An innovative solution for breast cancer textual big data analysis

N. Thiebaut [1], K. Neuberger [1], A. Simoulin[1], I. Ibnouhsein[1], N. Reix[2], M. Lodi[2] and C. Mathelin[2]

1. Quantmetry, 128 rue du Faubourg Saint-Honoré, Paris 75008, France
2. Hôpital de Hautepierre, 1 Avenue Molière, Strasbourg 67200, France

Work in progress, ClinicalTrials.gov Identifier: NCT02810093

Introduction

Hospitals continuously gather huge amounts of textual data in electronic health records. Storing health data in text format is convenient, but without processing, search and analysis operations on such data become tedious.

Here we present an innovative solution for the extraction of structured information out of a corpus of multidisciplinary meeting notes and hospital letters. It relies both on standard text mining methods, on the NegEx [1] algorithm for negation detection and a synonym detection method based on the word2vec [2] algorithm.

Input data: Multidisciplinary meeting notes

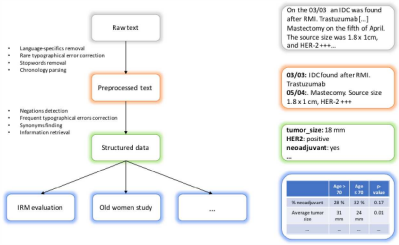
Our corpus is a collection of breast cancer treatment multidisciplinary meeting notes and hospital letters, written in French.

- 10,000 multidisciplinary meeting notes
- Hospitalisations from 2000 to 2016
- Various formats (.doc, .docx) and document structures

All documents were first anonymized, and the project is under the supervision of an ethics committee and CNIL (French National Commission on Informatics and Liberty).

Data transformation pipeline

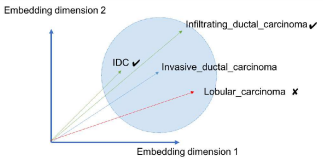
Starting from many Microsoft Word® documents, we have used the python-docx package to obtain JSON files reminiscent of the raw documents structure. Then the JSON files collection was used as input of the pipeline depicted below.



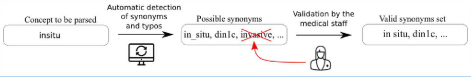
At the end of the pipeline, automatic hypothesis testing was performed to highlight statistically significant differences between sub-populations.

Synonym detection with the word2vec algorithm

To retrieve named entities from the documents, one must be able to handle synonyms, typographical errors and acronyms. To do so, we have adopted an iterative method in which the medical staff provided an initial list of equivalent formulations for all the concepts under investigation. We could further complete this initial list with similar words found using the word2vec algorithm.



Among similar words found using the word2vec word embedding, we find synonyms as expected, but also hyponyms, hyperonyms, antonyms, and semantically related words that appear in similar contexts. Those similar words were presented to the medical staff and validated before incorporation in the synonyms dictionary.



Extraction evaluation

In order to estimate the precision of our structuring methodology, we have compared the extracted values with two different logbooks maintained by the hospital. Comparison with the first logbook gives encouraging preliminary results with 95% agreement.

result example

We have used our structuring program to compare breast cancer characteristics for 7,071 patients older than 50 years old with invasive tumors and no diabete, and 473 patients similar patients with type 2 diabete. The results are shown in the table below.

Indicator	Non diabetic	Type 2 diabetic	p-value
Average tumor size	24.0 mm	27.7 mm	0.001
Estrogen receptor	82.8 %	87.4 %	0.016
Progesterone receptor	68.4 %	72.4 %	> 0.05
HER2+	13.4 %	10.9 %	> 0.05
SBR I grade	27.3 %	20.3 %	0.001
SBR II grade	44.7 %	51.2 %	0.003
SBR III grade	28.0 %	28.5 %	0.048

Conclusion

Our method allows for a versatile text structuring, and can be adapted to any language. We have built lexicons with the help of the medical staff and a word2vec based synonym detection algorithm. This method can be adapted to any corpus, and is complementary to the use of clinical healthcare terminologies such as SNOMED or UMLS.

References

[1] Chapman *et al.*, J Biomed Inform 2001; 34: 301-10.
[2] Mikolov *et al.*, Advances in neural information processing systems, 3111-3119

Reproduction policy