BigClin
pre-Final report

Sept 6th, 2019

I. Context and objectives

A well-known challenge for secondary use of Clinical Big Data is that much of detailed patient information is embedded in narrative text, mostly stored as unstructured data. The lack of efficient Natural Language Processing (NLP) resources dedicated to clinical narratives, especially for French, leads to the development of ad-hoc NLP tools with limited targeted purposes. Moreover, the scalability and real-time issues are rarely taken into account for these possibly costly NLP tools, which make them inappropriate in real-world scenarios.

The BigClin project proposes to address the essential need to leverage the above barriers when reusing unstructured clinical data at a large scale:
1) We propose to develop new clinical records representation relying on fine-grained semantic annotation thanks to new NLP tools dedicated to French clinical narratives.
2) Since, the aim is to efficiently map this added semantic information to existing structured data to be further analyzed in a Big data infrastructure, the project also addresses distributed systems issues: scalability, management of uncertain data and privacy, stream processing at runtime...
3) To evaluate the added value of these methods in several real clinical data and on real use-cases, including epidemiology and pharmaco-vigilance, clinical practice assessment and health care quality research, clinical trials.

II. Main scientific results

Development of annotated datasets

Clinical records cannot be shared, which is a real hurdle to develop and compare information extraction techniques. One important outcome of the project is the development of annotated corpora, that share the same linguistic properties than records, but can be freely distributed for research purposes. Several corpora and several types of annotation were proposed for French, Portuguese and English. See the Dataset section for more details.

These corpora will foster reproducible research on clinical text mining. They are already at the heart of the DeFT text-mining competition (https://deft.limsi.fr/2019/) that we organized (with other colleagues) in 2019.

Extracting information from clinical data

Several NLP techniques and tools have been developed within the project in order to identify relevant medical or linguistic information. They are all chiefly based on machine learning approaches, and for most of them, more specifically, on deep learning.
For instance, we have developed a new Part-of-Speech tagger and lemmatizer for French, especially suited to handle medical texts; it is freely available as a web-service at https://allgo.inria.fr .
The identification of negation and uncertainty is important to precisely understand the clinical texts. Thus, we have proposed neural techniques to find the negation/uncertainty cues and their scope (part of sentence concerned by the negation or uncertainty). It achieves state-of-the-art results on English, and is pioneer work for French and Portuguese for which it sets a new standard; it is available at https://allgo.inria.fr .
Other achievements in text-mining includes: numerical value extraction (finding concepts that are measured, such as lab results, numerical expressions, their units) in French, English and Portuguese, the identification of gender, age, outcome and admission reasons in French clinical texts...

Identification of diseases, findings, and disorders from clinical text is crucial to meet eligibility criteria. This information is also relevant for drug safety and medico economic studies . Therefore, neural approaches (LSTM and CNN algorithms) for ICD-10 categorization were developed using real and massive corpus of patient data coming from the EHR of the university hospital of Rennes. For this task, we did not use a direct annotation approach which is quite time consuming, overcosted and difficult to get without the contribution of proficient annotators (clinicians or healthcare professionals). We leveraged the retrospective DRG data (i.e PMSI in France) made available in the Clinical Data Center of Rennes hospital, as well as the corresponding discharge summaries from which the ICD-10 codes were previously manually extracted. Thus, the dataset used for this task included the comprehensive clinical data for all inpatients of 2015, which encompassed xxx patients, xxx stays and xxx documents. The preliminary results of the automatic coding are encouraging but should require more interdisciplinary interactions to fill the gap between these theoretical results and the development of operational tools expected by medical users. However, the contribution of massive dataset with indirect annotation could be easily replicated in other hospitals since all of them produce this kind of data.

Distributed real-time data stream processing

The need to analyze in real time large-scale and distributed data streams has recently became tremendous. In particular the identification of recent heavy-hitters (or hot items) is essential but highly challenging. This problem has been heavily studied during the last decades with both exact and probabilistic solutions. A lot of research works have been dedicated to the detection of top-k items in continuous and massive streams since the seminal work of Misra and Gries in 1982. Lines of work particularly focus on the space and time complexity of the detection algorithms in order to cope with the continuously increasing rates of data generation. While simple to state and fundamental for advanced analysis, the detection of top-k items in continuous and massive streams over a time sliding window and among distributed nodes is still an active research field. The distributed detection of frequent items over a sliding window presents two extra challenging aspects with respect to the centralized detection of frequent items since the inception of the stream: (i) treat time-decaying items as they enter and exit the sliding window, and (ii) produce mergeable local stream summaries in order to obtain a system-wide summary.
Our algorithm is based on a deterministic counting of the most over-represented items in the data streams, which are themselves probabilistically identified using a dynamically defined threshold. To fix a proper threshold, we combine two known theoretical results issued from the coupon collector and the leader election problems. To analyze the performance of our approach, we have considered simulated data following zipfian distributions, essentially because the family of power law distributions are widely observed in real world data sets and natural phenomena. Obtained results on simulated data show how impressive is the detection of the top-k items in a large range of distributions, which in addition to the fact that our algorithm is resilient to the permutation of items within the data stream, and fully adapted to a distributed sliding window schema, makes our approach innovative.

III. Other achievements

The BigClin project has served as a springboard to launch a collaboration with the Dept of health informatics in PUCPR, Curitiba, Brazil. With the help of additional funding for travel expenses and student exchanges from the CNRS (for France) and conFAP (for Brazil), we have built strong scientific links with C. Moro's team around the BigClin scientific objectives.

As already mentioned, our work within BigClin motivated us to propose a Text mining challenge for the clinical domain, with the help of T. Hamon and C. Grouin (LIMSI). This challenge, DeFT2019, relies on our CAS corpus on which 3 tasks were proposed: generation of keyword, search for expertise given a case and information extraction about the patient and its diseases. It was a success with 9 teams finally participating. A workshop will be held at PFIA conference in July 2019. Furthermore, we have been invited as editors to supervise a special issue of the journal RIDoWS to publicise this challenge and other biomedical text-mining works.

IV. Training

Two PhD were funded by the project:

- PhD of Clément Dalloux, Indexing and information extraction in clinical texts, started in Dec 2016. Defense expected in January 2020.
- PhD of Vasile Cazacu, Calcul distribué pour la fouille de données cliniques, started in February 2017. Defense expected in February 2020.

One 12-month post-doctoral position is to be hired in 2020 (delayed due to difficulty to hire a candidate with appropriate skills).

Student exchange: C. Dalloux (IRISA) spent one month in the Dept. of medical informatics of PUCPR, Curitiba, Brazil in 2017 to adapt his neural techniques to Brazilian clinical data.
Student exchange: L. Oliveira and Y. Gumiel (PUCPR) spent 3 months in the LinkMedia team, IRISA, in 2018 to work on information extraction for cohort selection in clinical trials.

Others

- C. Moro, L. Oliveira (PUCPR) and N. Grabar (STL) gave invited talks about their work in BigClin to students of the Master bio-informatics, Univ. Rennes 1.
- V. Claveau and N. Grabar gave a 24h course about biomedical NLP at the Master in medical informatics, Univ. Rennes 1.

V. Publications and software

Journals, proceedings, conferences and workshops

- Vincent Claveau and Ewa Kijak: **Direct vs. indirect evaluation of distributional thesauri**. *International Conference on Computational Linguistics, COLING* : 2016
- Vincent Claveau, Ewa Kijak. **Strategies to select examples for Active Learning with Conditional Random Fields***, CICLing 2017 - 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Apr 2017, Budapest, Hungary. pp.1-14.
- Clément Dalloux. **Détection de l'incertitude et de la négation : un état de l'art***, RECITAL 2017 - 18ème Rencontre des Étudiants Chercheurs en Informatique en Traitement Automatique des Langues*, Jun 2017, Orléans, France. pp.1-14
- Clément Dalloux, Vincent Claveau, Natalia Grabar. **Détection de la négation : corpus français et apprentissage supervisé***, SIIM 2017 - Symposium sur l'Ingénierie de l'Information Médicale*, Nov 2017, Toulouse, France. pp.1-8

- Vincent Claveau, Lucas Oliveira, Guillaume Bouzillé, Marc Cuggia, Claudia Cabral Moro *et al.* **Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation***, AIME 2017 - 16th Conference in Artificial Intelligence in Medecine*, Jun 2017, Vienne, Austria. pp.203-208.
- Natalia Grabar, Vincent Claveau. **Critères numériques dans les essais cliniques : annotation, détection et normalisation***, Actes de la conférence TALN 2017*, Jun 2017, Orléans, France
- Natalia Grabar, Vincent Claveau, Clément Dalloux. **CAS: French Corpus with Clinical Cases***, LOUHI 2018 - The 9th International Workshop on Health Text Mining and Information Analysis*, Oct 2018, Bruxelles, Belgium. pp.1-7
- Clément Dalloux, Natalia Grabar, Vincent Claveau, Claudia Moro. **Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien***, TALN 2018 - 25e conférence sur le Traitement Automatique des Langues Naturelles*, May 2018, Rennes, France. 1, Actes de la conférence TALN 2018 - Traitement Automatique de la Langue Naturelle
- Anne-Lyse Minard, Christian Raymond, Vincent Claveau. **IRISA at SMM4H 2018: Neural Network and Bagging for Tweet Classification***, SMM4H 2018 - Social Media Mining for Health Applications, Workshop of EMNLP*, Oct 2018, Brussels, Belgium. Pp.1-8
- Emmanuelle Anceaume, Yann Busnel, E. Shulte-Geers, Bruno Sericola, **Optimization results for a Generalized Coupon Collector Problem**, Journal of Applied Probability, 53(2): 622-629 (2016).
- Emmanuelle Anceaume, Yann Busnel, Bruno Sericola, **New results on a Generalized Coupon Collector Problem using Markov Chains**, Journal of Applied Probability, 52(2): 405-418 (2015).
- Emmanuelle Anceaume, Yann Busnel, **Lightweight Metric Computation for Distributed Massive Data Streams**, Transactions on Large-ScaleData and Knowledge-Centered Systems. Volume 33. 2017.
- Emmanuelle Anceaume, Yann Busnel, Vasile Cazacu. **Finding Top-k Most Frequent Items in Distributed Streams in the Time-Sliding Window Model***, DSN 2018 - 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2018.
- Emmanuelle Anceaume, Yann Busnel, Vasile Cazacu. **On the Fly Detection of the Top-k Items in the Distributed Sliding Window Model**, *NCA 2018 - 17th IEEE International Symposium on Network Computing and Applications*, 2018.
- Emmanuelle Anceaume, Yann Busnel, Vasile Cazacu. **L'art d'extraire des éléments du top-k en temps réel sur des fenêtres glissantes réparties**, *ALGOTEL 2019 - 21èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications*, Jun 2019, Saint Laurent de la Cabrerisse, France
- Natalia Grabar, Cyril Grouin, Thierry Hamon, Vincent Claveau, **Corpus annoté de cas cliniques en français**, *conference TALN 2019*, Toulouse, France
- Natalia Grabar, Cyril Grouin, Thierry Hamon, Vincent Claveau, **Actes de l'atelier et compétition DeFT 2019**, Juillet 2017, Toulouse, France
- Natalia Grabar, Cyril Grouin, Thierry Hamon, Vincent Claveau, **Recherche et extraction d'information dans des cas cliniques. Présentation de la campagne d'évaluation DEFT 2019**, *Proc. of the DeFT 2019 workshop*, Juillet 2019, Toulouse, France
- Clément Dalloux, Vincent Claveau, Natalia Grabar. **Détection de la négation : corpus français et apprentissage supervisé**, *journal TSI - Technique et science informatiques*, to appear 2019
- Cyril Grouin, Natalia Grabar, Vincent Claveau, Thierry Hamon, **Clinical Case Reports for NLP**, *workshop BioNLP 2019*, Nov 2019, Hong-Kong

Submitted

- N. Grabar, C. Dalloux, V. Claveau, **CAS: corpus of clinical cases in French**, Journal of Biomedical Semantics (JBMS)
- C. Dalloux, V. Claveau, N. Grabar, **Speculation and Negation detection in French biomedical corpora**, conference RANLP 2019
- C. Dalloux, V. Claveau, N. Grabar, L Oliveira , C Moro, Y. Gumiel, D. Carvalho, **Cross-lingual and Cross-domain Detection of Negation and of its Scope**, *journal Natural Language Engineering, special issue on Negation*
- Y. Gumiel, L. Oliveira, D. Carvalho, L. Ronnau, A. Pucca da Silva, V. Claveau, N. Grabar, M. Cuggia, C. Dalloux, C. Moro, **Multilingual algorithm for eligibility criteria detection in English and Portuguese Language,** journal *???*
- Y. Gumiel, L. Oliveira, V. Claveau, N. Grabar, E. Paraiso, C. Moro, D. Carvalho, **Temporal Relation Extraction in Clinical texts: A Systematic Review,** *journal ACM survey*

Software

- TagEx: Part-of-Speech tagger for French, especially suited for biomedical/clinical texts, available as a web-service at https://allgo.inria.fr
- NegDetect: detection of negation (cues and scope) for French, available as a webservice at https://allgo.inria.fr
- Age, gender, outcome and admission detection systems (used as baseline for DeFT19), to be made available on https://allgo.inria.fr

Datasets

- CAS: a corpus of clinical cases in French, annotated with keywords, patient information (age, gender), medical information (sign or symptoms, diseases, reason of admission, outcome), linguistic information (Part-of-Speech, UMLS concepts, negation, uncertainty)... Freely available for research purposes.
- Clinical trials: a corpus in English annotated with numerical information (concept that is measured, value, unit, quantifier, temporal validity), negation and uncertainty. Freely available for research purposes.
- ESSAI: idem for French
- TESTES: idem for Portuguese
- eHOP Dataset : corpus of real patient data including discharge summary and ICD10 coding.