Detecting Noun Phrases in Biomedical Terminologies: the first step in managing the evolution of knowledge

Adila Merabti[1], Lina F. Soualmia[1,2], and Stéfan J. Darmoni[1,2]

[1] CISMeF, TIBS LITIS laboratory EA 4108, Rouen University Hospital, France
[2] LIMICS, French National Institute for Health, INSERM UMR 1142, Paris, France
(adila.merabti, lina.soualmia, stefan.darmoni)@chu-rouen.fr

Abstract. In order to identify variations between two or several versions of Clinical Practice Guidelines, we propose a method based on the detection of noun phrases. Currently, we are developing a comparison approach to extract similar and different elements between medical documents in French in order to identify any significant changes such as new medical terms or concepts, new treatments etc. In this paper, we describe a basic initial step for this comparison approach i.e. detecting noun phrases. This step is based on patterns constructed from six main medical terminologies used in document indexing. The patterns are constructed by using a Tree Tagger. To avoid a great number of generated patterns, the most relevant ones are selected by choosing those that identify more than 80% of the six terminologies used in this study. These steps allowed us to obtain a manageable list of 262 patterns which have been evaluated. Using this list of patterns, 708 maximal noun phrases were found, among them, 364 are correct which represent a 51.41% precision. However by detecting these phrases manually, 602 maximal noun phrases were found which represent a 60.47% recall and by consequence a 55.57% F-measure. We tried to improve these results by increasing a number of patterns from 262 to 493. We obtained a total of 729 maximal noun phrases, with 365 which were correct, which corresponding to a 50.07% precision, 60.63% recall and 54.85% F-measure.

Keywords: Medical knowledge evolution, biomedical terminologies, Natural Language Processing, Clinical Practice Guidelines, noun phrases detection, patterns.

1 Introduction

This study was developed in the context of medical knowledge evolution, specifically in the context of document evolution such as Clinical Practice Guidelines (CPGs) and Electronics Health Records. It is important for the physician to have a clear view of this evolution. As an initial application, we chose to focus on Clinical Practice Guidelines (CPGs), which are medical documents that provide recommendations for the diagnosis and treatment of numerous diseases.

These guidelines are a fundamental reference source physicians [1]. Based on this option, many studies have been proposed to elucidate this problem. However, the methods used remain rather specific to the documents they treat and it is practically impossible to apply the same tools on other types of medical documents. Therefore, we were prompted to propose the implementation of a practical ergonomic tool which is flexible and can manage this evolution in any type of French medical document. This tool will serve as a basis for the evaluation of the proposed approach by comparing our results with those of the other existing methods [2]. This approach involves several steps. The basic step consist of detecting noun phrases by extracting patterns from different medical terminologies. It is the purpose of this study. The noun phrases are groups of words constructed around a single noun, which is called the headword of the phrase. We selected there specific types of phrases because they represent the general structure of medical terminology which will serve as the basis of our study. Once the various noun phrases identified the next step will be a comparison to detect any minor or major changes. For example in the case of chronic diseases that constantly evolve, the physician can use our tool and immediately learn of new practices without necessary having to search the new CPG. The paper is structured as follow: first we describe the material used, in the section 3 we describe the methods. The step of detecting noun phrases are detailed in section 4, and the evaluation is presented in section 5. In section 6 we presente the results of this study. Finally, we give conclusions in the section 7.

2 Material

To test our approach, we chose the Clinical Practice Guidelines (CPGs) as the basic medical documents to be compared. Medical terminologies are then used to build patterns such as : MeSH (Medical Subject Heading), ICD10(International Classification of Diseases)... and an annotating tool Tree Tagger [3]. We describe hereafter the structure of each element.

2.1 Clinical Practice Guidelines:

As previously mentioned CPGs contain recommendations for the diagnosis and treatment of numerous diseases. The American Institute of Medicine defines clinical practice guidelines as systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances [4].

2.2 Biomedical terminologies

A terminological system links together concepts of a domain and gives their associated terms, as well as their definition and code. It might also provide the designation of terminology, thesaurus, controlled vocabulary, nomenclature, classification, taxonomy or ontology. In [5], a terminology is defined as a set of words.

However a more precise definition of terminology has been proposed [6]: Terminologies are composed by a list of terms of one domain or a topic representing concepts or notions most frequently used or most characteristic. Thereby, the content and the structure of a terminology depends on its specific function. In a thesaurus the terms are for example organized alphabetically and the concepts may be designed with one or several synonyms. When the terms are associated to definitions, it constitutes a controlled vocabulary.
The main biomedical terminology references are as follow: MeSH, SNOMED Int, MedDRA, ICD10, ATC and FMA. All are annoteted with Tree Tagger [3].

MeSH. The Medical Subject Headings (MeSH) [7] is a biomedical thesaurus, created and updated by the US National Library of Medicine (NLM). It is used for indexing the bibliographic references of MEDLINE/PubMed. Originally in English, the MeSH has been translated into numerous other languages, such as French. It contains $545,082$ concepts (eg: *embryotomy*).

SNOMED Int. The Systematized Nomenclature Of MEDicine International (SNOMED Int) is used essentially to describe electronic health records [8]. It contains $208,769$ terms (eg: *Parkinson's disease*).

MedDRA. The Medical Dictionary for Regulatory Activities (MedDRA) [9] has been designed for the encoding of adverse drug reactions chemically induced. It contains a large set of terms (signs and symptoms, diagnostics, therapeutic indications, complementary investigations, medical and surgical procedures, medical, surgical, family and social history). It contains $45,663$ concepts (eg: *Marfan's syndrome*, *Asthma*).

ICD10. The International Classification of Diseases (ICD) is the standard diagnostic tool for epidemiology, health management and clinical purposes. This includes the analysis of the general health situation of population groups. It is used fior classifying diseases and other health problems recorded on many types of health and vital records including death certificates and health records. It contain $44,962$ terms(eg: *Turner syndrome*).

ATC. The Anatomical, Therapeutic and Chemical classification (ATC) is an international classification [10] used to classify drugs. The ATC classification is developed and maintained by the Collaborating Centre for Drug Statistics Methodology. In the ATC classification system, the drugs are divided into different groups according to the organ or the system on which they act and their chemical, pharmacological and therapeutic properties. It contains $11,105$ terms(eg : *Acebutolol* and *thiazides*).

FMA. The Foundational Model Anatomy (FMA)is an evolving formal ontology that has been under development at the University of Washington since 1994 [11]. It is the most complete ontology of human "canonical" anatomy. The FMA describes anatomical entities, most of which are anatomical structures composed

of many interconnected parts in a complex way. It contains more than 81,000 classes and 139 relationships connecting the classes, and over 120,000 terms(eg : *Arm*).

The knowledge terminological ressources MeSH, SNOMED INT, MedDRA, ICD10 and ATC exist in French and in English. FMA terms are currently being translated by CISMeF team(Catalogue et Index des Sites Médicaux de langue Française) [12].

2.3 Tree Tagger

Tree Tagger [3] is a tool for annotating text with part-of-speech (POS) and lemma information. It was developed by Helmut Schmid at the Institute for Computational Linguistics of the University of Stuttgart. The Tree Tagger has been successfully used to tag several languages (German, English, French, etc) and is adaptable to other languages if a lexicon and a manually tagged training corpus are available. It takes as input a text and as output it gives the POS tag (noun, adjective, verb, etc). For example in the Figure 1, the sentence ***Carbonate de sodium dihydroxyaluminium*** is tagged by carbonate=NOM, de=PRP, etc.

Input: carbonate de sodium dihydroxyaluminium.
Output:

word	pos	lemma
carbonate	NOM	carbonate
de	PRP	de
sodium	NOM	sodium
dihydroxyaluminium	NOM	dihydroxyaluminium

NOM = Noun; ADJ = Adjective; VER = Verb

Fig. 1. Example of tagging with a Tree Tagger

3 Methods

Among the works on the Clinical Practice Guidelines, there is Brigitte Seroussi's studies [2], which are based on the formalization of the CPGs in the form of decision tree structure, by comparing the basic CPG which are represented as rules production, with clinical situations and action plans. In our study we propose a generic method able to manage this evolution on all types of medical documents by ignoring the specificity of the processed document. Our approach includes several steps which are detailed in the Figure 2.

– Step 1: The selection of the CPGs on the same pathology and edited by the same organization with five years difference published for example.
– Step 2: The segmentation of the input text CPGs into sentences.
– Step 3: The correspondence of the sentences of both CPGs which will allow us to obtain the most similar sentences in output. In this step, we used similarity measures (Dice [13], Levenshtein [14] and Stoilos [15]) which calculated the similarity between all the characters that compose sentences of both CPGs.
– Step 4: The extraction of the maximal noun phrases by using a pattern constructed approach on medical terminologies tagged with Tree Tagger.
– Step 5: The comparison of noun phrases extracted from both CPGs based on the context of each phrase (right and left elements) to extract all the possible insertions, deletions and substitutions.

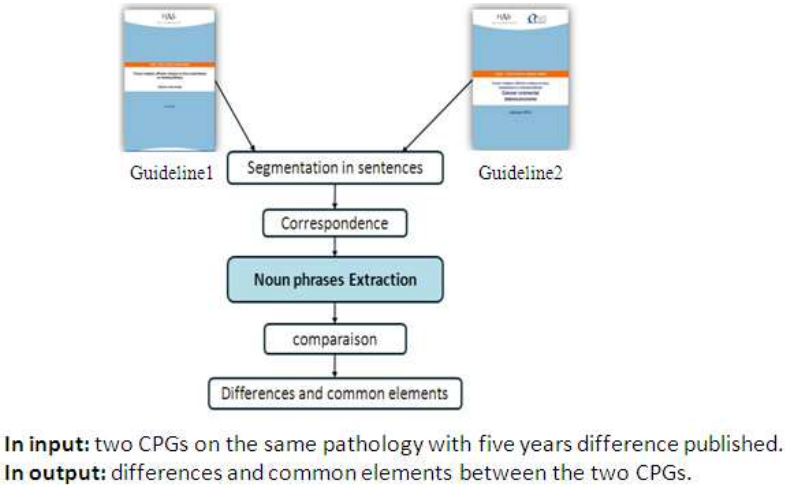

Fig. 2. Method plan

The following are the equations of the similarity measures used:

1. **Dice's coefficient:** This measure is used in statistics to determine the similarity between two samples X and Y. It is between 0 and 1. In our case, we calculate the coefficient between two sentences. For this, we defined two samples X and Y as the set of bigrams of each respective sentence x and y. A bigram is the union of two letters. The Dice's coefficient is defined by the equation (1).

$$Dice\,(X,Y) = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{1}$$

2. **Distance of Levenshtein:** This measure between two sentences x and y is defined as the minimum number of elementary operations that is required

to pass from the sentence x to the sentence y. There are three possible transactions: replacing a character with another (*ast**h**ma, ast**m**ma*), deleting a character (*ast**h**ma, astma*) and adding a character (*asthma, asth**m**ma*). This measure takes its values in the interval [0,∞[. The Normalized Levenshtein [16] ($LevNorm$) in the range $[0, 1]$ is obtained by dividing the distance of Levenshtein $Lev(x, y)$ by the size of the longest string and it is defined by the equation (2).

$$LevNorm(x, y) = 1 - \frac{Lev(x, y)}{Max(|x|, |y|)} \tag{2}$$

LevNorm $(x, y) \in [0, 1]$ because $Lev(x, y) \leq Max(|x|, |y|)$; with $|x|$ the length of the sentence x.

3. **Stoilos similarity function :** is based on the idea that the similarity between two entities is related to their commonalities as well as their differences. Thus, the similarity should be a function of both these features. It is defined by the equation (3).

$$Sim(x, y) = comm(x, y) - Diff(x, y) + Winkler(x, y) \tag{3}$$

Comm(x,y) stands for the commonality between the strings x and y, Diff(x,y) for the difference between x and y, and Winkler(x,y) for the improvement of the result using the method introduced by Winkler in [17]. The function of commonality is determined by the substring function. The biggest common substring between two sentences (MaxComSubString) is computed. This process is further extended by removing the common substring and by searching again for the next biggest substring until none can be identified. The function of commonality is given by the equation (4):

$$Comm(x, y) = \frac{2 \times \sum_i |MaxComSubString|}{|x| + |y|} \tag{4}$$

The function of Difference is defined in the equation (5) where $p \in [0, \infty[$ (usually p= 0.6), $|u_x|$ and $|u_y|$ represent the length of the unmatched substring from the strings sentences and y scaled respectively by their length:

$$Diff(x, y) = \frac{|u_x| \times |u_y|}{p + (1 - p) \times (|u_x| + |u_y| - |u_x| \times |u_y|)} \tag{5}$$

The Winkler parameter Winkler(x,y) is defined by the equation (6):

$$Winkler(x, y) = L \times P \times (1 - Comm(x, y)) \tag{6}$$

Where L is the length of common prefix between the sentences x and y at the start of the sentence up to a maximum of 4 characters and P is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. The standard value for this constant in Winkler's work is P=0.1 [17].

Detection and extraction of noun phrases using patterns based on different medical terminologies and Tree Tagger is the first step in the procedure. Once the detection is accomplished, we procede to the comparison step in an attempt to detect all possible changes. Many studies have been dedicated to the extraction of noun phrases. For example the ACABIT tool [18] is a program that allows a terminological acquisition on a pre-tagged and disambiguated corpus. This acquisition is performed in two steps: 1) a linguistic analysis of the corpus by transducers that produces candidate terms and 2) a statistical filtering step that sorts candidate terms from a reference corpus and valids terms. Another extractor YaTeA (Yet another Term Extractor) [19], is used to assist the process of identifying terms in a French and English, with visualization and configuration interfaces to reduce the complexity of the writing and editing configuration files.

4 Noun phrases detection

4.1 Extraction patterns of noun phrases

For the construction patterns of extraction step for noun phrases, we tagged the six biomedical terminologies detailed in section 2, with Tree Tagger. For example: the term ***douleur abdominale*** (***abdominal pain***) is tagged as follows:
douleur **NOM** douleur — abdominale **ADJ** abdomen.
All the corresponding patterns were automatically generated. For example: the corresponding pattern of ***douleur abdominale*** (***abdominal pain***), is ***NOM ADJ*** (***NOUN ADJECTIVE***). The duplicates were removed and to reduce the total number of patterns, only some of them were selected. Do do that, we relied on two selection criteria: the length of terms and their relevance. For the first criterion, we realized a statistical study by calculating for every terminology the percentage of terms which length was less than or equal to 1, 2 to 16 words by term, and we found that over 95% of all the terms have a length less than or equal to 8 words. The details are reported in the Table 1.
For the second criterion, which is the relevance of the patterns, we calculated for each pattern, the percentage of words represented by the latter. For example, the pattern ***NOM*** (***NOUN***) represents 22.16% of the MeSH terminology and ***NOM ADJ*** (***NOUN ADJ***) represents 13.90%. Therefore 36.06% of the terms may be represented with only these two patterns. Thus we kept the patterns which represent more than 80% of the chosen terminologies. The final list is composed by 262 patterns. Table 2 shows some of the results for the MeSH thesaurus.

4.2 Detection and extraction of noun phrases

The patterns are applied to the CPGs previously labeled with a Tree Tagger to extract the corresponding noun phrases. Since the list of the patterns contains imbricated patterns (e.g.: ***NOM*** is included in ***NOM ADJ*** (***NOUN*** and ***NOUN ADJECTIVE***)) this implies imbricated noun phrases. (e.g.: the

Table 1. Percentage of length of terms, Nb: Number of terms of length lower or equal to 1, 2 to 9 and the percentage of these terms.

	1	2	3	4	5	6	7	8	9
ATC	$Nb = 2,994$	3,568	7,531	8,512	9,683	10,273	10,587	10,823	10,928
	26.97%	32.14%	67.83%	76.67%	87.22%	92.53%	95.36%	97.49%	98.43%
MeSH	$Nb = 33,846$	76,596	103,836	120,717	130,029	135,449	138,204	139,739	140,559
	23.94%	54.17%	73.43%	85.37%	91.96%	95.79%	97.74%	98.83%	99.41%
ICD10	$Nb = 2,661$	9,479	16,370	21,972	25,906	28,856	30,896	32,493	33,720
	5.92%	21.08%	36.41%	48.87%	57.62%	64.18%	68.72%	72.27%	75.00%
FMA	$Nb = 1,817$	7,280	11,495	14,071	15,872	16,872	17,278	17,555	17,675
	10.24%	41.01%	64.76%	79.27%	89.41%	95.05%	97.34%	98.90%	99.57%
SND	$Nb = 27,406$	85,926	123,797	152,553	172,464	185,258	193,508	199,098	202,645
	13.13%	41.16%	59.30%	73.07%	82.61%	88.74%	92.69%	95.37%	97.07%
MDR	$Nb = 4,583$	16,310	26,968	35,540	40,617	43,108	44,350	44,916	45,263
	10.04%	35.72%	59.06%	77.83%	88.95%	94.41%	97.13%	98.37%	99.13%

noun phrases *maladie* (*disease*) and *Alzheimer* (*Alzheimer*) are included in the noun phrase *maladie d'Alzheimer* (*Alzheimer's disease*)). To avoid redundancies in noun phrases, we extract them from the maximal noun phrases according to their positions in the sentence. The process of the algorithm we propose is detailed in the Figure 3.

The position (pos) and the length of each extracted noun phrase is calculated. Then we look for the imbricated noun phrases and look if they have the same position, (i) if it is the case, the smallest noun phrase is deleted. (ii) If not, we look if the position of the smallest noun phrase is equal to the position of the biggest noun phrase added to the position of the smallest in the biggest noun phrase; The smallest noun phrase is then deleted.
For example, by applying this algorithm to the sentence : *Le médecin traitant assure la coordination des soins et la surveillance du patient en ambulatoire en lien avec l'équipe spécialisée*, we obtained: *médecin traitant*, *coordination des soins*, *surveillance du patient en ambulatoire*, *lien* and *équipe spécialisée.*(See Figure 4)

5 Evaluation of extraction of noun phrases

To evaluate this approach, two CPGs were used *Cancer colorectal, Février 2008* and *Cancer colorectal, Janvier 2012* and all the corresponding maximal noun phrases were automatically and manually extracted to be able to calculate: the precision, the recall and the F-measure. These measures are defined as follows:

$$Precision = \frac{Number\ of\ Correct\ maximal\ noun\ phrases\ detected}{Total\ number\ of\ maximal\ noun\ phrases\ detected} \quad (7)$$

Table 2. Patterns compliance. For example : 22.16% of MeSH terms are nouns(NOM), 13.90% are adjectives(ADJ).

Pattern	Percentage of terms	Cumulative percentage of terms
NOM	22.16%	22.16%
$NOM\ ADJ$	13.90%	36.06%
$NOM\ NOM$	06.24%	42.30%
$NOM\ PRP\ NOM$	04.74%	47.07%
$ADJ\ NOM$	03.57%	50.61%
$NOM\ NAM$	02.71%	53.32%
$NOM\ PRP : detNOM$	02.07%	55.39%

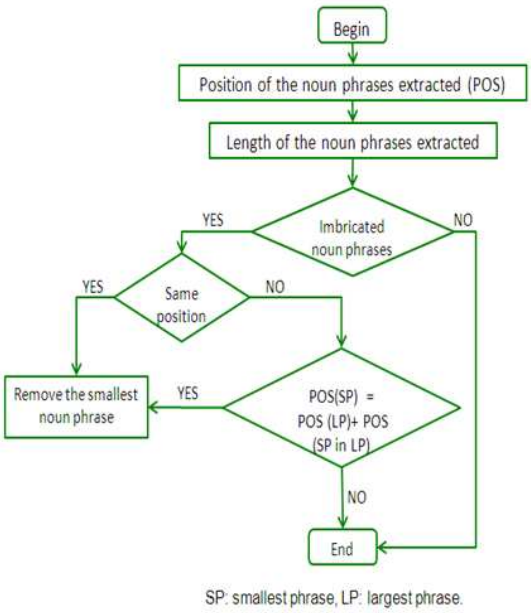

Fig. 3. Algorithm for the extraction of maximal noun phrases

$$Recall = \frac{Number\ of\ Correct\ maximal\ noun\ phrases\ detected}{Total\ number\ of\ maximal\ noun\ phrases\ in\ the\ text} \tag{8}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

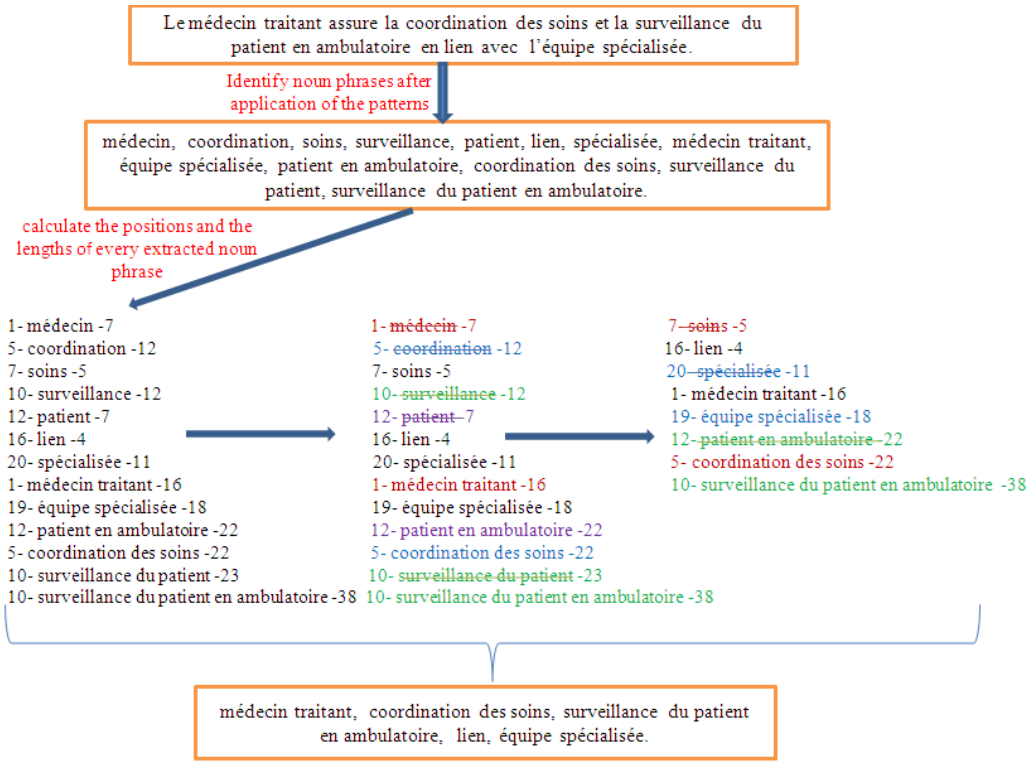

Fig. 4. Example of extraction the maximal noun phrases

In a first step, we used the list of 262 patterns. This list was built using two criteria which are: the length of the pattern and its relevance. The patterns covers more than 80% of all terms in terminologies. To evaluate the validity of this list, we increased it with 231 new patterns. The new list covered 85% of terms in the six terminologies.

6 Results and discussion

Using a list of 262 patterns, 708 maximal noun phrases were found, among them, 364 are correct which represent a 51.41% precision. However by detecting these phrases manually, 602 maximal noun phrases were found which represent a 60.47% recall and by consequence a 55.57% F-measure. We tried to improve these results by increasing a number of patterns from 262 to 493. We obtained a total of 729 maximal noun phrases, with 365 which were correct, which corresponding to a 50.07% precision, 60.63% recall and 54.85% F-measure. These new results are not different from the first results and they comforting the choice of a reduced list of 262 patterns.(see Table 3)

Table 3. Results

Number of Patterns	Precision	Recall	F-measure
262	51.41%	60.47%	55.57%
493	50.07%	60.63%	54.85%

Other options can be tested to improve these results. For example, the use of another tool for tagging (than Tree Tagger), such as GATE Tagger [20]. In another context, we are evaluating our approach of detection of noun phrases by applying it to other medical documents such as the Summary of Product Characteristics (SPCs), which are the legal documents approved as part of the marketing authorization for each drug. They are the basis of information for healthcare professionals on how to use the drug and they are updated throughout the life-cycle of the product as new data emerge. These documents are of interest especially in the comparison step because it exists for each drug, a corresponding SPC which is regularly updated and it will permit us to obtain the documents detailing the same drug but within a 4 or 5 year interval and will give a sense to the comparison and especially to the detection of possible changes.

7 Conclusion

In this paper we presented a useful tool for detection and extraction of noun phrases in order to develop a generic method for comparing medical documents. The method has been used in French but it can easily be applied to other languages. Furthermore, our method is based on the patterns constructed from six medical terminologies (MeSH, ATC, SNOMED INT, FMA, MedDRA and ICD10) that we tagged with a Tree Tagger. These patterns are applied on the CPGs as input to detect and extract the corresponding noun phrases. Then, in the comparison step, we compared different noun phrases and their left and right contexts in order to extract the insertions, additions and substitutions. Finally, a recapitulative file with the differences and the common elements between two or more CPGs inputs is created in order to implement an ergonomic tool to represent knowledge medical evolution. This tool will serve as basis for evaluation of the approach by comparing our results with those of the other existing approaches. The results we obtained conforted us to apply it on other kind of medical documents such as SPCs.

References

[1] Grimshaw, J.M., Russell, I.T.: Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. The Lancet, 342, (8883), 1317-1322, (1993)

[2] Bouaud, J., Séroussi, B., Brizon, A., Culty, T., Mentré and, F., Ravery, V.: How updating Textual Clinical Practice Guidelines impacts Clinical Decision Support Systems: a case study with Bladder Cancer Management. MedInfo, 829-833, (2007)
[3] Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In Proceedings of international conference on new methods in language processing, 12, 44-49, Manchester, UK (1994)
[4] Cheah, T.S.: The impact of clinical guidelines and clinical pathways on medical practice: effectiveness and medico-legal aspects. ANNALS-ACADEMY OF MEDICINE SINGAPORE, 27, 533-539, (1998)
[5] Roche, C.: Terminologie et ontologie. Armand Colin 48-62, (2005), in French.
[6] Lefèvre, P.: La recherche d'informations: du texte intégral au thésaurus. Hermes Science, (2000), in French.
[7] Nelson, S.J., Johnston, W.D., Humphreys, B.L.: Relationships in medical subject headings (MeSH). In Relationships in the organization of knowledge, Springer, 171-184, (2001)
[8] Cornet, R., de Keizer, N.: Forty years of SNOMED: a literature review. In BMC medical informatics and decision making, Springer 8, Suppl 1, 1-6, (2008)
[9] Brown, E.G., Wood, L., Wood, S.: The medical dictionary for regulatory activities (MedDRA). Drug Saf. 20, 109-117, (1999)
[10] Skrbo, A. , Begović, B. , Skrbo, S.: Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes. Medicinski arhiv, 58, 1 Suppl 2, 138, (2004)
[11] Rosse, C., Mejino Jr., José, L.V.: A reference ontology for biomedical informatics: the Foundational Model of Anatomy. Journal of biomedical informatics, 36, 6, 478-500, Elsevier (2003)
[12] Merabti, T., Soualmia, L., Grosjean, J., Palombi, O., Müller, JM., Darmoni, S.: Translating the Foundational Model of Anatomy into French using knowledge-based and lexical methods. In BMC medical informatics and decision making, 11, 1, 65, BioMed Central Ltd, (2011)
[13] Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26 (3), 297-302 (1945),
[14] Levenshtein, V.: Binary codes capable of correcting deletions, insertions and reversals. In Soviet Physics-Doklady, 10, (1965)
[15] Stoilos, G., Stamou, G., Kollias, S.: A string metric for ontology alignment. Proceedings of the International Semantic Web Conference, 624-637, (2005)
[16] Yujian, L., Bo, L.: A normalized Levenshtein distance metric. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(6), 1091-1095, (2007)
[17] Winkler, W.E.: The state of record linkage and current research problems. Technical report: Statistics of Income Division, Internal Revenue Service Publication, (1999)
[18] Daille, B.: Conceptual structuring through term variations, Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, 18, 9-16, Association for Computational Linguistics, (2003)
[19] Aubin, S., Hamon, T.: Improving Term Extraction with Terminological Resources. Advances in Natural Language Processing: 5th International Conference, Finland, 4139, 380, Springer, (2006)
[20] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE (Version 6), 978-0956599315, (2011),