Published in Working Notes of CLEF 2016 (Conference and Labs of the Evaluation forum), Évora, Portugal, 5-8 September, 2016, which should be cited to refer to this work.

BiTeM at CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction

Luc Mottin[1,2,*], Julien Gobeill[1,2], Anaïs Mottaz[1,3], Emilie Pasche[1,2], Arnaud Gaudinat[1,2], Patrick Ruch[1,2]

[1]BiTeM group, HES-SO/HEG Geneva, Information Science Department, 17 rue de la Tambourine, CH-1227, Carouge, Switzerland
[2]SIB Text Mining, Swiss Institute of Bioinformatics, 1 rue Michel-Servet, CH-1206, Genève, Switzerland
[3]HUG, Geneva University Hospitals, 4 rue Gabrielle-Perret-Gentil, CH-1205, Genève, Switzerland
*Corresponding author: Tel: +41 22 38 81 848; Fax: +41 22 38 81 701; Email: luc.mottin@hesge.ch

Abstract: BiTeM/SIB Text Mining (http://bitem.hesge.ch/) is a University research group carrying over activities in semantic and text analytics applied to health and life sciences. This paper reports on the participation of our team at the CLEF eHealth 2016 evaluation lab. The processing applied to each evaluation corpus (QUAREO and CépiDC) was originally very similar. Our method is based on an Automatic Text Categorization (ATC) system. First, the system is set with a specific input ontology (French UMLS), and ATC assigns a rank list of related concepts to each document received in input. Then, a second module relocates all of the positive matches in the text, and normalizes the extracted entities. For the CépiDC corpus, the system was loaded with the Swiss ICD-10 GM thesaurus. However a late minute data transformation issue forced us to implement an ad hoc solution based on simple pattern matching to comply with the constraints of the CépiDC challenge. We obtained an average precision of 62% on the QUAREO entity extraction (over MEDLINE/EMEA texts, and exact/inexact), 48% on normalizing this entities, and 59% on the CépiDC subtask. Enhancing the recall by expanding the coverage of the terminologies could be an interesting approach to improve this system at moderate labour costs.

Key words: Named-Entity Recognition, Automatic Text Categorization, Discontinuous Entity Extraction, Relocation, Statistical Training, Concept Normalization, UMLS, ICD-10.

1. Introduction

Biomedical data involves a large diversity and quantity of valuable knowledge for the medical research and practice. Thus, text-mining tools such as named-entity recognizers have been developed to effectively and efficiently access textual contents. Now, a dynamic way to improve the different systems implies to compare them on specific shared tasks as in CLEF such as in [1-3]. In 2016, the challenge was divided in three subtasks: entity recognition and normalization on the QUAREO corpus, and entity extraction on the CépiDC corpus, plus a replication track [4-5]. Both of the corpora are available in French and related to biomedical literature.

We report in this paper the contribution of the group to the eHealth task 2 (Multilingual Information) within the CLEF 2016 competition. Our team participated to most of these tracks, including MEDLINE and EMEA entity extraction (respectively labelled 2.Q.1 and 2.Q.2), MEDLINE and EMEA normalizedEntities (2.Q.3 and 2.Q.4), the CépiDC coding (2.C), and the replication track.

Our approach was to integrate an existing automatics categorizer (Ruch 2006) in the processing of corpora. By providing a ranked list of concepts for each unit of a corpus, we aim at testing the accuracy of this tool within a Named-Entity Recognition (NER) task.

2. Methods

2.1. QUAREO

2.1.1. Data

The QUAREO French medical corpus provided for this task includes two datasets [6]. The first one is composed of 833 article titles from MEDLINE. The second dataset contains four sets of instructions for use of medicines from the European Medicines Agency (EMEA), which are separated in 15 free-texts. Additionally, two other datasets were previously supplied to train, evaluate and adjust the systems.

Designed with a controlled language and strict rules, EMEA instructions represent a good assessment for the extraction of entities blurred into free-text. MEDLINE extracts contain fewer concepts, but might be a challenge since they come from different authors and journals that imply diverse writing style.

The Unified Medical Language System (UMLS) is a compilation of ontologies and software or services [7-8]. Required for the entity normalization, we used the standard French release of the UMLS Metathesaurus as exclusive dictionary to extract the biomedical entities with their Unique Concept Identifiers (CUIs) [9]. Thus, to set up our application we handled the release, freely available in April 2016, from the National Institute of Health website (www.nlm.nih.gov). With 397203 entries including synonyms, and 139771 unique concept, this terminology regroups concepts from nine sources in their French versions; Table 1 presents the distribution for each source.

Table 1 : Distribution of terms in the French UMLS Metathesaurus.

Source	# terms
MSHFRE	112571
MDRFRE	97896
LNC-FR-FR	88306
LNC-FR-BE	44451
LNC-FR-CA	42766
LNC-FR-CH	4940
WHOFRE	3717
MTHMSTFRE	1833
ICPCFRE	723

Ten groups of clinical entities are defined from the UMLS semantic types to provide a consistent categorization of biomedical concepts and support their normalization [10-11]. These semantic groups are: Anatomy, Chemicals & Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, and Procedures. Aware of nested entities that could be assigned to different groups [2], we used the training data to statistically assign the semantic types to the ten categories. Regarding the semantic types with no mapping, or weakly expressed, we decided to manually curate the corpus.

2.1.2. Automatic Text Categorization

CLEF 2016 was the opportunity to evaluate a tool we worked on several years ago [12]. Based on a specific thesaurus, the Categorizer operates on each text of the corpus one by one, and provides ranked lists of concepts. This ranking process combines a regular expression classifier with a vector-space classifier described in Ruch (2016).

2.1.3. Entity relocation

The second phase of our system aims at matching the new list of concepts with the input text using patterns. Each line is divided in word tokens and the program considers that multiple-words entities can be discontinuous, with one or many nested words. Concretely, the system will successively try to find each term from the biomedical concept identified with the line tokens. This also implies to take care of repeated and overlapping entities. When a prediction is completely retrieved in the text, the system recovers the offset position (positions of the first and last characters), and prepares a new entry in the output respecting the BRAT format.

2.1.4. Entity normalization

Normalization was processed directly with the matching. As the ATC predict a list of possible entities derived from the UMLS concepts, UMLS CUIs are associated with every proposition. Thus, for each prediction matched in the text, the system can immediately assign a unique CUI.

2.2. CepiDC

2.2.1. Data

The CépiDC corpus compiles 110869 lines related to causes of death, and reported by physicians, within a single CSV file. The corpus is structured in such a way that one sentence is repeated when multiple causes should be distinctly encoded.

The International Classification of Diseases (ICD), maintained by the World Health Organisation (WHO), is an international standard including causes of mortality. The ICD-10 GM is the Swiss national version of this vocabulary [13], and we used it as basis to set the system. Aiming to expand the coverage of the primary thesaurus, we upgraded it by adding new entities (new translations from the English ICD-10, see examples in Table 2). We also included new synonyms from the training dictionaries (from 2006 to 2013) with their ICD-10 codes. Finally, to avoid false positives potentially induced by short terms and acronyms, the expansion was limited to terms longer than three characters.

Table 2 : U84 and children translation granularity from ICD-10 GM.

ICD-10 code	ICD-10 WHO (English)	ICD-10 GM (French)	Translation proposed
U84	Resistance to other antimicrobial drugs	Virus de l'Herpès résistants aux virostatiques	Résistance aux autres antimicrobiens
U84.0	Resistance to antiparasitic drug(s)	-	Résistance aux médicaments antiparasitaires
U84.1	Resistance to antifungal drug(s)	-	Résistance aux médicaments antifongiques
U84.2	Resistance to antiviral drug(s)	-	Résistance aux médicaments antiviraux
U84.3	Resistance to tuberculostatic drug(s)	-	Résistance aux médicaments antituberculeux
U84.7	Resistance to multiple antimicrobial drugs	-	Résistance à de multiples médicaments antimicrobiens
U84.8	Resistance to other specified antimicrobial drug	-	Résistance à un autre antimicrobien précisé
U84.9	Resistance to unspecified antimicrobial drugs	-	Résistance à un antimicrobien non précisé

2.2.2. Pattern Matching

Our system uses pattern matching to test the different concepts, from the thesaurus, with each line in the input. First, this method prioritizes the exact match that fit the whole text, and then the longer entities.

3. Results and discussion

Performances of the systems are evaluated with the common metrics used in Natural Language Processing [14]. Precision represents the proportion of retrieved concepts that exactly match the gold benchmark prepared for these documents, while Recall represents the proportion of relevant concepts that were exactly extracted by the system. F-measure, also called harmonic mean, evaluates the accuracy of the system using both of the Precision and the Recall. Scores are calculated according to the following formulas.

$$Precision = \frac{true\ positives}{true\ \ positives + false\ positives}$$

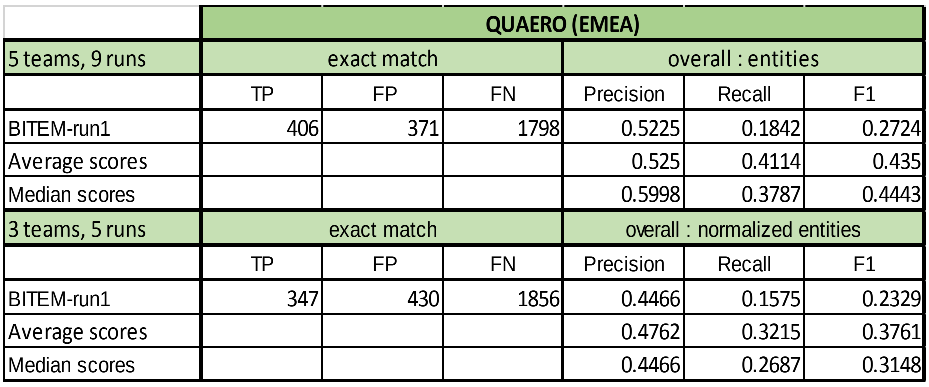
$$Recall = \frac{true\ positives}{true\ \ positives + false\ negatives}$$

$$F - measure\,(\beta) = (1 + \beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

Moreover, an exact match is attributed for the entity recognition when the entity type and span (starting position + ending position) correspond to the gold benchmark. Regarding the normalized entity recognition, the UMLS CUIs must also coincide with the reference benchmark. Inexact matches are credited when at least one word overlap from the prediction overlaps the span from the certificated benchmark.

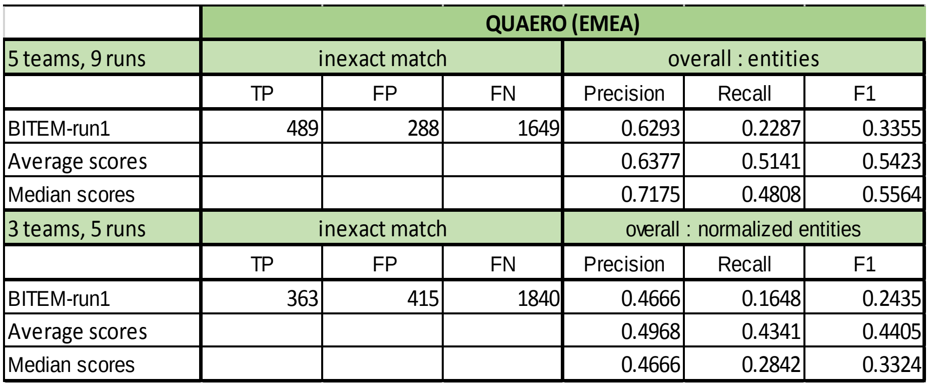
The results from the competitive phase disclosed in mid-May are reported in figures from 1 to 5. Our system provides substantially better results on MEDLINE than EMEA corpus, with F-scores of respectively 50% and 27% on the plain entity recognition. However, the recall may indicates that the basic French UMLS limits the coverage. This one is obviously not sufficient to extract all the concepts of interest, especially on the EMEA corpus that implies more drugs and pharmaceuticals.

On the other hand, to pre-process the ontology must have played a significant role to reach a F-score of 55% (precision 59% and recall 53%) by deploying an ad hoc solution for the CéPIDC coding task.



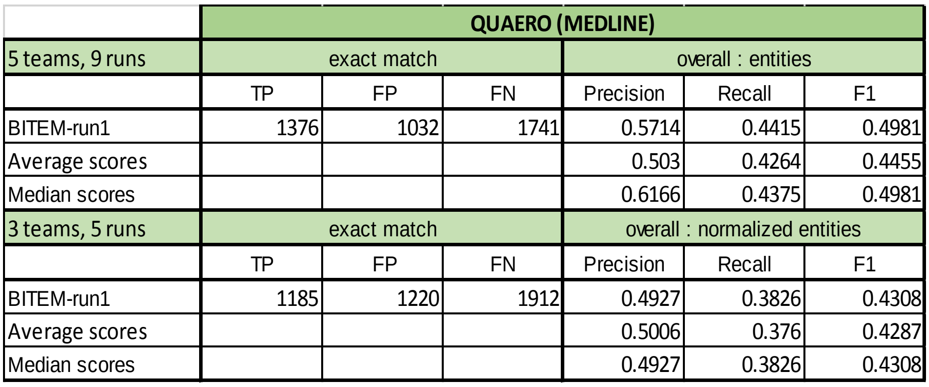
	QUAERO (EMEA)					
5 teams, 9 runs	exact match			overall : entities		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	406	371	1798	0.5225	0.1842	0.2724
Average scores				0.525	0.4114	0.435
Median scores				0.5998	0.3787	0.4443
3 teams, 5 runs	exact match			overall : normalized entities		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	347	430	1856	0.4466	0.1575	0.2329
Average scores				0.4762	0.3215	0.3761
Median scores				0.4466	0.2687	0.3148

Figure 1 : System results for the plain entity recognition and the normalized entity recognition tasks on the QUAREO/EMEA corpus, regarding the exact matches.



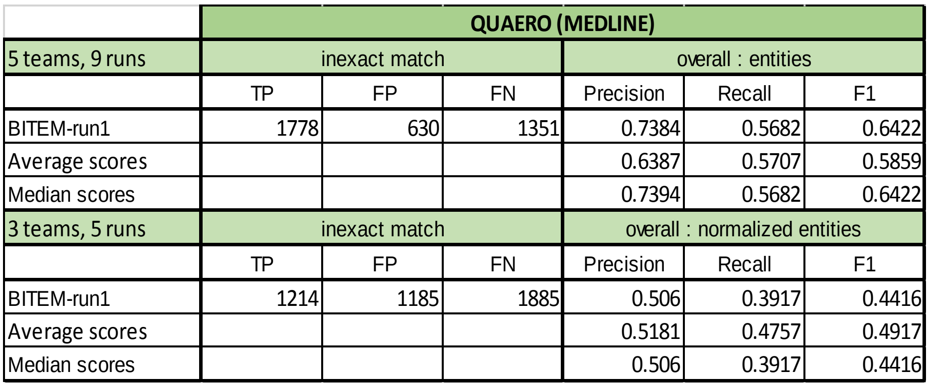
	QUAERO (EMEA)					
5 teams, 9 runs	inexact match			overall : entities		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	489	288	1649	0.6293	0.2287	0.3355
Average scores				0.6377	0.5141	0.5423
Median scores				0.7175	0.4808	0.5564
3 teams, 5 runs	inexact match			overall : normalized entities		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	363	415	1840	0.4666	0.1648	0.2435
Average scores				0.4968	0.4341	0.4405
Median scores				0.4666	0.2842	0.3324

Figure 2 : System results for the plain entity recognition and the normalized entity recognition tasks on the QUAREO/EMEA corpus, regarding the inexact matches.



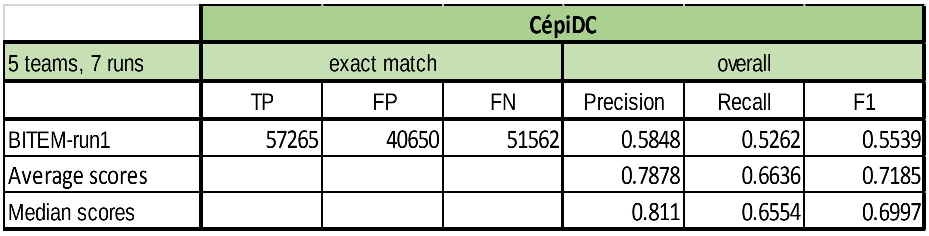
	QUAERO (MEDLINE)					
5 teams, 9 runs	exact match			overall : entities		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	1376	1032	1741	0.5714	0.4415	0.4981
Average scores				0.503	0.4264	0.4455
Median scores				0.6166	0.4375	0.4981
3 teams, 5 runs	exact match			overall : normalized entities		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	1185	1220	1912	0.4927	0.3826	0.4308
Average scores				0.5006	0.376	0.4287
Median scores				0.4927	0.3826	0.4308

Figure 3 : System results for the plain entity recognition and the normalized entity recognition tasks on the QUAREO/MEDLINE corpus, regarding the exact matches.



	QUAERO (MEDLINE)					
5 teams, 9 runs	inexact match			overall : entities		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	1778	630	1351	0.7384	0.5682	0.6422
Average scores				0.6387	0.5707	0.5859
Median scores				0.7394	0.5682	0.6422
3 teams, 5 runs	inexact match			overall : normalized entities		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	1214	1185	1885	0.506	0.3917	0.4416
Average scores				0.5181	0.4757	0.4917
Median scores				0.506	0.3917	0.4416

Figure 4 : System results for the plain entity recognition and the normalized entity recognition tasks on the QUAREO/MEDLINE corpus, regarding the inexact matches.



	CépiDC					
5 teams, 7 runs	exact match			overall		
	TP	FP	FN	Precision	Recall	F1
BITEM-run1	57265	40650	51562	0.5848	0.5262	0.5539
Average scores				0.7878	0.6636	0.7185
Median scores				0.811	0.6554	0.6997

Figure 5 : System results for the coding task on the CépiDC corpus.

4. Conclusion

Our results in the QUAREO subtask could certainly be improved by working with the English version of the UMLS, which covers much more terminology (128 sources), such as the NCI thesaurus or dictionaries specific to the drugs. Text sample would be translated in English using APIs (such a method has been proposed in past CLEF eHealth workshops), and the resulting coverage improvement could be significant. Another way to improve our system on QUAREO might have been to exploit the training datasets to exercise the Categorizer.

Regarding the CepiDC corpus, ATC did not achieved good results (e.g. forgetting many exact matches) due to an issue at data pre-processing stages. Our ad hoc pattern matching method brought relatively good results for the precision as well as the recall, but it would be interesting to prepare a subsequent run using the Categorizer.

5. References

1- Braschler M. (2000) CLEF 2000 — Overview of Results. In Cross-Language Information Retrieval and Evaluation, Springer Berlin Heidelberg, 2069, 89-101.

2- Goeuriot L., Kelly L., Suominen H., et al. (2015) Overview of the CLEF eHealth Evaluation Lab 2015. In Information Access Evaluation. Multilinguality, Multimodality, and Interaction, Springer International Publishing, 9283, 429-443.

3- Huang C.C., Lu Z. (2015) Community challenges in biomedical text mining over 10 years: success, failure and the future. In Brief Bioinform, 17, 132-144.

4- Overview of the CLEF eHealth Evaluation Lab 2016. Upcoming publication.

5- CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction. In CLEF 2016 Working Notes. CEUR-WS, Vol-1609.

6- Néveol, A., Grouin, C., Leixa, J.,et al. (2014) The Quaero French medical corpus: A Resource for Medical Entity Recognition and Normalization. In Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, 24-30

7- Barr C.E., Komorowski H.J., Pattison-Gordon E., et al. (1988) Conceptual Modeling for the Unified Medical Language System. In Proceedings of the Annual Symposium on Computer Application in Medical Care, 1988, 148-151.

8- Humphreys B.L., Lindberg D.A.B., Schoolman H.M., et al. (1998) The Unified Medical Language System: an informatics research collaboration. In JAMIA, 5 (1), 1–11.

9- Tuttle M., Sherertz D., Erlbaum M., et al. (1989) Implementing Meta-1: The First Version of the UMLS Metathesaurus. In Proceedings of the Annual Symposium on Computer Application in Medical Care, 1989, 483-487.

10- McCray A.T., Burgun A., Bodenreider O. (2001) Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. In Studies in health technology and informatics, 84(0 1), 216-220.

11- Yan Chen Y., Gu H., Perl Y., et al. (2008) Structural group auditing of a UMLS semantic type’s extent. In Journal of Biomedical Informatics, 42(1), 41-52.

12- Ruch P. (2006) Automatic assignment of biomedical categories: toward a generic approach. In Bioinformatics, 22 (6), 658-664.

13- Jetté N, Quan H., Hemmelgarn B., et al. (2010) The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. Medical Care, 48(12), 1105-1110.

14- Manning,C.D. and Schütze,H. (1999) Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, 268-269.