Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation

Vincent Claveau[1], Lucas Emanuel Silva Oliveira[2], Guillaume Bouzillé[3], Marc Cuggia[3], Claudia Maria Cabral Moro[2], Natalia Grabar[4]

[1] IRISA - CNRS, Rennes, France
[2] PUCPR - Pontifícia Universidade Católica do Paraná, Curitiba, Brazil
[3] INSERM/LTSI, HBD; CHU de Rennes; Université Rennes 2
[4]CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
vincent.claveau@irisa.fr, lucas.oliveira@pucpr.br,
guillaume.bouzille@chu-rennes.fr, marc.cuggia@chu-rennes.fr,
c.moro@pucpr.br, natalia.grabar@univ-lille3.fr

Abstract. Clinical trials are fundamental for evaluating therapies and diagnosis techniques. Yet, recruitment of patients remains a real challenge. Eligibility criteria are related to terms but also to patient laboratory results usually expressed with numerical values. Both types of information are important for patient selection. We propose to address the processing of numerical values. A set of sentences extracted from clinical trials are manually annotated by four annotators. Four categories are distinguished: C (concept), V (numerical value), U (unit), O (out position). According to the pairs of annotators, the inter-annotator agreement on the whole annotation sequence CVU goes up to 0.78 and 0.83. Then, an automatic method using CFRs is exploited for creating a supervised model for the recognition of these categories. The obtained F-measure is 0.60 for C, 0.82 for V, and 0.76 for U.

Keywords: Natural Language Processing, Supervised Learning, Clinical Trials, Patient Eligibility, Numerical Criteria

1 Introduction

In the clinical research process, recruitment of patients for clinical trials (CTs) remain an unprecedented challenge, while they are fundamental for evaluating therapies or diagnosis techniques. They are the most common research design to test the safety and efficiency of interventions on humans. CTs are based on statistical inference and require appropriate sample sizes from well identified population. The challenge is to enroll a sufficient number of participants with suitable characteristics to ensure that the results demonstrate the desired effect with a limited error rate. Hence, CTs must define a precise set of inclusion and exclusion criteria (eg, age, gender, medical history, treatment, biomarkers). With paper files and EHRs as the main sources of information, only human operators

are capable to efficiently detect the eligible patients [3]. This is a laborious and costly task, and it is common that CTs fail because of the difficulty to meet the necessary recruitment target in an acceptable time [6]: almost half of all trial delays are caused by participant recruitment problems. Only 18% in Europe, 17% in Asia-Pacific, 15% in Latin America, and 7% in the USA complte enrollment on time [4]. The existing enrollment systems are facing the gap between the free text representation of clinical information and eligibility criteria [11, 17]. Most of them propose to fill in this gap manually, while automatic NLP methods may help to overcome this issue.

The traditional NLP work is dedicated to the recognition and extraction of terms. Yet, there is an emerging work on detection of temporality, certainty, and numerical values. Such information has the purpose to complete, enrich and more generally make more precise the terminological information. In the general language, framework for automated extraction and approximation of numerical values, such as height and weight, has been proposed [5]. It uses relation patterns and WordNet and shows the average precision up to 0.84 with exact matching and 0.72 with inexact matching. Another work proposes two extraction systems: rule based extractor and probabilistic graphical model [9] for extraction of life expectancy, inflation, electricity production, etc. It reaches 0.56 and 0.64 average F-measure for the rule-based and probabilistic systems, respectively. On the basis of a small set of clinical annotated data in French, a CRF model is trained for the recognition of concepts, values, and units. Then, a rule-based system is designed for computing the semantic relations between these entities [2]. The results obtained show average F-measure 0.86 (0.79 for concepts, 0.90 for values and 0.76 for units). On English data, extraction of numerical attributes and values from clinical texts is proposed [13]: after the extraction of numerical attributes and values with CRFs, relations for associating these attributes to values are computed with SVMs. The system shows 0.95 accuracy with entities and 0.87 with relations. Yet another work is done on cardiology radiological reports in English [10] and achieves 93% F1-measure. In contrast with these studies, here we focus on clinical trial protocols written in English.

2 Material

Clinical Trials. In December 2016, we downloaded protocols of the whole set of CTs from *www.clinicaltrials.com*. The corpus counts 211,438 CTs. We focus on inclusion and exclusion criteria (more than 2M sentences).

Reference Annotations 1,500 randomly selected sentences are annotated by 3 annotators with different backgrounds (medical doctor and computer scientists). Each sentence is annotated by at least two of them. On such typical sentences:

- *Absolute neutrophil count ≥ 1,000 cells/µl.*
- *Exclude if T3 uptake is less than 19%; T4 less than 2.9 ((g/dL); free T4 index is less than 0.8.*

the annotators have to mark up three categories of entities: C (concepts *Absolute neutrophil count, T3 uptake, T4, free T4 index*), V (numerical values *≥ 1,000, less than 19, less than 2.9, less than 0.8*), U (units *cells/µl, %, g/dL*).

3 Methods

The main objective of the methods is to create an automatic model for the detection of numerical values (concept, value and unit).

Inter-annotator agreement. In order to assess the inter-annotator agreement, we compute Cohen's κ [1] between each pair of annotators. The final version is obtained after a consensus is reached among the annotators.

Automatic annotation. Conditional Random Fields (CRFs) [7] are undirected graphical models that represent the probability distribution of annotation y on observations x. They are widely used in NLP thanks to their ability to take into account the sequential aspect and rich descriptions of text sequences. CRFs have been successfully used in many tasks casted as annotation problems: information extraction, named entity recognition, tagging, etc. [18, 12, 14]. From training data, CRF models learn to assign a label to each word of a sentence such that the tag sequence is the most probable given the sentence given as input. We want the CRFs to learn to label words denoting a concept with the tag C, values with V, units with U, while every other words will receive a void label noted O. In order to handle multi-word concepts, values and units, we adopt the so-called BIO scheme: the first word of a multi-word concept is labeled as B_C (B stands for *beginning*), next words are labeled as I_C (I stands for *inside*), the same for values and units. To find the most probable label of a word at position i in a sentence, CRFs exploit features describing the word (for example, Part-of-Speech tags, lemmas [15], graphemic clues) and its context (words and features at positions $i-1$, $i+1$, $i-2$) up to 4 words. The CRF implementation used for our experiments is Wapiti [8], which is known for its efficiency.

Evaluation of automatic annotation. The evaluation is performed against the reference data and is measured with token errors (percentage of words wrongly labeled with respect to the human annotation) and F-measure [16].

4 Results and Discussion

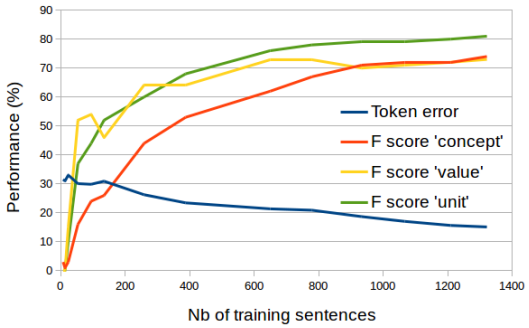
Inter-annotator agreement. In Table 1, we indicate the inter-annotator agreement for each pair of annotators and taking into account two tagsets: whole set of tags and the tagset without concepts. The figures indicate that A1 and A2 show the highest agreement: both have important experience in medical area. When concepts are not taken into account the agreement is even better: manual annotation of concepts is more complicated than annotation of the two other categories. With annotations from A1 and A2, the consensual annotation is built. This version of data is used for training and evaluation of the supervised model.

Automatic annotation. In Figure 1, we present the evaluation of automatic annotation in terms of token errors and F-score for each category. In order to estimate the ideal amount of the required training data, we also display the evolution of the performance according to the size of the training data used. First, the global error rate tends to decrease. Since its decrease continues, more training data would help reaching better results. Among the categories aimed,

Table 1. Inter-annotator agreement (Cohen's κ) on the whole and reduced tag sets

	A1 vs. A2	A1 vs. A3	A2 vs. A3
κ whole tagset	0.78	0.51	0.47
κ without 'concept'	0.83	0.60	0.64

Fig. 1. CRF annotation performance (globally in terms of token errors, and by category in terms of F-score) according to the amount of training data (number of sentences)



the best performance is obtained with units, while the concept category is the most difficult to detect. For these two categories, the performance continues to grow up: a larger set of annotated data would be helpful. As for the value category, its evolution is less linear and finally it seems to find a "plateau" with no more apparent evolution. Otherwise, the detection efficiency of this category is in between the two other categories. The obtained F-measure is 0.60 for C, 0.82 for V, and 0.76 for U.

5 Conclusion and Future Work

Recruitment of patients for CTs is a difficult task. We proposed a contribution to this task. We generate an automatic model for the detection of numerical values, composed of thee items (concept C, value V and unit U), in narrative text in English. These results are evaluated against reference data and show F-measure 0.60 for C, 0.82 for V, and 0.76 for U. We have several directions for future work: to normalize the units; to build resources and rules for their standardization (*cell/mm3* instead of *cell/cm3*); to prepare a larger set of reference annotations; to complete these annotations with temporal information; to apply the models for enrollment of patients in French and Brazilian hospitals.

Acknowledgements. This work was partly funded by CNRS-CONFAP project FIGTEM for Franco-Brazilian collaborations and a French government support granted to the CominLabs LabEx managed by the ANR in Investing for the Future program under reference ANR-10-LABX-07-01.

References

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics 34(4) (2008)
2. Bigeard, E., Jouhet, V., Mougin, F., Thiessard, F., Grabar, N.: Automatic extraction of numerical values from unstructured data in EHRs. In: MIE (Medical Informatics in Europe) 2015. Madrid, Spain (2015)
3. Campillo-Gimenez, B., Buscail, C., Zekri, O., Laguerre, B., Le Prisé, E., De Crevoisier, R., Cuggia, M.: Improving the pre-screening of eligible patients in order to increase enrollment in cancer clinical trials. Trials 16(1), 1–15 (2015)
4. Center Watch: State of the clinical trials industry: A sourcebook of charts and statistics. Tech. rep., Center Watch (2013)
5. Davidov, D., Rappaport, A.: Extraction and approximation of numerical attributes from the web. In: 48th Annual Meeting of the Association for Computational Linguistics. pp. 1308–1317 (2010)
6. Fletcher, B., Gheorghe, A., Moore, D., Wilson, S., Damery, S.: Improving the recruitment activity of clinicians in randomised controlled trials: A systematic review. BMJ Open 2(1), 1–14 (2012)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (ICML) (2001)
8. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 504–513. Association for Computational Linguistics (July 2010), `http://www.aclweb.org/anthology/P10-1052`
9. Madaan, A., Mitta, A., Mausam, Ramakrishnan, G., Sarawagi, S.: Numerical relation extraction with minimal supervision. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
10. Nath, C., Albaghdadi, M., Jonnalagadda, S.: A natural language processing tool for large-scale data extraction from echocardiography reports. PLoS One 11(4), 153749–64 (2016)
11. Olasov, B., Sim, I.: Ruleed, a web-based semantic network interface for constructing and revising computable eligibility rules. In: AMIA Symposium. p. 1051 (2006)
12. Pranjal, A., Delip, R., Balaraman, R.: Part Of speech Tagging and Chunking with HMM and CRF. In: Proceedings of NLP Association of India (NLPAI) Machine Learning Contest (2006)
13. R., S.P., Mandhan, S., Niwa, Y.: Numerical atrribute extraction from clinical texts. CoRR 1602.00269 (2016), `http://arxiv.org/abs/1602.00269`
14. Raymond, C., Fayolle, J.: Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In: Actes de la conférence Traitement Automatique des Langues Naturelles. Montréal, Canada (2010)
15. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proc of International Conference on New Methods in Language Processing. pp. 44–49 (1994)
16. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
17. Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P., Elhadad, N., Johnson, S., Lai, A.: A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 21(2), 221–30 (2014)
18. Wang, T., Li, J., Diao, Q., Wei Hu, Y.Z., Dulong, C.: Semantic event detection using conditional random fields. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06) (2006)