LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: Track 1.B De-Identification of Clinical Texts at Character and Token Levels

Cyril Grouin

LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: Track 1.B De-Identification of Clinical Texts at Character and Token Levels

Cyril Grouin, PhD
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France

Introduction

Track 1.B of the CEGS N-GRID 2016 NLP Shared-Tasks consisted in automatically de-identifying clinical texts. The training dataset is composed of 600 texts annotated into 7 main categories of PHI *(AGE, CONTACT, DATE, ID, LOCATION, NAME, PROFESSION)*. The test dataset is composed of 400 unannotated texts. A tokenization issue occurred in both corpora: a high proportion of end-of-lines were removed, producing run-in words, i.e., sequences of two tokens where it may happen that only one token must be de-identified (e.g., "11/20/2073CPT").

Methods

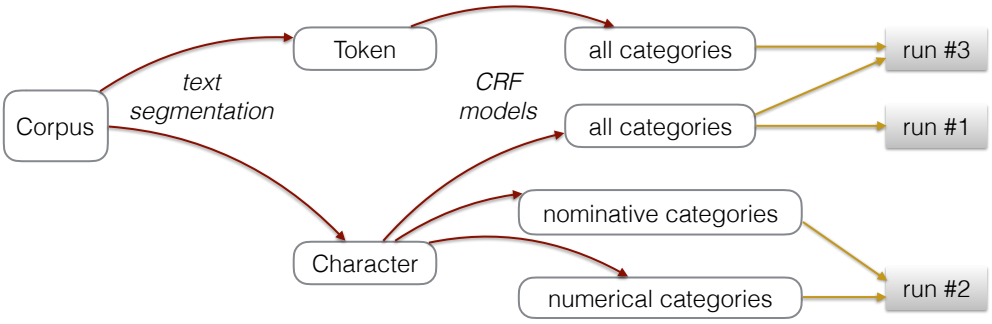

Figure 1: Type of segmentation (token/character-based), type of models (generic/specific), and official submissions

Our approach relies on Wapiti [1], a CRF-based system. Since a tokenization issue occurred in all texts, we decided to process the clinical texts using two types of text segmentation (see Figure 1). We produced both character-based models (predictions for each character) and token-based models (predictions for each token, using a tokenization based on white spaces). For character-based models, we produced both a generic model to predict all categories of PHI and specific models to predict nominative categories* only or numerical categories† only. We used the following features:

- Character-based CRF models: character, unigrams of the six previous and the six next characters, token the character belongs to, unigrams of previous and next tokens, character is a punctuation, character is a digit;
- Token-based CRF models: token, unigrams of two previous and next surrounding tokens, presence of punctuation mark and digit in the token, typographic case of the token, and cluster ID from an automatic unsupervised clustering of all tokens of the training corpus into 360 clusters using Brown's algorithm [2].

We defined three configurations for this participation (see Figure 1):

- Run #1: one generic character-based CRF model. The aim is to process all categories using a single model.
- Run #2: combination of two specific character-based CRF models, a nominative model and a numerical model. The aim is to process PHIs of the nominative and numerical categories more accurately using specific models.
- Run #3: combination of two generic CRF models, a character-based model and a token-based model. The aim is to maximize the benefit of each type of text segmentation to deal with the initial tokenization issue. In

*Nominative categories concern the following sub-categories: *City, Country, Doctor, Hospital, Organization, Patient, Profession, State, Street.*
†Numerical categories concern the following sub-categories: *Age, Date, Fax, License, Medical Record, Phone, Zip.*

the combination process, in case of overlap between predictions made by the two models (e.g., the sequence "March 2090Spiritual/Religion" by the token-based model vs. the sequence "March 2090" by the character-based model), we gave priority to predictions produced by the character-based model due to its better accuracy.

Results

Table 1 presents the results we achieved on the test dataset for each one of our three configurations.

Category	Run #1 (character)			Run #2 (numer+nominat)			Run #3 (character+token)		
	R	P	F	R	P	F	R	P	F
All	0.7090	**0.9106**	0.7973	0.7211	0.8915	0.7973	**0.7487**	0.9048	**0.8193**
CONTACT	0.8560	**0.8652**	**0.8606**	0.7893	0.8630	0.8245	**0.8587**	0.8610	0.8598
NAME	0.5620	**0.9240**	0.6989	0.5965	0.9167	0.7227	**0.6392**	0.9121	**0.7516**
DATE	0.9510	**0.9797**	0.9651	0.9488	0.9402	0.9445	**0.9585**	0.9765	**0.9674**
AGE	0.7952	**0.8593**	0.8260	0.8139	0.8394	0.8264	**0.8366**	0.8462	**0.8414**
PROFESSION	0.5182	**0.8966**	0.6568	**0.5912**	0.8120	0.6842	0.5643	0.8914	**0.6911**
ID	0.3958	0.7037	0.5067	0.3542	**0.9444**	0.5152	**0.4167**	0.7143	**0.5263**
OTHER	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LOCATION	0.4967	0.8061	0.6146	0.4983	**0.8374**	0.6248	**0.5551**	0.8133	**0.6598**

Table 1: Official results on the test dataset (token level: R=Recall, P=Precision, F=F-measure). Best results are in bold

Discussion

We achieved our best results when combining character and token-based CRF models (global F-measure of 0.8193, run #3). This configuration allows us to obtain the highest values of F-measure and recall for almost all PHI categories. The first configuration (character-based model only) allows us to obtain the highest values of precision for almost all categories. Nevertheless, the global F-measure (0.7973) as well as the F-measure values of all categories are lower (except for CONTACT) than those obtained on the third configuration. Finally, the combination of nominative and numerical character-based models did not improve the results, except for PROFESSION where the recall is the highest.

Nevertheless, for long sequences, token-based models outperform character-based models. As an example, the sequence "construction worker" is correctly identified using the token-based model while the character-based model only identifies the first element of the sequence ("construction"), the sequence being too long for this kind of model. In this case, the priority given to predictions from the character-based model during the combination process would produce an incorrect output. We estimate a voting procedure may improve the combination process.

Conclusion

To cope with tokenization issues in the clinical texts of the corpus, we produced both character-based CRF models (to identify frontiers of sequences with more accuracy) and token-based CRF models (to process long sequences) to automatically de-identify clinical texts. The combination of predictions made by character and token-based models, with priority given to predictions from the character-based model in case of overlapping sequences, allowed us to achieve our best results (global recall of 0.7487 and F-measure of 0.8193). There is room for improvement, especially for the NAME category for which we achieved a recall of 0.6392 using our best configuration. New features must be identified to process the false positives (e.g., last names written in lower case: "Dr. denton" and composed names).

References

1. Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *Proc of ACL*, pages 504–13, Uppsala, Sweden, July 2010.
2. Peter F Brown, Vincent J Della Pietra, Peter V de Souza, Jenifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–79, 1992.