MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220058

Comparison of MetaMap, cTAKES, SIFR, and ECMT to Annotate Breast Cancer Patient Summaries

Akram Redjdal[a], Jacques Bouaud[b,a], Joseph Gligorov[c,d] and Brigitte Séroussi[a,d,e]

[a] *Sorbonne Université, Université Sorbonne Paris Nord, Inserm, UMRS_1142, LIMICS, Paris, France*
[b] *AP-HP, DRCI, Paris, France*
[c] *Sorbonne Université, Institut Universitaire de Cancérologie, Paris, France*
[d] *AP-HP, Hôpital Tenon, Paris, France*
[e] *APREC, Paris, France*

Abstract

Most clinical texts including breast cancer patient summaries (BCPSs) are elaborated as narrative documents difficult to process by decision support systems. Annotators have been developed to extract the relevant content of such documents, e.g., MetaMap and cTAKES, that work with the English language and perform concept mapping using UMLS, SIFR and ECMT, that work for the French language and provide concepts using various terminologies. We compared the four annotators on a sample of 25 French BCPSs, pre-processed to manage acronyms and translated in English. We observed that MetaMap extracted the largest number of UMLS concepts (15,458), followed by SIFR (3,784), ECMT (1,962), and cTAKES (1,769). Each annotator extracted specific valuable information, not proposed by the other annotators. Considered as complementary, all annotators should be used in sequence to optimize the results.

Keywords:

Information Extraction, Decision Support, Breast Cancer

Introduction

Multidisciplinary tumor boards (MTBs) provide a collaborative and multidisciplinary approach to cancer care, bringing together surgery, oncology, radiology, and pathology specialists to optimize decision making and care coordination. But, the benefits of tumor boards, which have long been taken for granted, are recently being challenged. Positive outcomes from MTBs depend on the presence of qualified and effective faculty, good preparation of patient cases, format and structure of the meeting, efficient leadership, and contributive interactions among MTB physicians [1].

Clinical decision support systems (CDSSs) are software components that aim to support clinicians in their decision-making process. CDSSs have been used in many medical areas. They have proven to improve the quality of patient care by increasing the compliance of clinician decisions with clinical practice guidelines (CPGs) [2]. However, CDSS routine use for decision-making shows varied performance and needs to be further studied [3].

DESIREE [4] is a European project that aims at developing web-based services for the management of primary breast cancer by MTBs. While evaluating the guideline-based component of DESIREE, we found that for some patient cases the system did not provide recommendations or gave inappropriate recommendations [5]. These patient cases were considered as "complex cases".

Assuming that "one size does not fit all", our goal is to create a decision support system that includes two modules: (i) a case-based decision support module for complex clinical cases that builds on the recall of similar patient cases and of the decisions previously made to produce treatment plans, and (ii) a classical guideline-based system for the non-complex cases. This system is expected to be used at MTBs of the Tenon hospital (AP-HP, Paris, France).

Decision support systems usually rely on structured clinical data processing. However, breast cancer patient summaries (BCPSs) used to orally present a patient history during MTBs are expressed as natural language clinical notes. Because, we were lacking an annotated corpus of BCPSs, we decided to use automatic semantic annotators and indexers to structure the relevant content of natural language BCPSs. Currently, two systems [6] are widely used in the biomedical field for the English language, MetaMap [7] and cTAKES [8]. Since we will work on a corpus of French BCPSs, we considered two systems that work for the French language, i.e., SIFR [9] and ECMT [10].

We implemented the four annotators and compared the structured data produced by each of them on a sample of BCPSs used in French with the annotators able to process French summaries, and with the same sample of BCPSs translated beforehand in English for the annotators able to process English summaries. The aim was to assess whether annotators' outputs were redundant or complementary and to identify the best one.

Methods

Annotator tools

MetaMap [7] was developed by the National Library of Medicine to map biomedical texts to concepts in the Unified Medical Language System (UMLS). The tool uses a hybrid approach combining natural language processing, knowledge-intensive approach, and computational linguistic techniques [11].

cTAKES [8] (Clinical Text Analysis and Knowledge Extraction System) uses rule-based and machine learning techniques to extract information from clinical text. Both MetaMap and cTAKES use UMLS to extract and standardize medical concepts.

ECMT (Extracteur de Concepts Multi-Terminologique – http://ecmt.chu-rouen.fr) is a web service inspired by the CISMef algorithm for information retrieval with the Doc'CISMeF search engine and F-MTI [10] which is a multi-terminology automatic indexer. ECMT works for the French language. There are two query modules: one default module based on a bag of words algorithm [10] and one expanded module based on textual indexing, using Oracle text indexing. ECMT does not allow one to choose the ontology to be used by the annotation process. The annotator works with seven terminologies, and supports semantic expansion features [12].

SIFR (Semantic Indexing of French Biomedical Data Resources - http://bioportal.lirmm.fr/annotator) Annotator [9] is an openly available web service enabling both recognition and contextualization of concepts from 30 medical terminologies and ontologies. The annotator service processes textual descriptions, tags them with relevant biomedical ontology concepts including UMLS, expands the annotations using the knowledge embedded in ontologies, and contextualizes the annotations before returning them to the users in several formats .

Corpus of breast cancer patient summaries

We had access to a sample of 643 BCPSs available as textual unstructured documents. They provide a portrait of patients with all relevant information that MTB clinicians need to know to make the best patient-specific therapeutic decision. BCPSs contain different types of information: reason for presentation, type of tumor, biometric data, personal history, family history, TNM classification, etc. However, unstructured formats make information extraction complicated. First, there are many abbreviations, acronyms, and specialized terms. Secondly, a variety of terms may be used, that may not correspond to a general domain [6], depending on the health professional specialty of the BCPS's author.

Pre-treatment and text translation

In order to run cTAKES and MetaMap, we had to translate BCPSs from French to English. However, there were a lot of acronyms related to the oncological field, and it was nearly impossible for a general automated translator to find the right translation for acronyms. To solve this issue, we performed web scraping to search for French medical acronyms and their signification. We found such information mainly on Wikipedia and DoctoLib's Dictionary. We created a local dictionary with medical acronyms and their definition. Then, we replaced acronyms in the BCPSs with their definition in the dictionary to obtain a "translatable" text. We finally used a pre-trained Opus-MT [13] translation model. As a result, all BCPSs were available in French and English in a text format (.txt) which is the input format for the four annotators.

Executing the annotators on the BCPS corpus

Many components bundled in cTAKES are available and can be used in multiple ways. We used the Default Clinical Pipeline, which produced the most commonly desired output. This includes annotations for anatomical sites, signs and symptoms, procedures, diseases and disorders, and medications. For each annotation, normalized UMLS CUIs are provided, as well as markers for negation, uncertainty, and subject.

MetaMap maps text parts to concepts from the UMLS Metathesaurus. Text is processed through a series of modules and broken down into components that include sentences, phrases, lexical elements, and tokens. Variants are generated from the resulting phrases, and candidate concepts from the UMLS Metathesaurus are retrieved and evaluated against their phrases, taking into account negation.

ECMT is licensed by the University Hospital of Rouen, France and it was possible to process data locally. The ECMT output is an XML file that contains the identified concepts in the seven terminologies used by ECMT. The current version does not provide UMLS CUIs directly nor contextual information (such as negativity).

SIFR bioportal is openly accessible and offers docker packaging. The annotator can be used via a REST-API, the workflow generates a final json-ld output or converts it to different formats (e.g., BRAT). The result contains UMLS CUIs for each extracted concept and includes a module for identifying negations [14].

Post-treatment of the outputs of each system

cTAKES outputs are XMI files, readable using the UIMA CAS Visual Debugger (CVD). We processed the files using a Python parser and generated a table with four columns: extracted concept, UMLS CUI, negation, and subject.

For MetaMap, we used the Python wrapper Pymm for extracting candidate and mapping concepts. Pymm parses the XML output of MetaMap. Extracted information included matched word, UMLS CUI, negation, semantic type.

For the two French annotators, we implemented a script that takes the output of each system and processes it:

- SIFR: the UMLS CUI, the matched concept, and negation were extracted from the json files
- ECMT: the matched concept, the terminology from which it was retrieved, and its code in that terminology were extracted from the XML output.
  In order to compare the results by using the UMLS CUIs to avoid problems in the translation of the matched text, we reused SIFR on the words and phrases matched by ECMT so we could obtain, if available, the UMLS CUIs for the concepts extracted by ECMT.

Comparison of annotators

Having the UMLS CUI for each concept extracted by each annotator, we compared the results produced by each system. We identified which concepts were extracted and by which annotator, and also studied the intersections of the coverage of each annotator by BCPS. For each BCPS, we computed:

- The number of concepts extracted by each annotator
- The number of concepts retrieved by each combination of annotators that forms the following partition:

  ECMT alone
  SIFR alone
  cTAKES alone
  MetaMap alone
  ECMT + SIFR
  ECMT + cTAKES
  ECMT + MetaMap
  SIFR + cTAKES
  SIFR + MetaMap
  MetaMap + cTAKES
  ECMT + SIFR + cTAKES
  ECMT + SIFR + MetaMap
  ECMT + MetaMap + cTAKES

MetaMap + SIFR + cTAKES
ECMT + SIFR + cTAKES + MetaMap

Figure 1 shows the pipeline used to run the four systems on BCPSs.

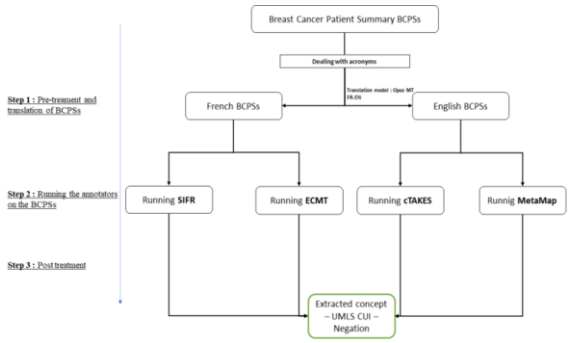

Figure 1: Pipeline for executing the four annotators on French BCPSs

Results

The four annotators were compared on a sample of 25 BCPSs randomly extracted from the corpus. Results are shown in Table 1 for four random BCPSs among the 25.

On the 25 BCPS-selection, we observed that MetaMap was the system that extracted the largest number of UMLS concepts (15,458), SIFR is the annotator with the second higher number of concepts extracted (3,784), followed by ECMT (1,962), and cTAKES (1,769).

Coverage intersections between systems for the same language (e.g. ECMT + SIFR, or cTAKES + MetaMap) are more frequent than in case of "multilingual" intersection (e.g. ECMT + cTAKES, or SIFR + MetaMap, etc.) (See Table 1).

Table 1. Comparison of the different annotators on four different BCPSs.

# concepts	BCPS1	BCPS2	BCPS3	BCPS4
ECMT alone	45	147	42	90
SIFR alone	91	261	84	192
cTAKES alone	52	109	30	59
MetaMap alone	487	1030	309	542
ECMT + SIFR	45	146	42	89
ECMT + cTAKES	17	35	14	22
ECMT + MetaMap	28	75	19	53
SIFR + cTAKES	24	44	18	30
SIFR + MetaMap	45	106	30	74
MetaMap + cTAKES	48	99	27	56
ECMT + SIFR + cTAKES	17	35	14	22
ECMT + SIFR + MetaMap	28	75	19	53
ECMT + MetaMap + cTAKES	17	35	13	21
MetaMap + SIFR + cTAKES	24	44	17	29
ECMT + SIFR + cTAKES + MetaMap	17	35	13	21

Within the sample of 25 BCPS , a total of 184 different UMLS CUIs were detected by all four systems. Figure 2 shows the distribution of these 184 concepts grouped by their UMLS semantic type.

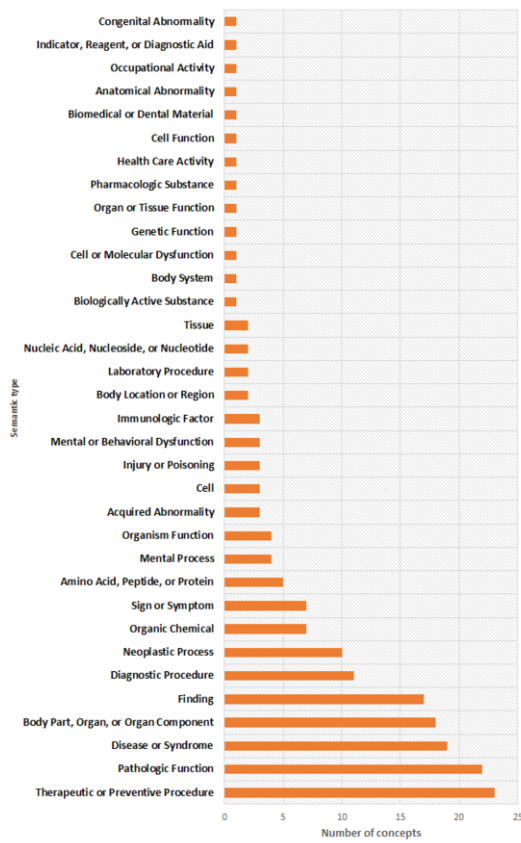

Figure 2: Distribution of the concepts extracted by the four systems.

Discussion

When analysing all extracted concepts, we found out that, in addition to retrieving the greatest number of concepts, MetaMap extracted multiple useful data for the cancer domain, e.g.,

hormonal receptors values or tumor size when mentioned in the text.

cTAKES did not exhibit as good results as MetaMap on our selection but this might be explained by the fact that we used the default clinical pipeline. cTAKES allows to create and customize pipelines according to the context. With adaptations, it could therefore deliver better results, since it has already demonstrated good performance for a breast cancer extraction system [15].

We also observed from the intersections between annotators of the same language that they have a shared conceptual coverage, larger than between annotators of different languages. For the French language, SIFR extracted most of the concepts extracted by ECMT, but also retrieved more concepts. The reason may originate from the large number of medical terminologies it uses.

The fact that the version of ECMT we used did not provide the UMLS CUIs was constraining for comparing annotators, but using SIFR to extract CUIs was a practical solution to overcome this difficulty.

Finally, concepts that were extracted by all systems for all patients represent common concepts that can be systematically found in clinical summaries of breast cancer patients. We noticed that the most frequent concepts extracted by the four annotators were represented by therapeutic procedures (chemotherapies, therapeutic radiology procedures, mastectomy, etc.), pathologic functions (axillary lymphadenopathy, inflammation, etc.), diagnostic and imaging procedures (mammography, MRI, biopsy, etc.), body parts and associated diseases, and that the less frequent concepts were related to tissues, cell functions, or lesion information. This latter point suggests that such information, which is indeed of critical importance for cancer care, might be expressed in varied ways that not all annotators detect.

It is also important to mention that each annotator can extract different specific categories of information about the patient:

- MetaMap could produce more information like HER status, hormonal receptors and numeric information like tumor size.
- ECMT can annotate phrases that are specific to the French language since it uses French terminologies for the indexation.
- SIFR offers access to multiple terminologies and ontologies, including those that are specific to the oncological field (e.g., MuEVo).
- cTAKES can be customized according to the clinical needs and can be performant in terms of breast cancer clinical data extraction (DeepPhe pipeline [15]).

The raw use of annotators of the full content of BCPSs only provides a list of concepts, asserted or negated (except for ECMT at present), but misses contextual information that could support a better data structuration. In practice, BCPSs are structured and follow a template including sections like: identity of the patient, reason of presentation, type of tumor, biometric data, personal history, family history, usual treatments, anamnesis, clinical examination, lesion assessment, TNM classification, neo-adjuvant treatment, response to the neo-adjuvant treatment, surgery, anatomic pathology of surgery, adjuvant treatment, proposed care plan, and MTB final decision.

Using this knowledge on BCPS contents, we could use the annotators with a particular focus that would help to interpret the extracted concepts and build a detailed structured data representation, which could be further used by software components like CDSSs or artificial intelligence algorithms.

Conclusions

To extract valuable information from BCPSs written in French language, the use of automatic annotators can be time saving and efficient. We compared four biomedical text annotators on a sample of 25 BCPSs. Despite variations in performance, we believe that the combination of annotators for the French language, and even English annotators, along with aggregations, might permit retrieval of complementary information. However, the result of such annotators might be enhanced by additional knowledge of textual contents to build a better structured data representation and efficiently feed decision support components.

Acknowledgements

The authors thank the University Institute of Health Engineering (IUIS – Sorbonne University) for financing this research and the AP-HP health data warehouse for supporting this work.

References

[1] El Saghir NS, Keating NL, Carlson RW, Khoury KE, Fallowfield L. Tumor boards: optimizing the structure and improving efficiency of multidisciplinary management of patients with cancer worldwide. Am Soc Clin Oncol Educ Book. 2014:e461-6. doi: 10.14694/EdBook_AM.2014.34.e461. PMID: 24857140.

[2] Kwan JL, Lo L, Ferguson J, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ. 2020;370:m3216.

[3] Beauchemin M, Murray MT, Sung L, Hershman DL, Weng C, Schnall R. Clinical decision support for therapeutic decision-making in cancer: A systematic review. *Int J Med Inform*. 2019;130:103940.

[4] Bouaud J, Pelayo S, Lamy JB, Prebet C, Ngo C, Teixeira L, Guézennec G, Séroussi B. Implementation of an ontological reasoning to support the guideline-based management of primary breast cancer patients in the DESIREE project. Artif Intell Med. 2020 Aug;108:101922.

[5] Redjdal A, Bouaud J, Guézennec G, Gligorov J, Seroussi B. Reusing Decisions Made with One Decision Support System to Assess a Second Decision Support System: Introducing the Notion of Complex Cases.

[6] Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak. 2018 Sep 14;18(Suppl 3):74. doi: 10.1186/s12911-018-0654-2. PMID: 30255810; PMCID: PMC6157281.

[7] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *JAMIA*. 2010;**17**(3):229–236

[8] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13. doi:

10.1136/jamia.2009.001560. PMID: 20819853; PMCID: PMC2995668.

[9] Sakji S, Gicquel Q, Pereira S, Kergourlay I, Proux D, Darmoni S, Metzger MH. Evaluation of a French medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. Stud Health Technol Inform. 2010;160(Pt 1):252-6. PMID: 20841688..

[10] Tchechmedjiev A, Abdaoui A, Emonet V, Zevio S, Jonquet C. SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. BMC Bioinformatics. 2018 Nov 6;19(1):405. doi: 10.1186/s12859-018-2429-2. PMID: 30400805; PMCID: PMC6218966.

[11] Aronso A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annu Symp Proc. 2001;2001:17–21.

[12] Pereira S, Névéol A, Kerdelhué G, Serrot E, Joubert M, Darmoni S. Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. AMIA symp. 2008:586-590.

[13] J. Tiedemann et S. Thottingal, « OPUS-MT – Building open translation services for the World », in Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisboa, Portugal, nov. 2020, p. 479 480

[14] Mirzapour M, Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. French FastContext: A publicly accessible system for detecting negation, temporality and experiencer in French clinical notes. J Biomed Inform. 2021 Mar 15;117:103733. doi: 10.1016/j.jbi.2021.103733. Epub ahead of print. PMID: 33737205.

[15] Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, Harris D, Hochheiser H, Lin C, Chavan G, Jacobson RS. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. Cancer Res. 2017 Nov 1;77(21):e115-e118. doi: 10.1158/0008-5472.CAN-17-0615. PMID: 29092954; PMCID: PMC5690492.

Address for correspondence

Akram Redjdal
LIMICS U1142
15, rue de l'école de médecine,
75006 Paris, France
E-mail: redjdalakram300@gmail.com