Efficient Reduction of Overgeneration Errors for Automatic Controlled Indexing with an Application to the Biomedical Domain

Samassi Adama[1], Brou Konan Marcellin[2], Gooré Bi Tra[3], Prosper Kimou[4]
Ecole Doctorale Polytechnique
Institut National Polytechnique (INP-HB)
Yamoussoukro, Côte d'Ivoire

***Abstract*—Studies on MetaMap and MaxMatcher has shown that both concept extraction systems suffer from overgeneration problems. Over-generation occurs when the extraction systems mistakenly select an irrelevant concept. One of the reasons for these errors is that these systems use the words to weight the terms of the concepts. In this paper, an Integer Linear Programming model is used to select the optimal subset of extracted concept mentions covering the largest number of important words in the document to be indexed. Then each concept mentions that this set is mapped to a unique concept in UMLS using an information retrieval model.**

***Keywords*—Concept extraction; concept recognition; automatic controlled indexing; controlled vocabulary; information retrieval**

I. Introduction

With the fast evolution of scientific publications in the biomedical field, the availability of these resources on the Internet requires automatic indexing methods. For this, specialized search engines such as PubMed in the USA and CisMEF in France have been developed. They use the concepts of the MeSH (Medical Subject Headings) thesaurus to index their document resources. This allows users (health professionals, students, patients) to retrieve relevant documents.

Indexing is the representation of a document by terms (keywords, sentences or concepts). It can be done manually or automatically. Manual indexing is efficient but it is costly in terms of time and human resources. On the other hand, automatic indexing is less efficient, but it is faster than manual indexing and requires less time and human resources.

Several automatic indexing approaches have been proposed. They can be classified into three categories: approaches based on the free extraction of terms, approaches based on controlled vocabularies (ontology, thesaurus, dictionary) and hybrid approaches that combine these two approaches. The first category represents the document by the keywords it contains without using controlled vocabulary. The second category such as MaxMatcher [3] and MetaMap [4] uses the terms of a controlled vocabulary to index the document. Approaches in this category can be subdivided into four sub-categories: (1) language-based approach, (2) machine learning approach, (3) statistic-based approach and (4) approach-based on searching in a dictionary. Language rule-based approaches define particular rules for describing terms that designate concepts. These rules are set manually by experts and depend on the characteristics of the language used.

Approaches based on machine learning use corpora manually annotated to train classifiers who consider several characteristics of textual instances to associate technical terms with predefined classes. However, these approaches are dependent on the availability of training data.

Most approaches based on statistical measures combine statistical information such as the frequency of the terms (TF), the inverse of the document frequency (IDF).

Dictionary-based approaches use terminological resources to compare textual instances with the entries (terms) of concepts in the dictionary. This search is based on the exact or partial matching between the textual fragments of the document and the entries (terms designating concepts) of the dictionary. However, exact matching leads to sub-generation problems. The partial matching leads to over-generation problems. Over-generation occurs when the system mistakenly considers a concept to be relevant because it contains a word with a high weight. This is the result of the use of simple words in the rough comparison between dictionary concepts and noun groups identified in the text as candidate concepts. The sub-generation is linked to ignorance of relevant concepts. This is the result of the strict comparison between the nominal groups of the text and the concepts of the dictionary.

In this article, the authors focus on dictionary-based approaches. The goal is to reduce the over-generation errors of concepts extraction systems based on search in a dictionary. The proposed approach is based on recent methods [1,2] of solving over-generation errors.

According to Boudin et al. [2], one of the reasons for over-generation errors is that in extraction systems, candidate concepts are selected based on the weight of their constituent words as in Zhou et al. [3]. As a result, an irrelevant concept containing a significant word can be selected. The selection of the candidate terms according to their constitutive words makes it possible to reduce the over-generation errors provided that the weight of each word is calculated only once in the set of extracted terms [2]. Thus, Boudin et al. [2] proposed an integer linear programming model for extracting

key terms. This model reduces over-generation errors by weighting candidate terms as a set rather than independently. In this model, key terms are selected based on their constituent words and the weight of each unique word. The main contribution of this paper is the proposal of a method for reducing the over-generation errors of extraction systems based on the search in a dictionary. It is summarized in these points:

- Identification of some problems related to the extraction of concepts based on the search in a dictionary.
- Identification of methods of literature that can provide solutions to these problems.
- Proposal of a method for reducing the over-generation errors in an extraction system.

The rest of the article is organized as follows. In Section 2 the authors present MetaMap and MaxMatcher, two state-of-the-art concept retrieval systems based on research in a controlled vocabulary. Next, they present two methods of solving the over-generation errors identified in the results of the extraction of the two previous systems. In Section 3 they describe their approach for reducing over-generation errors of search-based retrieval approaches in a dictionary. Section 4 presents the discussion. They conclude and present some ideas for future work in Section 5.

II. Related Work

In this section, the authors first introduce MetaMap and MaxMatcher, two state-of-the-art retrieval approaches based on search in a dictionary of terminology concepts. Next, they present two methods of solving the over-generation errors identified in the results of the extraction of the two previous systems.

A. MetaMap

MetaMap [4] is a tool for extracting the concepts of UMLS from biomedical documents. The MetaMap extraction process consists of the following five main steps:

- Identification of nominal groups in the text using an analysis grammatical.
- Generation of variants (synonyms, acronyms, ...) for each group nominal using the SPECIALIST Lexicon resource of the UMLS,
- Selection of candidate concepts: a concept with at least one word found in one of the variants is retrieved (this leads to over-generation and sub-generation problems),
- Concept evaluation: The candidate concepts are compared with the original text using the following four measures: centrality, variation, coverage and consistency. The candidate concepts are finally ordered according to the final score.
- Correspondence construction: for each document, the concepts are assigned according to their similarity score with it.

Among the disadvantages of MetaMap we have the over-generation problems, the under-generation issues and the data processing time.

1) Over-generation errors: Over-generation occurs when the system mistakenly selects an irrelevant concept. This is the result of the use of simple words in the comparison between dictionary concepts and noun groups identified in the text as candidate concepts. For example, for the nominal group "ocular complications", MetaMap selects the three concepts "Ocular", "Complications" and "Complications Specific to Antepartum or Postpartum" because they share at least one word.

2) Sub-generation errors: The sub-generation is linked to the non-selection of relevant concepts. This is the result of the strict comparison between each nominal group of the text and the concepts of the dictionary. For example, for the expression "gyrb and p53 protein", MetaMap can't identify the word "gyrb" as a protein because it is registered in the UMLS as "gyrb protein".

3) Data processing time: Another disadvantage of MetaMap is its data processing time. Indeed, this tool uses a set of sophisticated linguistic methods such as grammatical analysis, the generation of variants, the search in the whole of the Metathesaurus, as well as the calculation of several statistical measures.

B. MaxMatcher

Zhou et al. [3] proposed MaxMatcher, a generic extraction approach based on the approximate search for strings in a dictionary of terms designating concepts. The basic idea of this approach is to index documents with only the most significant words of the UMLS meta-thesaurus concepts.

1) Concept Recognition: For a document, MaxMatcher cuts it into sentences and then identifies biological concept names (terms). For a given text, a set of rules to identify the boundary of a biological concept name. A biological concept term should begin with a noun, a number, or an adjective while ending with a noun or a number. It can not contain any boundary words including: punctuations (except hyphen, period, and single quote), verbs, and conjunctions and prepositions (except “of”). Whenever a boundary word is encountered, a candidate concept term reaches its end and it is then extracted.

2) Concept Normalization: The task of mapping a biological term to a concept in a controlled vocabulary, typically to the standard thesaurus in the Unified Medical Language System (UMLS), is known as medical concept normalization.

After the concept recognition step, MaxMatcher identifies the extracted terms that correspond to concept entries (terms) in a dictionary of biological concept terms.

Let $t = \{w_1, w_2, ..., w_m\}$ be a candidate term (extracted from the text) consisting of a set of simple words, $N(w)$ the number of concepts whose variant names contain word w, w_{ji} the i-th word in the j-th variant name of the concept. The

similarity between each of its words w_i and a concept c of UMLS, denoted by a set of n variant names (terms) $\{v_1, v_2, ..., v_n\}$, is defined in [3] as follows :

$$I(w_i, c) = \max\{I(w_i, v_j) | j \leq n\} \quad (1)$$

where :

$$I(w, v_j) = \begin{cases} \sum_i \frac{1/N(w)}{1/N(w_{ji})} & \text{if } w \in v_j \\ 0 & \text{else} \end{cases} \quad (2)$$

According to equation (1), candidate concepts are selected based on the weight of their constituent words. As a result, an irrelevant concept containing a significant word can be selected. This is an over-generation error. It is the result of partial matching method used by MaxMatcher. However, according to Boudin et al. [2], the selection of the candidate terms according to their constitutive words makes it possible to reduce the over-generation errors provided that the weight of each word is calculated only once in the set of extracted terms.

C. Automatic Keyphrase Extraction Approaches

Keyword extraction is the task of automatically identifying a set of terms that best describe a text document [10]. Automatic keyword extraction has been found to be useful for many natural language processing applications such as information retrieval, automatic indexing and classification of text documents, automatic summarization [11,12]. However, state-of-the-art keyword extraction systems suffer from over-generation errors.

According to **Boudin et al**. [2], the selection of key terms according to their constituent words makes it possible to reduce over-generation errors, provided that the weight of each word is calculated only once in the set of these terms. The key-term extraction model they propose has three steps: (1) Extraction of candidate terms using heuristic rules (2) weighting of words using supervised or unsupervised methods (3) optimal subset of key terms by integer linear programming.

Jia et al. [1] proposed an unsupervised method of extracting key terms. According to these authors, unsupervised methods for extracting existing key-words suffer from over-generation error because they generally identify keywords and then return as keywords the terms of the text containing these keywords. In other words, key word extraction systems first assign scores to the words, then rank the candidate key terms based on the sum of the weights of their constituent words. To overcome this problem, Jia et al. proposed a weighting scheme that is applied directly to candidate key terms by exploiting some of their properties such as informativeness and positioning preference.

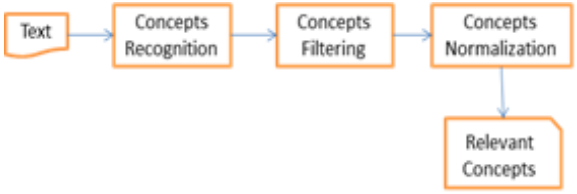

Fig. 1. The Proposed Approach.

III. Proposed Approach

The concept extraction approach proposed in this section is based on three steps (Figure 1): (1) Concepts Recognition; (2) Concepts Filtering; (3) Concept Normalization.

A. Concepts Recognition

Concepts recognition consists in the identification of the mentions (terms) of biological concepts in a textual document. For a given text, we used a set of rules to identify the boundary of a biological concept term (concept name) as in [3]. A biological concept name should begin with a noun, a number, or an adjective while ending with a noun or a number. It can not contain any boundary words including: punctuations (except hyphen, period, and single quote), verbs, and conjunctions and prepositions (except "of"). In other words, whenever a boundary word is encountered, a candidate concept mention reaches its end.

Let $V = \{T_1, T_2, ..., T_P\}$ be the set of recognized concept mentions extracted from the document. All these concepts are not relevant with respect to this document. We must select the optimal subset of these mentions covering the largest number of important words in the document.

B. Concepts Filtering

Filtering concepts consists in selecting of the optimal subset of concept mentions covering the largest number of important words in a document. The authors used the model of Boudin et al. [2] to find the optimal subset of concept mentions. The model is defined as

$$\max \sum_i \omega_i x_i - \lambda \sum_j \frac{(l_j - 1)c_j}{1 + \text{substr}_j} \quad (3)$$

$$\text{s.t} \sum_j c_j \leq N \quad (4)$$

$$c_j Occ_{ij} \leq x_i, \quad \forall i, j \quad (5)$$

$$\sum_j c_j Occ_{ij} \geq x_i, \quad \forall i \quad (6)$$

$$x_i \in \{0,1\} \quad \forall I \qquad c_j \in \{0,1\} \quad \forall j$$

where w_i (computed using Equations (7,8) is the weight of a word i, x_i and c_j are two binary variables indicating the presence of word i and candidate concept j in the set of extracted concepts, l_j is the size of concept j, $substr_j$ is the number of times concept j appears as a substring in the other concepts, Occ_{ij} is an indicator of the occurrence of word i in concept j and N the number of candidate concepts.

1) Word weighting Functions: The performance of the model of Boudin et al. [2] depends on how word weight w_i is estimated. It can be computed using one of the following unsupervised weighting functions: BM25 [7] and TFxIDF [6].

TF.IDF

$$TF \times IDF(t, d) = tf(t, d) \times \log(\frac{N}{n}) \quad (7)$$

Where tf(t, d) is the frequency of the word t in a document d, N is the number of documents in the corpus, n is the number of document containing t.

BM25

$$BM25(t,d) = tf(t,d)\frac{\log(\frac{N-n+0.5}{n+0.5})}{tf(t,d)+k_1((1-b)+b\frac{l_d}{avgl_d})} \quad (8)$$

where tf(t, d) is the frequency of the word t in a document d, N is the number of documents in the corpus, n the number of documents containing the word t, l_d the length of a document d, n the number of documents containing the word t, $avgl_d$ the average document length (number of words in the document) ; k_1 and b are free parameters.

Once the optimal set of concept mentions is found, each of them needs to be normalized, if possible, with a unique identifier (CUI) from the Unified Medical Language System (UMLS) metathesaurus.

C. Concepts Normalization

The task of mapping a concept mention in a text to a semantically equivalent concept in a biological knowledge base (like UMLS) is known as concept normalization. In this study, each concept mention is mapped to a Concept in the UMLS metathesaurus. This way, a semantic meaning is associated to each of them. In the UMLS each concept is given a Concept Unified Identifier (CUI). Each synonym and abbreviation of this concept is called Term. A term is either Preferred Term (PT) or Synonym (SY) (figure 2).

Concept normalization is challenging because: (1) the same word or term can be used to refer to different concepts, and (2) the same concept can be referred to by different words or terms, (3) the different expressions (terms) of a concept are not necessary all present in the knowledge base.

Let $T = \{T_1, T_2, ..., T_N\}$ be the optimal subset of the set V(the set of recognized concept mentions) containing the N concept mentions extracted from the document. In the proposed concept normalization method, each concept mention T_i is treated as a query, while the concepts in the UMLS are treated as documents that are searched to find the relevant concepts. So, all the concepts in the UMLS metathesaurus are indexed.

Formally, a concept mention is modeled as a sequence T_i of one or more words $\{t_1, t_2, ..., t_n\}$. A concept in UMLS is modeled as a concept-document C_j, which is a sequence of one or more words $\{t_1, t_2, ..., t_n\}$.

CUI = C2612523
PT = urate biosynthetic process
SY = urate formation
SY = urate synthesis
SY = uric acid biosynthetic process
SY = urate biosynthesis
SY = urate anabolism

Fig. 2. The Concept Urate Biosynthetic Process Identified by the Unique Identifer C2612523 in UMLS Metathesaurus.

The authors cast the concept normalization task in an information retrieval problem as in [5,8]. All the concepts in the UMLS metathesaurus are indexed. Thus, given a concept mention T_i (a query), retrieve the relevant concept-documents (concepts) $C_1, C_2, ..., C_k$ from this index. Standard Information Retrieval (IR) models can be used on the concept mapping problem. In this paper we used the BM25 to rank the concept-documents for a given concept mention T. Thus, the score of a concept-document (a concept in UMLS) for a concept mention T is defined as

$$BM25(T,C) = \sum_{t\in(T\cap C)} BM25(t,C) \quad (9)$$

with :

$$BM25(t,C) = tf(t,C)\frac{\log(\frac{N-n+0.5}{n+0.5})}{tf(t,C)+k_1((1-b)+b\frac{l_C}{avgl_C})} \quad (10)$$

where tf(t, C) is the frequency of the word t in a concept C, N is the number of concepts in the UMLS metathesaurus, n the number of concepts containing the word t, l_C the length of a concept C, n the number of concepts containing the word t, $avgl_C$ the average concept length (number of words in the concept) ; k_1 and b are free parameters.

Finally, each concept mention T is mapped to the concept C which has the maximum score.

IV. Discussion

Dictionary-based concept extraction is the state-of-the-art approach to biomedical literature indexing. In this work, the authors are interested in reducing the over-generation errors of MaxMatcher [3], which is a concept extraction system based on search in a dictionary. One reason these errors is that this system ranks extracted concepts according to the weights of their component words (equation 1). This approach poses a major problem: if a word is very important then all the terms that contain it will also be considered important. As a result, irrelevant concepts containing a significant word are selected. To reduce the number of these irrelevant concepts, the idea is to search for the optimal subset of extracted concepts covering the largest number of important words in the document. According to Boudin et al. [2], finding this optimal set of concepts is a combinatorial optimization problem, and can be formulated as an integer linear program. However, this approach [2] is supervised. Thus it requires large amounts of labeled training data. At the same time, unsupervised systems like [1] have poor accuracy and do not generalize well.

Jia et al. [1] proposed to directly weight candidate key terms by considering some of their properties such as informativeness and positioning preference. Such approach can be used to reduce over-generation errors. Since ambiguous concept mention can be mapped to multiple concepts in the referenced ontology (UMLS) depending on the context, one of the main challenges in the concept normalization task consists in the disambiguation of these cases [9].

V. CONCLUSION

In this paper, authors analyzed some works on MetaMap and MaxMatcher, two concept extraction systems based on the search for strings in a dictionary of terms designating concepts. They found that both systems suffer from over-generation errors. So they proposed to use an Integer Linear Programming model to select the optimal subset of extracted concepts that are relevant to the document to index. Then each concept mention of this set is mapped to a unique concept CUI in the UMLS metathesaurus. The authors cast the mapping task in an information retrieval problem, using a concept mention as query and the concepts in ULMS as documents.

In future work, we plan to test our method using the OHSUMED collection. Since ambiguous concept mention can be mapped to multiple concepts in the referenced ontology (UMLS) depending on the context, one of the main challenges in the concept normalization task consists in the disambiguation of these cases.

REFERENCES

[1] Jia, H., & Saule, E. (2018). Addressing Overgeneration Error: An Effective and Effcient Approach to Keyphrase Extraction from Scientific Papers. In BIRNDL@ SIGIR (pp. 60-73).

[2] Boudin, F. (2015, July). Reducing over-generation errors for automatic keyphrase extraction using integer linear programming. In ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction.

[3] Zhou, X., Zhang, X., & Hu X. (2006). MaxMatcher: Biological concept extraction using approximate dictionary lookup. PRICAI 2006: trends in artificial intelligence, 1145-1149.

[4] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association.

[5] Mirhosseini, S., Zuccon, G., Koopman, B., Nguyen, A., & Lawley, M. (2014, November). Medical free-text to concept mapping as an information retrieval problem. In Proceedings of the 2014 Australasian Document Computing Symposium (p. 93). ACM.

[6] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28:11–21.

[7] Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. In TREC-7, pages 199–210.

[8] Ruch, P. (2005). Automatic assignment of biomedical categories: toward a generic approach.

[9] Leal, A., Martins, B., & Couto, F. (2015). ULisboa: Recognition and normalization of medical concepts. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 406-411).

[10] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

[11] Balaji, J., Geetha, T. V., & Parthasarathi, R. (2016). Abstractive summarization: A hybrid approach for the compression of semantic graphs. International Journal on Semantic Web and Information Systems (IJSWIS), 12(2), 76-99.

[12] Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1262-1273).