluation reference (all ICD codes) and secondary reference (only external causes). [1]: results for the raw format. [2]: results for the aligned format. [*]: non-official run.

	all ICD codes			only external causes		
Method	Prec.	Recall	F1	Prec.	Recall	F1
English test[1]						
Early Fusion ($C = 1$)	0.9402	0.7251	0.8187	0.8800	0.1746	0.2914
Bo2G	0.9291	0.7169	0.8093	1.000	0.1587	0.2740
Task average	0.6700	0.5820	0.6220	0.4050	0.2670	0.2610
French test[1]						
BoW+Bo2G (avg)[*]	0.8827	0.5388	0.6692	0.7803	0.2903	0.4232
Bo2G[*]	0.8818	0.5357	0.6665	0.7667	0.2834	0.4139
Task average	0.4747	0.3583	0.4059	0.3668	0.2474	0.2921
French test[2]						
BoW+Bo2G (avg)[*]	0.8750	0.6140	0.7216	0.7479	0.3234	0.4515
Bo2G	0.8744	0.6106	0.7191	0.7400	0.3182	0.4450
Task average	0.6479	0.5555	0.5933	0.5051	0.3109	0.3663

The late fusion scheme has been applied to all feature types. Additionally, the top three feature types were selected to investigate the results without features that have a low classification performance (threshold is F1 = 0.7). As a consequence, the top three features types are bag of words (BoW), bag of 2-grams (Bo2G), and bag of 3-grams (Bo3G). For the English language, the best F1 score is 0.7684 using BoW+Bo2G+Bo3G in connection with majority vote. However, the best precision with 0.8807 was achieved using BoW+Bo2G+Bo3G and the rule average. In case of French, BoW+Bo2G and the rule average was superior with a F1 score of 0.7775 whereas the best precision with 0.8931 was obtained using BoW+Bo2G+Bo3G and the rule maximum.

The two best performing methods of each language in the development phase were then applied on the corresponding test sets. The results are shown in Table 2. The main evaluation reference for the task refers to all ICD10 codes. Additionally, external causes, characterized by the codes V01 to Y98, are considered as a secondary reference. In this case, the evaluation addresses a specific type of deaths such as violent deaths which are avoidable.

Regarding the English test set, the best method was early fusion which achieved a F1 score of 0.8187 (all ICD codes) and 0.2914 (external causes). For the French test set, the highest F1 score was obtained using late fusion of BoW+Bo2G in connection with the rule average (raw format: 0.6692 for all ICD codes and 0.4232 for external cases; aligned format: 0.7216 for all ICD codes and 0.4515 for external cases). However, the best results for the French test set are non-official, because they were submitted after the task deadline. Consequently, as shown in Table 2, the only official result for the French test set is obtained

using the feature type Bo2G with a F1 score of 0.7191 (all ICD codes) and 0.4450 (external causes).

5 Conclusions

We presented our methodology for Task 1 of the CLEF eHealth Evaluation Lab 2017 where the goal is to automatically assign codes (ICD10) to text content of death certificates. The corpus is made of two versions regarding the language: English and French.

Our approach is focused on fusion methods in conjunction with support vector machines for ICD10 code classification. We composed a set of features based on bag of words, bag of 2-grams, bag of 3-grams, latent Dirichlet allocation, and the ontologies of WordNet and UMLS. Three different methods were evaluated: each feature type separately (no fusion), early feature-level fusion, and late fusion. For the English test set, the best F-measure was 0.8187 using early fusion. For the two French test sets, we achieved 0.6692 and 0.7216 using late fusion in connection with the rule average for bag of words and bag of 2-grams.

However, further improvements could be achieved by more knowledge bases and other appropriate features from the field of Natural Language Processing. Moreover, the holistic system could benefit from other machine learning methods such as artificial neural networks, Naive Bayes, or k-nearest neighbors. Finally, fusion schemes can be optimized by input weights and the consideration of correlations between the inputs.

References

1. Herms, R., Richter, D., Eibl, M., Ritter, M.: Unsupervised language model adaptation using utterance-based web search for clinical speech recognition. *CLEF 2015 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2015.
2. Song, Y., He, Y., Liu, H., Wang, Y., Hu, Q., He, L., Luo, G.: ECNU at 2016 eHealth Task 1: Handover Information Extraction. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2016.
3. Quiroz, L., Mennes, L., Dehghani, M., Kanoulas, E.: Distributional Semantics for Medical Information Extraction. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2016.
4. Ebersbach, M., Herms, R., Lohr, C., Eibl, M.: Wrappers for Feature Subset Selection in CRF-based Clinical Information Extraction. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2016.
5. Dermouche, M. Looten, V., Flicoteaux, R., Chevret, S., Velcin, J., Taright, N.: ECSTRA-INSERM@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2016.
6. Mottin, L., Gobeill, J., Mottaz, A., Pasche, E., Gaudinat, A., Ruch, P.: BiTeM at CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2016.

7. Zweigenbaum, P., Lavergne, T.: LIMSI ICD10 coding experiments on CépiDC death certificate statements. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2016.
8. Cabot, C., Soualmia, L., Dahamna, B., Darmoni, S.: SIBM at CLEF eHealth Evaluation Lab 2016: Extracting Concepts in French Medical Texts with ECMT and CIMIND. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2016.
9. Lohr, C., Herms, R.: A Corpus of German Clinical Reports for ICD and OPS-based Language Modeling. *Proceedings of the Controlled Language Applications Workshop (CLAW) at LREC 2016*, pages 20–23, 2016.
10. Suominen, H., Kelly, L., Goeuriot, L., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Zuccon, G., Palotti, J.: Overview of the CLEF eHealth Evaluation Lab 2017. *CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS)*, Springer, September, 2017.
11. Névéol, A., Anderson, R. N., Cohen, K. B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Zweigenbaum, P.: CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French. *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, September, 2017.
12. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
13. Miller, G.: WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, ACM, 1995.
14. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, Oxford Univ Press, 2004.
15. Aronson, A.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, American Medical Informatics Association, 2001.
16. Castellano, G., Gunes, H., Peters, C., Schuller, B.: Multimodal affect recognition for naturalistic human-computer and human-robot interactions. *The Oxford handbook of affective computing*, Oxford University Press, USA, 2014.
17. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.