LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: Track 1.A De-Identification of Unseen Clinical Texts

Cyril Grouin

▶ **To cite this version:**

Cyril Grouin. LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: Track 1.A De-Identification of Unseen Clinical Texts. Workshop Challenges in Natural Language Processing for Clinical Data, Nov 2016, Chicago, United States. hal-01831223

HAL Id: hal-01831223
https://hal.science/hal-01831223

Submitted on 18 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: Track 1.A De-Identification of Unseen Clinical Texts

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

Introduction

Track 1.A of the CEGS N-GRID 2016 NLP Shared-Tasks consisted in automatically de-identifying unseen clinical texts. The organizers gave access to 600 unannotated clinical texts on June 6th. Participants were expected to submit their outputs no later than June 10th. The categories of de-identification were based on those defined during the i2b2/UTHealth 2014 NLP Challenge on de-identification of clinical texts [1]. Nevertheless, clinical texts were different in the two shared-tasks in terms of content and text pre-processing (tokenization and end-of-line formatting).

Methods

Since we participated in the i2b2/UTHealth 2014 NLP Challenge, and because categories of de-identification were similar, we simply reused the system we designed [2] and applied on the i2b2/UTHealth 2014 NLP Challenge data. Our approach relies on Wapiti [3], a CRF-based system, and post-processing rules.

CRF system We used the following features: token, length of the token, typographic case of the token, presence of punctuation mark and digit in the token, part-of-speech of the token provided by the Tree Tagger [4], identification of the token in a list of trigger words *(Dr, M.D., Mrs, Name, yo, y/o, avenue, st, street)*, and cluster ID from an automatic unsupervised clustering of all tokens of the training corpus into 320 clusters using Brown's algorithm [5]. We also defined bigrams of features for token, typographic case, and POS. An automatic feature selection was performed ($l1$ regularization).

The CRF model was trained on 2014 data to predict 21 sub-categories of de-identification. As in 2014, we merged some of the sub-categories (in italics) into generic categories (in upper case) to produce the final outputs: CONTACT *(email, fax, IP address, phone, URL)*; NAME *(doctor, patient, username)*; DATE *(date)*; AGE *(age)*; PROFESSION *(profession)*, ID *(ID num, medical record)*; LOCATION *(city, country, hospital, organization, other, state, street, zip)*.

Post-processing We reused the 77 post-processing rules we defined in 2014 to correct the CRF output. These post-processing rules deal with the following specific cases: identification of multi-word expressions in lexicon (city, country, profession and organization name), identification of multi-token sequences (phone number, date, patient and doctor name introduced or followed by a trigger word, etc.), deletion of elements out of the scope of the PHI (only numerical values for ages, no title for medical doctor names).

Additionally, we produced lists of PHIs from the 2014 i2b2/UTHealth training dataset. In the full post-processing stage, tokens were also searched within these lists of known PHIs while in the limited post-processing stage, we did not use them.

Design of experiments For our participation to the CEGS N-GRID 2016 Shared-Tasks, we reused the three configurations we defined while participating in the 2014 i2b2/UTHealth NLP Challenge:

- Run #1: CRF model from 2014 and full post-processing rules
- Run #2: CRF model from 2014 only (no post-processing rules)
- Run #3: CRF model from 2014 and limited post-processing rules

Results

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

Category	Run #1 (full post-proc)			Run #2 (no post-proc)			Run #3 (limited post-proc)		
	R	P	F	R	P	F	R	P	F
All	**0.4901**	0.5439	0.5156	0.4192	0.6657	0.5145	0.4899	**0.7234**	**0.5842**
CONTACT	**0.4304**	**0.2750**	**0.3356**	0.1130	0.5909	0.1898	**0.4304**	**0.2750**	**0.3356**
NAME	**0.6042**	0.4035	0.4838	0.5989	0.5454	0.5709	**0.6042**	**0.6931**	**0.6456**
DATE	**0.5951**	0.7389	0.6592	0.5226	0.6860	0.5933	**0.5951**	**0.8189**	**0.6893**
AGE	**0.4362**	**0.8924**	**0.5860**	0.3855	0.9493	0.5483	**0.4362**	**0.8924**	**0.5860**
PROFESSION	**0.3434**	0.4675	0.3960	0.1007	**0.9540**	0.1822	**0.3414**	0.4848	**0.4007**
ID	0.0345	1.0000	0.0667	0.0345	1.0000	0.0667	0.0345	1.0000	0.0667
OTHER	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LOCATION	**0.3712**	0.4295	0.3982	0.3153	0.6685	0.4285	**0.3712**	**0.6874**	**0.4821**

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

Discussion

We achieved our best results with the third run (F=0.5842). The second run achieved the worst results (no post processing rules) and the first run obtained slightly lower results than the third one due to the fact that the lists used in the post-processing rules were based on annotations from 2014. In 2014, we obtained our best results using the first run (R=0.7332, P=0.8937, F=0.8055), closely followed by results from the third run.

Since clinical texts and formatting characteristics were different between 2014 and 2016, the model we applied did not suit the clinical texts of 2016. The differences between the two corpora were of two types. First, one can identify three types of formats in 2014 clinical texts while 2016 texts only refer to the last type: fixed-width columns, double spacing between each line of text, single spacing between each line of text. Second, a tokenization issue occurred in 2016 texts: 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"). Since our approach needs a perfect tokenization, such an issue bore a strong impact on the success of our system.

Conclusion

Simply reusing a CRF model trained on one type of clinical text on a new type of clinical text without any adaptation produced lower results than on the original texts (loss of 22 F-measure points between 2014 and 2016). Nevertheless, this experiment illustrates that the type of system we designed, which belongs to the general class of supervised machine learning systems, must be tuned to the target data in order to obtain useful results.

References

1. Amber Stubbs, Christopher Kotfila, and Ozlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *J Biomed Inform*, 58:S11–S19, 2015.
2. Cyril Grouin. Clinical records de-identification using CRF and rule-based approaches. In *i2b2 Work Proc*, 2014.
3. Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *Proc of ACL*, pages 504–13, Uppsala, Sweden, July 2010.
4. Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *New Meth in Language Proc*, 1994.
5. 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.