BigClin

Big data analytics for unstructured Clinical data

IRISA/Inria LinkMedia
IRISA/Inria Cidre & Dionysos
INSERM/LTSI Health Big Data
CNRS/STL

Follow us:

NLP: detecting negation

Objectives : detecting the scope of negation in clinical texts

- Neural networks: cues and scope
- Annotated corpus in French (cues, scope)
- Comparison to existing systems for English
- Analysis of errors and limits of the approach

Data:

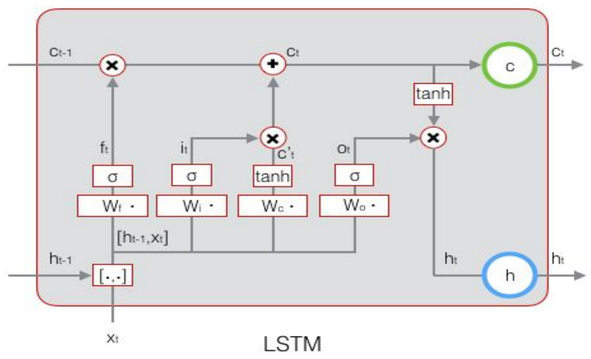
- **protocols of clinical trials**
- **820 negative sentences among 5,394 phrases (inclusion criteria)**

Sources: Gustave Roussy hospital, Institut National du Cancer

- *absence de [ganglion métastatique] / absence of [metastatic lymph node]*
- *En cas d'in[opérabilité] et/ou im[possibilité de réirradier] ,... / In case of in[operability] and / or im[possibility to reirradiate], …*
- *L'[abdomen] est souple et sans [défense]. / The [abdomen] is flexible and [defense]less.*

Negation: methods

Deep learning



LSTM

Instance I(n,c,t) of each word: a vector *n* (word-embedding), a vector *c* (cue-embedding), a vector *t* (PoS-embedding)

Embeddings with dim k=50

Hidden layer with 2x200 units (Bidirectional LSTM)

50 epochs for training

60% of the data for training, 15% validation , 25 % testing

Performance scores: precision P, recall R, F1-score

Negation: results

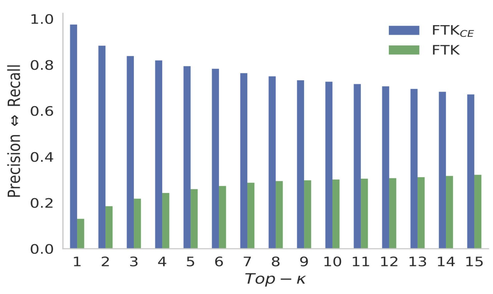
	annotated words			exact scope		
	P	R	F1	P	R	F1
FR Valid	93,72	86,22	89,81	100	72,48	84,04
FR Test	88,29	84,68	86,45	100	53,55	69,75
EN own	91,24	87,10	89,12	100	62,5	76,92
EN Fancellu	92,62	85,13	88,72	99,40	63,87	77,7

- Method tested on French, English, and Brazilian Portuguese
- Foreseen extension to uncertainty (*may, could...*)

Publications/software/dataset:

1. C. Dalloux, N. Grabar, V. Claveau. *Détection de la négation : corpus français et apprentissage supervisé.* p. 1-8. SIIM 2017 (Symposium sur l'Ingénierie de l'Information Médicale)
2. C. Dalloux, V. Claveau, N. Grabar, C. Moro. *Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien.* p. 1-9. TALN 2018
3. Online web-service: https://allgo.inria.fr/webapps/173
4. Dataset: annotated data made available

Distr. computing: stream analytics



FTK_{CE} : Automated detection of the k most frequent elements (Top-k), in massive multi-source distributed data streams, with a sliding window approach to take into account only recent events. Algorithm based on a deterministic counting of over-represented elements, which are identified probabilistically.
Possibility to add a semantic cost function on the received data to extract Top-k data with a specific medical interest.

Publications:

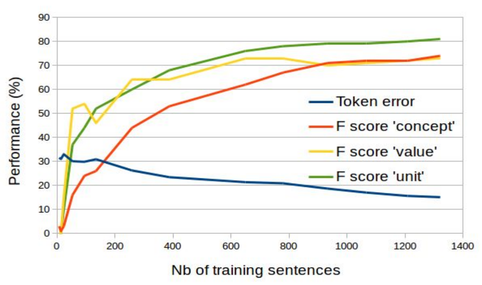
1. E. Anceaume, Y. Busnel, V. Cazacu. *Finding Top-k Most Frequent Items in Distributed Streams in the Time-Sliding Window Model.* DSN 2018 (International Conference on Dependable Systems and Networks)

very large data in the medical field

propose and adapt automatic methods

for their processing and analysis

NLP: detection of numerical values



- Concept, value, unit (*Absolute netrophil count 1,000 cells/l. Exclude if T3 uptake is less than 19%; T4 less than 2.9 (g/dL); free T4 index is less than 0.8.*)
- Normalization of units: cigarettes per day, packs per week

Publications:

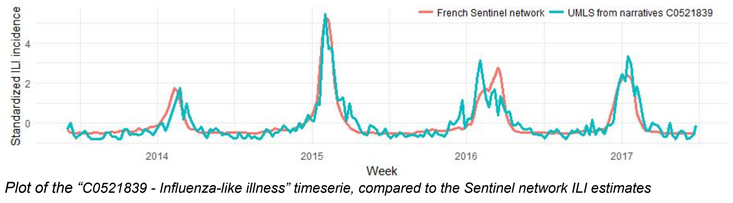
1. N. Grabar, V. Claveau. *Critères numériques dans les essais cliniques : annotation, détection et normalisation.* p. 1-8. TALN 2017 (Traitement Automatique des Langues)
2. V. Claveau, LE Silva Oliveira, G. Bouzillé, M. Cuggia, CM Cabral Moro, N. Grabar. *Numerical eligibility criteria in clinical protocols: annotation, automatic detection and*

Use case 1: real-time syndromic surveillance from EHRs

Aim: To automatically extract medical concepts from narratives to feed forecasting models of Inluenza-like illness (ILI) with the most relevant information.
Methods: Automatic mapping between 40 million narratives from the Rennes academic hospital clinical data warehouse (CDW) and UMLS CUIs: full-text queries of UMLS terms in narratives.
Evaluation: Pearson correlation distance between the reference ILI surveillance signal from the Sentinel network and weekly timeseries from all structured data available in the CDW: ICD-10, Lab results, procedures and UMLS extracted concepts.
Results:

Data sources and terminology	Code and label	Pearson distance
NLP (UMLS)	C0521839: Influenza-like illness	0.32
DRGs (ICD 10)	J11.1: Influenza with other respiratory manifestations, virus not identified	0.48
DRGs (ICD 10)	J10.0: Influenza with pneumonia, other influenza virus identified	0.49
NLP (UMLS)	C0949936: Influenzavirus A	0.53
NLP (UMLS)	C0029347: Influenza A virus	0.53
Lab results (local terminology)	H3: Haemoglutinin H3 gene	0.53

Top 6 concept timeseries most correlated with the ILI timeseries from the Sentinel network



Plot of the "C0521839 - Influenza-like illness" timeserie, compared to the Sentinel network ILI estimates

Use case 2: cohort selection (N2C2)

N2C2: Natural language Processing for Clinical Data challenge

Aim: Cohort selection for clinical trials: Decide if a patient (document) satisfies each of the 13 inclusion criteria

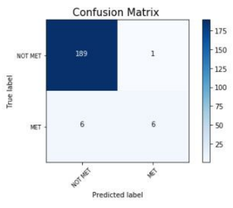
Dataset: 300 sets of longitudinal patient records:
- provided by Partners Healtcare
- annotated by medical professionals

13 inclusion criteria related to: treatments, medication, disorders, health problems, social characteristics and behaviour, lab results
Classes: for each patient and each criteria

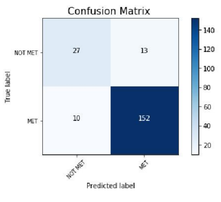
Framework:
- features: identification of medical concepts (CUIs), either bag-of-CUIs or frequency of CUIs, temporality
- machine learning: Naïve Bayes, SVM, Logistic regression, Neural Networks
- rule based: regex, hand-encoded patterns for classes with few or no examples

N2C2: results and future prospects

MI-6MOS (Myocardial infarction in the past 6 months)



ALCOHOL-ABUSE (Current alcohol use over weekly recommended limits)



Official evaluation : main measure - overall

```
********************************** TRACK 1 **********************************
                 ---------- met ----------     ------ not met -------     -- overall ---
                 Prec.   Rec.   Speci.  F(b=1)   Prec.   Rec.    F(b=1)   F(b=1)  AUC
Abdominal        0.8947  0.5667 0.9643  0.6939   0.8060  0.9643  0.8780   0.7860  0.7655
Advanced-cad     0.5938  0.4222 0.6829  0.4935   0.5185  0.6829  0.5895   0.5415  0.5526
Alcohol-abuse    0.3333  0.3333 0.9759  0.3333   0.9759  0.9759  0.9759   0.6546  0.6546
Asp-for-mi       0.8219  0.8824 0.2778  0.8511   0.3846  0.2778  0.3226   0.5868  0.5801
Creatinine       0.9375  0.6250 0.9839  0.7500   0.8714  0.9839  0.9242   0.8371  0.8044
Dietsupp-2mos    0.5652  0.5909 0.5238  0.5778   0.5500  0.5238  0.5366   0.5572  0.5574
Drug-abuse       0.5000  0.3333 0.9880  0.4000   0.9762  0.9880  0.9820   0.6910  0.6606
English          0.9012  1.0000 0.3846  0.9481   1.0000  0.3846  0.5556   0.7518  0.6923
Hba1c            1.0000  0.6000 1.0000  0.7500   0.7846  1.0000  0.8793   0.8147  0.8000
Keto-1yr         0.0000  0.0000 0.9884  0.0000   1.0000  0.9884  0.9942   0.4971  0.4942
Major-diabetes   0.7500  0.6977 0.7674  0.7229   0.7174  0.7674  0.7416   0.7322  0.7326
Makes-decisions  0.9747  0.9277 0.3333  0.9506   0.1429  0.3333  0.2000   0.5753  0.6305
Mi-6mos          0.2500  0.1250 0.9615  0.1667   0.9146  0.9615  0.9375   0.5521  0.5433
                 ------------------------------   ------------------------  ----------
Overall (micro)  0.8177  0.7429 0.8847  0.7785   0.8317  0.8847  0.8574   0.8179  0.8138
Overall (macro)  0.6556  0.5465 0.7563  0.5875   0.7417  0.7563  0.7321   0.6598  0.6514

                                  86 files found
```

Future prospects:
- improve the current results with English clinical data
- adapt the system to Brazilian Portuguese and to French
- test the system on other clinical trials / use cases