Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-126

Evaluation of the Terminology Coverage in the French Corpus LiSSa

Chloé CABOT[a], Lina F. SOUALMIA[a,b], Julien GROSJEAN[a], Nicolas GRIFFON[a,b] and Stéfan J. DARMONI[a,b,1]
[a] *Normandie Univ., TIBS - LITIS EA 4108, Rouen University and Hospital, France*
[b] *French National Institute for Health, INSERM, LIMICS UMR-1142, France*

Abstract. Extracting concepts from medical texts is a key to support many advanced applications in medical information retrieval. Entity recognition in French texts is moreover challenged by the availability of many resources originally developed for English texts. This paper proposes an evaluation of the terminology coverage in a corpus of 50,000 French articles extracted from the bibliographic database LiSSa. This corpus was automatically indexed with 32 health terminologies, published in French or translated. Then, the terminologies providing the best coverage of these documents were determined. The results show that major resources such as the NCI and SNOMED CT thesauri achieve the largest annotation of the corpus while specific French resources prove to be valuable assets.

Keywords. Information extraction, Semantics, Natural Language Processing, Data storage and retrieval, Vocabulary controlled

1. Introduction

Indexing medical documents such as clinical reports as well as biomedical articles is a key to various information retrieval tasks in medical information management. Automatic indexing can deal with the increasing amount of new material being produced in biomedical fields that has made manual indexing slow and expensive. Annotating medical documents and the following applications is actually a frequent topic in English-speaking scientific literature. Various annotating tools are available for English text, as well as the resources provided by the National Library of Medicine in association with the Unified Medical Language System (UMLS). Several vocabulary-controlled approaches for indexing documents have been proposed. Aronson et al. use MetaMap and the tri-gram method to extract UMLS terms, and then refine them to MeSH concepts [1]. Natural Language Processing (NLP) techniques can be also applied to annotate documents with UMLS [2]. Gurulingappa et al. use the JSRE system combining Support Vector Machines (SVMs) with different kernels specially designed for the NLP and relation extraction [3]. Vector space model (VSM) is also a common approach that can be mixed with NLP techniques. Jonnalagadda et al. adopt this approach to identify UMLS concepts in the i2b2/VA concept extraction corpus [4]

[1] Corresponding author, Stéfan J. DARMONI, Rouen University Hospital, 1 rue de Germont, 76000 Rouen, France; E-mail: stefan.darmoni@chu-rouen.fr

French-speaking texts do not benefit from such various tools and resources. French is lowly represented in the UMLS [5] As provided in the 2016AA release, the French UMLS thesaurus manages 9 resources while 128 resources are available in English, providing a French concept for 85,685 concept unique identifiers. Only 3.11% of English UMLS terms are available in French and while each English term has an average of 2 synonyms, only 1.54 synonyms are available for each French term.

Since 2005, our team develops the Health Terminology/Ontology Portal (HeTOP) [6] providing an access to 55 terminologies in French and English, partially translated into French. A major application of this multi-lingual portal includes a multi-terminology automatic indexing tool called ECMT [8] based on HeTOP resources.

The aim of this study is to analyze the coverage of 32 terminologies available in French in the HeTOP on the French medical corpus LiSSa [7]. These 32 terminologies were selected among the 55 available terminologies as the most relevant for this task. This corpus was indexed with the ECMT tool to help reduce (i) the amount of terminologies used in automatic indexing, (ii) the noise generated by using multiple terminologies, especially with some specific types of concepts and (iii) the amount of redundant concepts.

2. Methods

2.1. Automatic Indexing with ECMT

The ECMT tool is designed to identify clinical concepts in biomedical documents using terminologies included in HeTOP. ECMT relies on the "bag-of-words" algorithm and also on pattern-matching designed for discharge summaries, procedure reports or laboratory results which contain symbolic data (presence or absence), numerical data and units of measurement [8].

Each concept identified in a document and its metadata (the concept type, original identifier, terminology) was stored for subsequent analysis. Prior to the analysis of the coverage, the indexing terms, which presented the highest occurrence frequencies throughout the corpus, were manually reviewed to detect common and regular indexing errors, and excluded in relevant cases.

2.2. The French Medical Corpus LiSSa

The corpus of the bibliographic database LiSSa[2] contains more than 850,000 articles in French. Among them 50,000 articles were randomly selected and each title, abstract and set of keywords were indexed using the ECMT tool with 32 terminologies. These resources as well as the versions used are available in HeTOP[3]. The source language of these resources varies: 13 terminologies are published originally in French while 19 have been totally or partially translated.

[2] http://www.lissa.fr
[3] http://www.hetop.eu

3. Results

The amount of all concept occurrences identified in each terminology is determined for each document category: titles, abstracts and sets of keywords. The results are detailed in Table 1 and Figure 1. Distinct concepts (i.e. counted only once)

Table 1. Terminology coverage of the French corpus LiSSa for each document category.

Titles		Abstracts		Keywords	
Terminology	**Concepts**	**Terminology**	**Concepts**	**Terminology**	**Concepts**
NCIt	150,224	NCIt	2,040,356	NCIt	52,423
MeSH	106,170	SNOMED CT	1,543,456	MeSH	50,937
SNOMED Int.	96,771	MeSH	1,238,133	TSP	47,237
SNOMED CT	95,409	TSP	1,089,331	SNOMED Int.	45,879
TSP	84,989	SNOMED Int.	827,714	SNOMED CT	36,771
MedDRA	45,164	LOINC	502,964	MedDRA	25,802
LOINC	36,483	MedDRA	395,434	LOINC	15,491
FMA	24,300	FMA	182,398	ICNP	13,585
ICNP	24,244	ICNP	161,341	FMA	8,819
ICD-10	14,022	CLADIMED	87,924	HPO	7,633
Others	77,273	Others	641,766	Others	40,583
Total	755,049	Total	8,710,817	Total	345,160

identified in each terminology were also determined for each document category. The results are detailed in Table 2.

The five terminologies obtaining the most indexing terms in each document category, NCIt, SNOMED CT, SNOMED Int., MeSH and TSP are consistently the same for each group. The NCI thesaurus obtains the best coverage in all document categories, while the Thésaurus Santé Publique (TSP), a French Public Health thesaurus is the only French ressource to appear in the first third of the ressource ranking. More specialized resources such as HRDO (rare diseases) or ATC (chemical therapeutics) achieve much less coverage than expected. The five first terminologies giving the best coverage of the corpus add up 65% to 70% of the whole indexing term set depending on the document category. However, some smaller resources published originally in French achieve a good coverage of the corpus in spite of a limited French indexing terms. These resources such as the CISMeF thesaurus [9] or the Q-Codes classification [10] are actually developped to fit clinical and non-clinical information in abstracts and complete larger terminologies as the MeSH or the SNOMED Int.

Table 2. Terminology coverage by distinct concepts of the French corpus LiSSa for each document category.

Titles		Abstracts		Keywords	
Terminology	**Concepts**	**Terminology**	**Concepts**	**Terminology**	**Concepts**
MeSH	11,324	MedDRA	19,281	MeSH	5,908
SNOMED Int.	10,396	SNOMED Int.	17,968	SNOMED Int.	4,290
MedDRA	8,644	MeSH	17,353	NCIt	4,249
SNOMED CT	7,996	SNOMED CT	17,192	MedDRA	4,024
NCIt	7,247	NCIt	12,683	SNOMED CT	3,491
TSP	3,617	TSP	5,799	TSP	2,801
LOINC	1,822	FMA	3,903	LOINC	1,012
FMA	1,815	LOINC	3,372	FMA	935
ICD-10	1,758	HPO	2,997	ICD-10	870
HPO	1,488	ICD-10	2,812	HPO	823
Others	7,186	Others	11,612	Others	4,082
Total	63,293	Total	114,982	Total	32,485

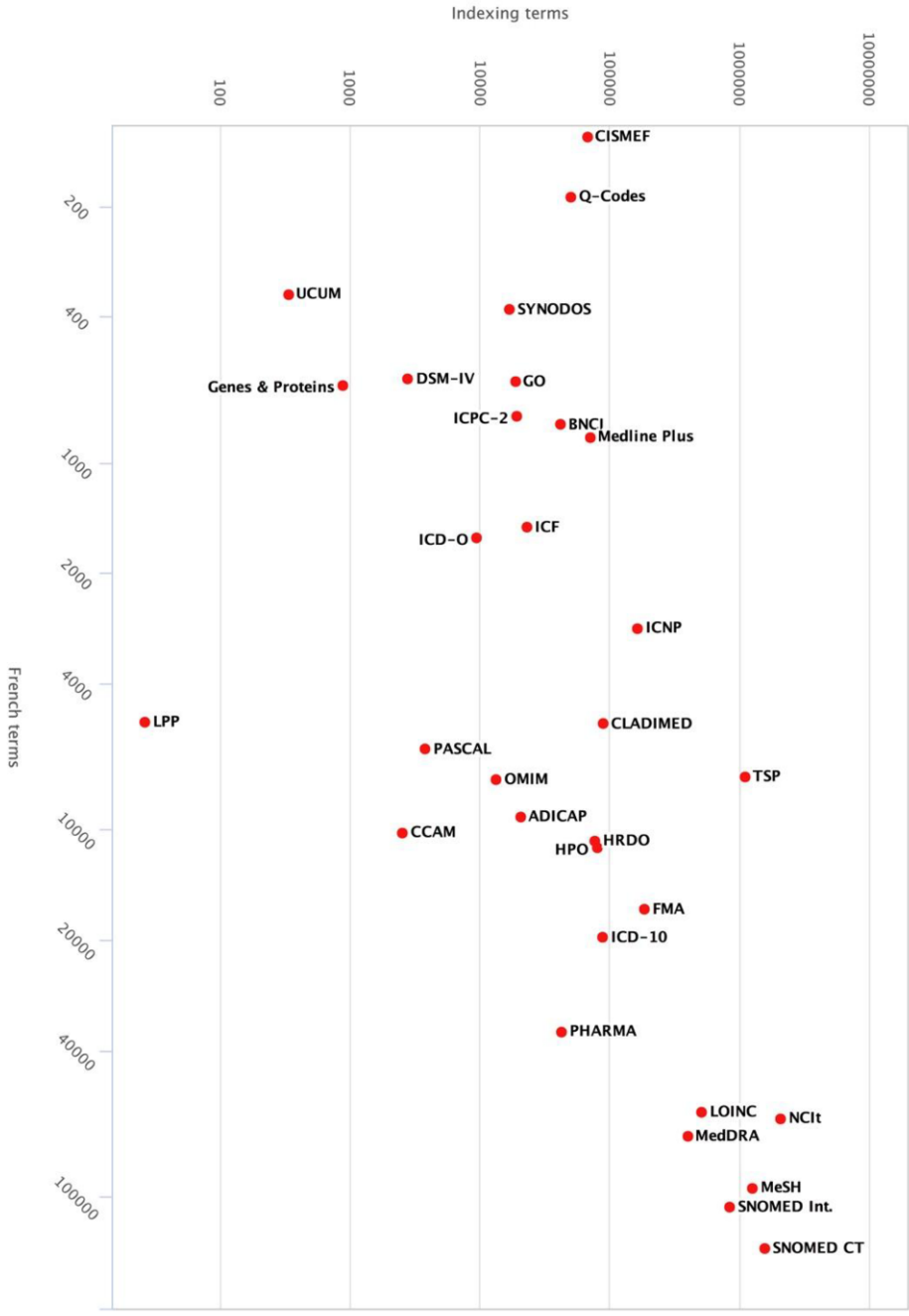

Figure 1. Terminology coverage of article abstracts in the LiSSa corpus.

4. Discussion and Conclusion

LiSSa is a bibliographic database in French providing a large corpus of titles, abstracts and authors' keywords. ECMT was able to annotate these three corpora. Overall, NCIt is surprisingly ranked first in the three corpora although only 60000 terms are now translated in French while over 90000 are translated for the MeSH and 137000 for SNOMED CT. When analyzing with distinct concepts, the ranking is very different (MeSH ranked first for titles, MedDRA for abstracts). This coverage of distinct concepts should be refined at the concept scale to evaluate the concept redundancy between the top ten ranked resources. This study is still ongoing and a phase of manual annotation of the corpus by field experts to validate the automatic indexing results and refine these observations is currently processed. In the near future, we will reproduce the same study on a corpus of discharge summaries. Terminologies developed for this purpose should be better ranked, in particular SNOMED CT and ICD-10.

5. Acknowledgments

The LiSSa project was partially granted by the ANR TecSan program (ANR-14-CE17-0020).

References

[1] Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, and Willie J Rogers. The NLM Indexing Initiative's Medical Text Indexer. *Studies in health technology and informatics*, 107(Pt 1):268–272, 2004.

[2] Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics*, 2 Suppl 5(Suppl 5):S4, October 2011.

[3] Harsha Gurulingappa, Abdul Mateen-Rajput, and Luca Toldo. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):15, December 2012.

[4] Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics*, 45(1):129–140, February 2012.

[5] A Névéol, J Grosjean, S J Darmoni, and P Zweigenbaum. Language Resources for French in the Biomedical Domain. *LREC*, 2014.

[6] J Grosjean, T Merabti, and B Dahamna. Health multi-terminology portal: a semantic added-value for patient safety. *Stud Health Technol*, 2011.

[7] Nicolas Griffon, Matthieu Schuers, Lina Fatima Soualmia, Julien Grosjean, Gaetan Kerdelhué, Ivan Kergourlay, Badisse Dahamna, and Stéfan Jacques Darmoni. A Search Engine to Access PubMed Monolingual Subsets: Proof of Concept and Evaluation in French. *Journal of Medical Internet Research*, 16(12):e271, December 2014.

[8] L F Soualmia, C Cabot, B Dahamna, and S J Darmoni. SIBM at CLEF e-Health Evaluation Lab 2015. 2015.

[9] Magaly Douyère, Lina F Soualmia, Aurelie Neveol, Alexandrina Rogozan, Badisse Dahamna, Jean-Philippe Leroy, Benoit Thirion, and Stéfan J Darmoni. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Information & Libraries Journal*, 21(4):253–261, 2004.

[10] M Jamoulle. Using the International Classification for Primary Care (ICPC) and the Core Content Classification for General Practice (3CGP) to classify conference abstracts. *Rev Port Med Geral Fam*, 2013.