Construction of an Attribute-Value Representation for Semi-structured Medical Findings Knowledge Extraction
*

Daniel de Faveri Honorato[1], **Everton Alvares Cherman**[1], **Huei Diana Lee**[1],
Maria Carolina Monard[2], **Feng Chung Wu**[1]

[1]Bioinformatics Laboratory – LABI
Engineering and Exact Sciences Center – CECE
Western Paraná State University – UNIOESTE
Itaipu Technological Park – PTI
P.O. Box 961, 85870-650 – Foz do Iguaçu, PR, Brazil

[2]Computational Intelligence Laboratory – LABIC
Institute of Mathematical and Computing Sciences – ICMC
University of São Paulo – USP
P.O. Box 668, 13560-970 – São Carlos, SP, Brasil

Abstract

Data Mining is a process related to analysis, understanding and knowledge extraction from databases. In order to perform this process it is usually necessary to represent the data in the so called attribute-value format. This work proposes an extension of a methodology which supports, through a semi-automatic process, the construction of a table in the attribute-value format from information contained in medical findings which are described in natural language (Portuguese). A case study in which the methodology has been applied to a collection of Upper Digestive Endoscopies' medical findings is presented. Results show the suitability of our proposal.

Keywords: Text Processing, Information Extraction, Document Processing

1 Introduction

With the advance of technology the amount of digitally stored information increases constantly. In order to perform a more complete analysis of such information, it is necessary to properly represent the information so it can be processed and a model which represents the embedded knowledge within the data can be constructed, since a manual analysis is not possible. Data Mining – DM — is a process that can be used to support this task. DM aims at identifying novel, and useful patterns embedded in databases [2].

The DM process is iterative and incremental, and it is usually composed by three phases: pre-processing, pattern/knowledge extraction, and post-processing. The first one is frequently the most expensive, consuming around 80% of the whole DM process [11]. This phase is responsable by data preparation, reduction, and transformation. It is also usually necessary to represent the data in the attribute-value format. In order to perform the pattern/knowledge extraction task it is necessary to select a suitable pattern extraction algorithm. This is a usually iterative task due to the necessary adjustments of the selected algorithm. Once the model is constructed it is then evaluated and validated during the post-processing phase.

With the available technological development there is a large amount of information that can be stored by medical institutions. Medical findings — MF — which describe information in natural language, are usually

*This work was supported by Fundação Parque Tecnológico Itaipu – FPTI – and CNPq – Brazil.

stored in a semi-structured format. However, it is convenient to describe the data in the attribute-value format for the application of the DM process. Thus, the MF information should be interpreted and transformed into the attribute-value format. This transformation, besides expensive, is exposed to the subjective interpretation of the person performing it [3, 7]. Therefore, processes that support the semi-automation of this task provide the benefit of time reduction in mapping new findings to the attribute-value format, besides helping the standardization of MF information treatment [6].

The Information Extraction — IE — research area [8] is related to this work, whose methods are based on syntactical and semantical restrictions. IE methods enable the construction of structured representations of unstructured texts described in natural language having a well-defined grammar. Some related work can be found in [1, 4, 12, 14], which use different techniques for the transformation of MF unstructured information into the attribute-value format. In this work we are considering semi-structured medical findings in which the language grammar, specifically the Portuguese language, does not play a leading role. The methodology proposed in this work, which is an extension of [5], was idealized to support the construction of an attribute-value table from semi-structured medical findings described in natural language (Portuguese).

This work is organized as follow: next Section presents the original methodology [5]. Section 3 presents the extensions developed in this work and Section 4 presents a case study applying the new methodology to Upper Digestive Endoscopies — UDE — semi-structured medical findings, followed by Section 5 presenting the conclusions.

2 Description of the Initial Methodology

As stated previously, the methodology proposed in [5], implemented in Perl [13] using the object-oriented paradigm, aimed at construction an attribute-value table from semi-structured medical findings. In [5] the implemented methodology was applied to MF of Upper Digestive Endoscopies (Figure 2), specifically for information related to esophagus. This methodology consists of two phases, as illustrated in Figure 1.

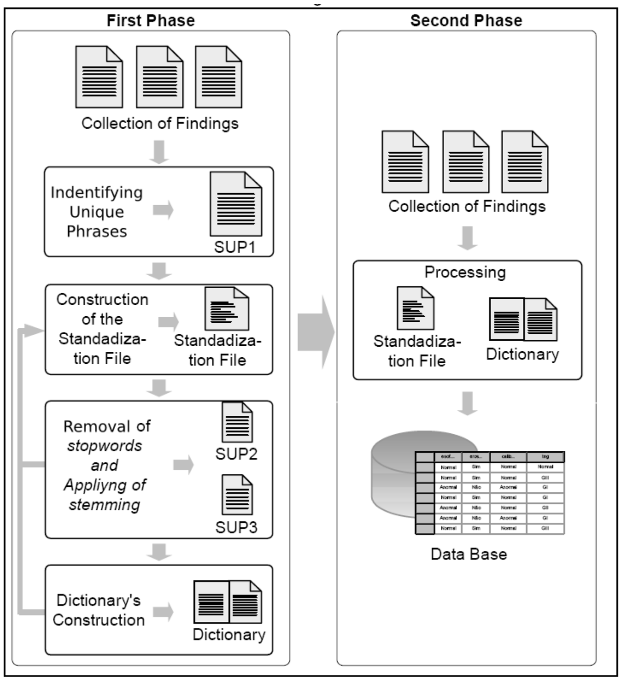
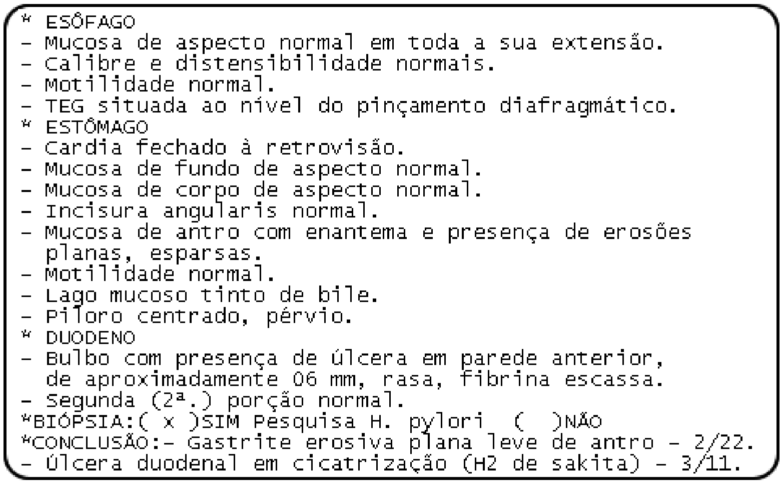


Figure 1: Initial proposal.

In the first phase, a dictionary is constructed with the aid of experts. The contents of the dictionary is mapped from patterns found in the findings. In the second phase, the dictionary is used to map the findings into an attribute-value table using a pattern matching process. It is important to notice that the support of experts is fundamental during the first phase since the construction of the dictionary is strongly based on patterns found in the medical findings and on the domain knowledge provided by experts. Both phases are described next.



```
* ESÔFAGO
- Mucosa de aspecto normal em toda a sua extensão.
- Calibre e distensibilidade normais.
- Motilidade normal.
- TEG situada ao nível do pinçamento diafragmático.
* ESTÔMAGO
- Cardia fechado à retrovisão.
- Mucosa de fundo de aspecto normal.
- Mucosa de corpo de aspecto normal.
- Incisura angularis normal.
- Mucosa de antro com enantema e presença de erosões
  planas, esparsas.
- Motilidade normal.
- Lago mucoso tinto de bile.
- Piloro centrado, pérvio.
* DUODENO
- Bulbo com presença de úlcera em parede anterior,
  de aproximadamente 06 mm, rasa, fibrina escassa.
- Segunda (2ª.) porção normal.
*BIÓPSIA:( x )SIM Pesquisa H. pylori  (  )NÃO
*CONCLUSÃO:- Gastrite erosiva plana leve de antro - 2/22.
- Úlcera duodenal em cicatrização (H2 de sakita) - 3/11.
```

Figure 2: Example of UDE medical findings.

2.1 First Phase

The dictionary construction is performed in four iterative and interactive steps, which are described next.

Identification of unique phrases: In UDE medical findings the information is stored in phrases, each one referring to a diagnosis or an observation from the physician about the performed exam. Therefore, the first step of the dictionary construction aims at joining all different phrases of all medical findings in a single file. During this stage all the phrases of the collection of medical findings are extracted and processed. Identical phrases in the collection are discarded so that only a copy of unique phrases is kept. At the end of this phase, the first set of unique phrases is obtained — SUP1—, which will be the base for the construction of the dictionary.

Construction of the standardization file — SF: The standardization of the data contained in the findings is necessary due to the frequent use of synonyms to describe information in the medical findings, and the presence of phrases that express information in a manner different from the one that will be used in the dictionary. After SUP1 is obtained, it is possible to identify part of the information that can be standardized. In Table 1 two examples of the standardization process are presented.

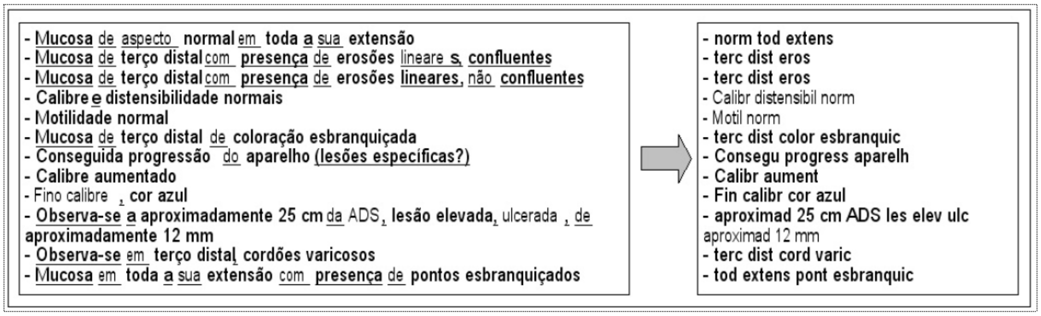
Table 1: Examples of standardization.

Before standardization	**After standardization**
Coloração esbranquiçada	Anormal
Calibre e distensibilidade normais	Calibre normal Distensibilidade normal

The terms "coloração esbranquiçada" in the second line of Table 1, which is used to describe the characteristics of a biological material, state that the biological material is abnormal. The phrase "calibre e distensibilidade normais" in the third line indicates two distinct events, *i.e.*, "calibre" is "normal" (calibre normal) and "distensibilidade" is "normal" (distensibilidade normal). Therefore, the standardization process should map this phrase into two different phrases, one for each event. For example, "calibre normal" and "distensibilidade normal". The construction of the standardization file is performed as soon as information that can be standardized is identified by the expert, and the process continues until the end of the first phase of the proposed methodology. The standardization process enables the mapping of the information contained in the medical findings into a standard format that will be used by the dictionary and by the table filling process during the second phase of the methodology.

Removal of *Stopwords* — RS — and Application of *Stemming* — AS: The objective of this step is to support the process of identifying patterns used by experts to map information in the medical findings. Words that are not important for the analysis of the text of the findings are removed. This kind of words, known as *stopwords*, are kept in a stoplist and are mainly composed by conjunctions, prepositions, articles and words defined by experts as not relevant for the current objective of the process. The removal of *stopwords* from SUP1 generates SUP2. During this step, the stemming process, which may help to eliminate redundancy from SUP2 is used. This process consists of the identification of different inflections from the same word and the substitution of these words by their common radical [7]. This process generates SUP3.

Figure 3 illustrates the removal of *stopwords* and application of stemming on a fragment of SUP1 extracted from esophagus information of the UDE domain. The *stopwords* in the figure are underlined. As can be observed, after the completion of these tasks, the second and third phrases are the same. Thus, one phrase was removed from SUP3. It is important to notice that this process is also supervised by domain experts since there is no guarantee that words having the same stem will have the same meaning. Thus, both SUP2 and SUP3 are considered to help the experts during the analysis of unique phrases to find patterns, and to help in the decision of how information should be organized into the dictionary structure.



- Mucosa de aspecto normal em toda a sua extensão
- Mucosa de terço distal com presença de erosões lineare s, confluentes
- Mucosa de terço distal com presença de erosões lineares, não confluentes
- Calibre e distensibilidade normais
- Motilidade normal
- Mucosa de terço distal de coloração esbranquiçada
- Conseguida progressão do aparelho (lesões específicas?)
- Calibre aumentado
- Fino calibre , cor azul
- Observa-se a aproximadamente 25 cm da ADS, lesão elevada, ulcerada , de aproximadamente 12 mm
- Observa-se em terço distal, cordões varicosos
- Mucosa em toda a sua extensão com presença de pontos esbranquiçados

- norm tod extens
- terc dist eros
- terc dist eros
- Calibr distensibil norm
- Motil norm
- terc dist color esbranquic
- Consegu progress aparelh
- Calibr aument
- Fin calibr cor azul
- aproximad 25 cm ADS les elev ulc aproximad 12 mm
- terc dist cord varic
- tod extens pont esbranquic

Figure 3: Original fragment of SUP1 and the fragment after the removal of *stopwords* and application of *stemming*.

Dictionary construction: As already mentioned, the dictionary is used as the base to construct the attribute-value table. In other words, it helps placing the information contained in the medical findings into the attribute-value table. Therefore, before constructing the dictionary it is necessary to define which features (attributes) should be included in the database. Frequently, medical findings from specific areas have information organized in the form of anatomical structure and their associated characteristics. This characteristic is especially observed in Upper Digestive Endoscopy findings which were used to apply the methodology. Figure 4 illustrates an example of this kind of mapping.

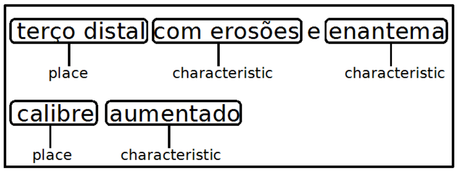


Figure 4: Example of places and characteristic(s).

This way, in the proposed methodology the structure of the dictionary is organized by places and their characteristics. Thus, the dictionary construction is performed in conjunction with the domain experts considering the existing information in SUP3 and the standardization file. The analysis of theses files together with the support of the experts is used in order to identify the local characteristic relations that will be the base to generate the attributes. Figure 5 illustrates the dictionary structure.

As presented in Figure 5, the list P of places stores the name of a certain place P_i and each place holds a list of one or more associated characteristics (C_{ia}), $\forall\ a \geq 1$. The list of characteristics stores, in addition to the name of the characteristic, the corresponding index for the position of the attribute in the Record of Table — RT — and the value that should be stored in the corresponding field of RT.

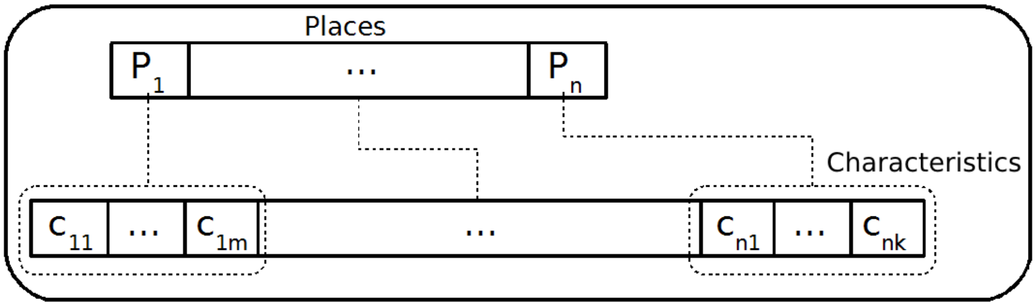


Figure 5: Initial dictionary structure.

2.2 Second Phase

The objective of this phase is to process the collection of medical findings, considering the information mapped within the dictionary structure (places and characteristics), and to fill in the attribute-value table. Each finding refers to a RT. The storing process is performed by a searching and filling algorithm which is applied as illustrated in Figure 6.

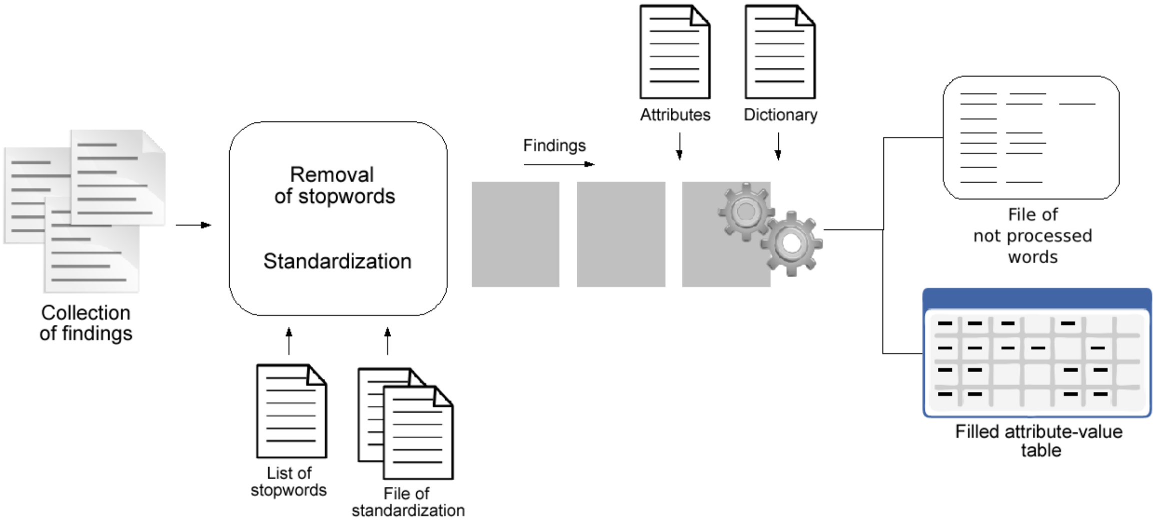


Figure 6: Attribute-value table construction process.

The storing process first receives a medical finding as an entry. Then, the removing of the *stopwords* and application of the standardization using the SF are performed for this finding and a phrase is extracted. The storing process is performed by pattern matching between the dictionary structure and the extracted phrase. For example, assuming that the place “terço distal” and the characteristic “com erosão” are stored in the dictionary structure, and the phrase “terço distal com erosão” was extracted from the finding, the search and filling algorithm will unify the terms “terço distal” from the dictionary with the terms “terço distal” from the phrase. After having identified the location, the algorithm will try to identify which characteristics will be unified in the dictionary structure and in the phrase. In the illustrated example, the characteristic “com erosão” in the list of characteristics associated to the place “terço distal” is presented. This characteristic will be unified with the terms “com erosão” from the phrase. If the place and characteristic match with the terms from the phrase, the attribute in the table will be filled with the specified value in the dictionary structure. It is interesting to notice that all information that is not processed by the dictionary during the table filling process is saved for later processing. This information can be used to detect the existence of new attributes that were not present in the set of findings used to first construct the dictionary. If these attributes are considered relevant by the domain expert than new attributes are added to the dictionary structure. This process is repeated until all the phrases of the findings are completely processed. In the end, the fulfilled RT with the findings information is inserted in the table and a new iteration is started with the processing of the next collection finding. Once this process is over, a table containing the patterns identified from the medical findings is obtained.

3 Dictionary Expansion

This Section presents the expansion built for the methodology proposed in [5]. Two main changes were proposed. The first one refers to an expansion in the original dictionary construction during the first phase. The second change is related to some modifications in the search and filling algorithm for the second phase. The methodology presented in Section 2 covers the mapping of structured information presented in the form of places and characteristics, such as in the portion of esophagus in UDE findings. However, during the analysis of the stomach portion (*ESTOMAGO in Figure 2), a necessary extension of the dictionary was identified. Figure 7 illustrates the main differences between esophagus and the stomach.

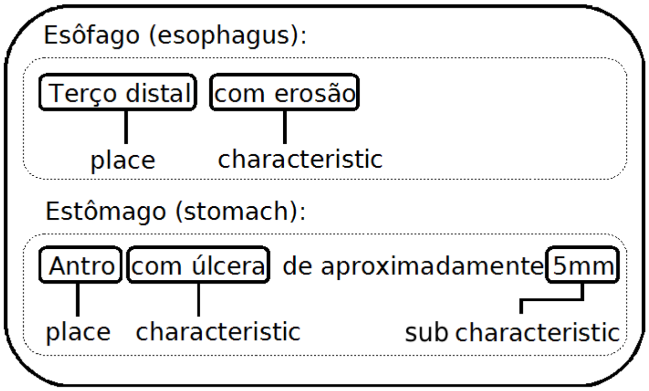


Figure 7: Examples of information referring to esophagus and stomach in a UDE medical finding.

As illustrated in this figure, in order to enable the application of the methodology for stomach information, it is necessary to modify the dictionary structure by adding a new level of information. We named this new level "sub characteristic level". Figure 8 illustrates the proposed expansion schema for the dictionary structure.

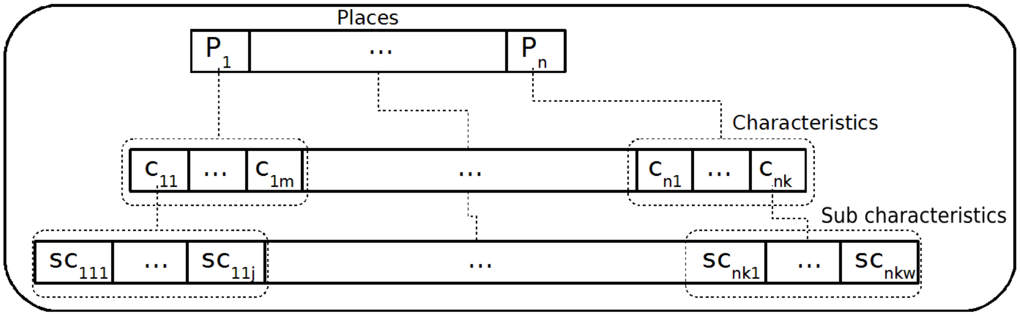


Figure 8: Expanded structure of the dictionary: schematic representation with three levels of information.

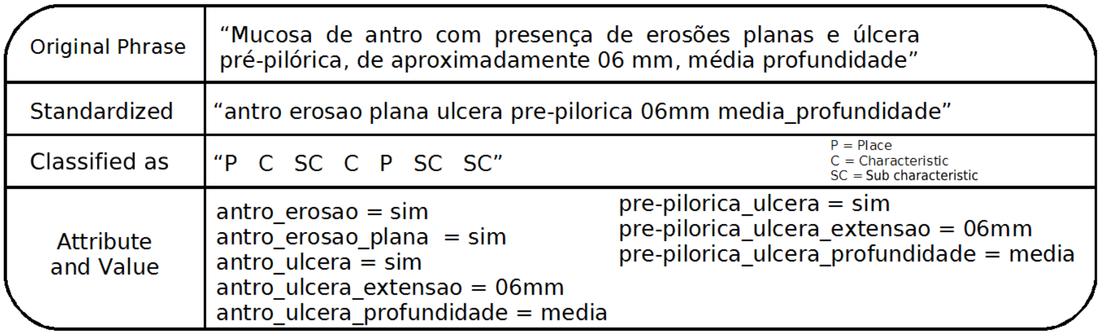
The first and the second levels are identical to the ones of the first methodology. On the other hand, this new structure has a new list of one or more sub characteristics associated to each characteristic. Similarly to the list of characteristics, the list of sub characteristics stores information about the attribute and the value of the attribute that should be inserted in the attribute-value table. Due to this modification in the dictionary structure, there is an increase in the number of attributes since the relation of the sub characteristic of the local-characteristics also represents an attribute. Thus, it was necessary to design and implement a new search and filling algorithm. This new algorithm aims at combining the places, characteristics and sub characteristics described in the findings. This is done by searching attribute values defined for the attribute-value table in these relations. Next, the proposed extensions are discussed in detail.

3.1 Information Distribution

After the analysis of unique phrases, the existence of two patterns was detected: a general pattern, according to the pattern identified in the initial methodology, and another pattern, present in the distribution of the words in the phrases, specifically in the case of the stomach. In the first observed pattern each phrase was referring to a diagnosis or an observation from the physician about the performed exam. With respect to the distribution of the words, the presence of a place and a characteristic is essential, otherwise the information contained in the phrase does not represent any attribute that can be filled in in the table. Since the sub characteristic is a specification of a characteristic, its presence is not strongly required although it may make the information description more detailed, and consequently more attributes in the attribute-value table will be created. After the phrases analysis together with the experts, it was identified that the characteristics which are described in phrases do not refer to all the places present in the phrases, necessarily. In the stomach portion, for example, the information is disposed in a different manner. Thus, it is necessary to associate them correctly, in order to keep the original interpretation. The following patterns were identified based on the results analysis:

- For all characteristics one or more places are necessary such that consecutive places share the characteristic;
- A characteristic can only be a specification of one or more consecutive places found afterwards case there is no characteristic after or before these places;
- All the described sub characteristics are preceded by a related characteristic.

Figure 9 shows a phrase which exemplifies the patterns previously described.



Original Phrase	"Mucosa de antro com presença de erosões planas e úlcera pré-pilórica, de aproximadamente 06 mm, média profundidade"	
Standardized	"antro erosao plana ulcera pre-pilorica 06mm media_profundidade"	

Classified as	"P C SC C P SC SC"	P = Place
		C = Characteristic
		SC = Sub characteristic

Attribute and Value	antro_erosao = sim
	antro_erosao_plana = sim
	antro_ulcera = sim
	antro_ulcera_extensao = 06mm
	antro_ulcera_profundidade = media
		pre-pilorica_ulcera = sim
		pre-pilorica_ulcera_extensao = 06mm
		pre-pilorica_ulcera_profundidade = media

Figure 9: Standardization and classification of an original phrase.

In this figure it can be observed that the places "antro" and "pré-pilórica" share the characteristic "úlcera", although the characteristic "erosões" belongs only to "antro". Furthermore, the sub characteristic "planas" is associated to "erosões", and the sub characteristic "06mm" and "média profundidade" to "ulcera".

3.2 The Search and Filling Algorithm

Figure 10 illustrates the process performed by the newly developed search and filling algorithm. The process is similar to the one applied in the previous methodology. The main update refers to the incorporation of four auxiliary structures and the modification of the pattern matching algorithm. In this process, the pre-processing task is similar to the initial methodology. After each finding is pre-processed, a processing of each phrase is realized. This is done for all the findings in the collection.

The phrase is evaluated from left to right. Using the dictionary information, the words categories such as places, characteristics and sub characteristics are identified. If a word does not belong to any of these categories, it is then added to a file of non-processed phrases for further analysis by the experts. Four auxiliary structures were defined for the identification of the relations between places, characteristics and sub characteristics since these relations (a place with one or more characteristics) are not present in the original methodology. Theses structures are used to memorize places, characteristics and sub characteristics identified during the initial processing of the phrase.

The four new structures are: Indefinite Places Set— IPS; Definite Places Set — DPS; Characteristics Set — CS; and Sub characteristics Set — SCS. All characteristics and sub characteristics retrieved from the phrases are

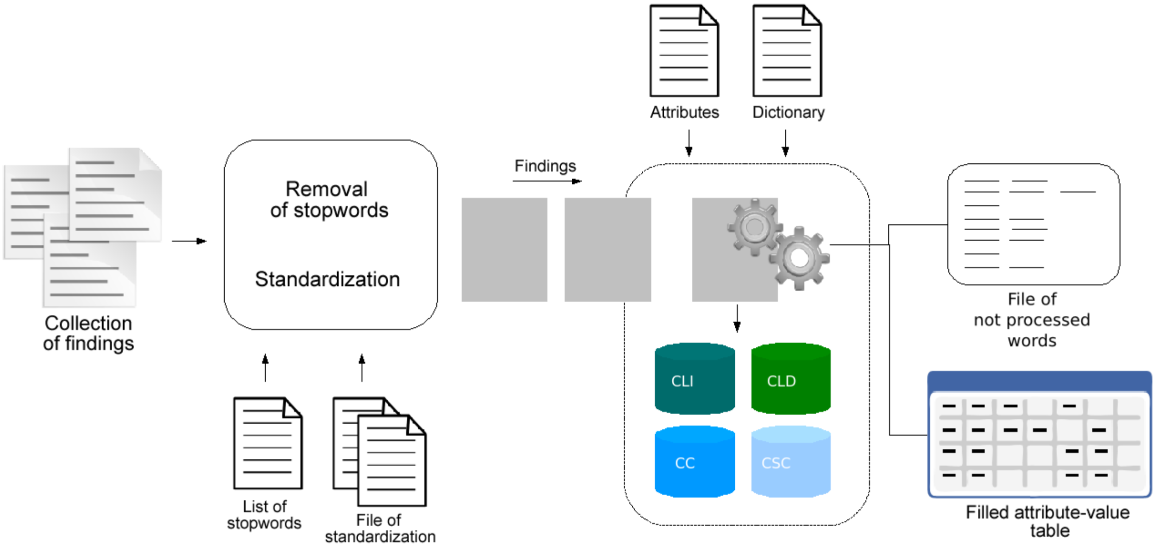


Figure 10: The new process of attribute-value table construction.

stored in CS and SCS, respectively. IPS is used to store the places identified in the phrase until a characteristic or a sub characteristic is identified. If a characteristic is identified, the places are sent to CLD and will be associated with the identified characteristic. If a sub characteristic is identified, the places are also sent to CLD, but they will be associated with the preceding characteristic. Once the information is identified, the filling of the attribute-value table is performed. The filling is performed by searching the built dictionary. During this search the places of DPS and the characteristics of CS are associated. Also, the characteristics of CS are associated to the sub characteristics of the SCS. The search allows the identification of the attributes that should be inserted in the table as well as the values of these attributes. This process is performed for all phrases in the findings and for the collection of findings. The result is a filled table with the patterns identified in the findings.

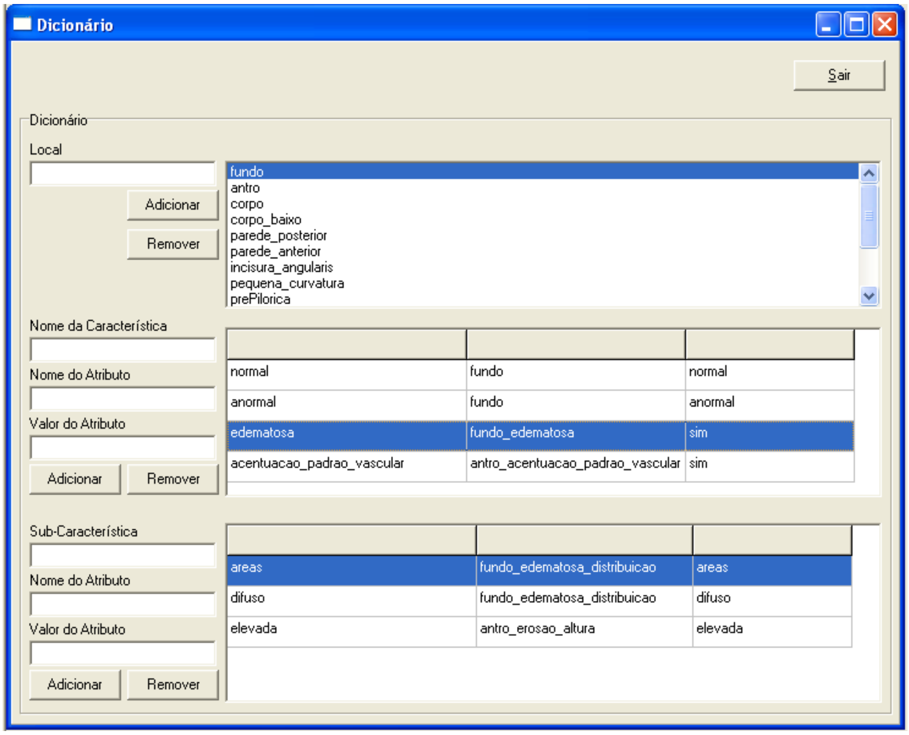
3.3 Computational Tool

A Computational Tool — CT was developed for the application of the new methodology. This new tool also solves several problems found in the implementation of the previous methodology [5] where the information was mapped into text files and separated only by textual marks. For example, in the previous methodology information stored in the dictionary was mapped as *[place*$_1$ *characteristics*$_1$*]* *[next]* *[place*$_2$ *characteristics*$_2$*]...[next][place*$_n$ *characteristics*$_n$*]*, where the word *[next]* represents a textual mark. This kind of mapping caused an additional difficulty because a simple typing error or changing the order of pieces of information could cause an erroneous interpretation by the algorithms implemented. Another problem with the previous implementation was related to the construction of the different files (dictionary file, standardization file, list of attributes and list of *stopwords*), where theses files were created separately. However, it is important to have a joint vision of them when building these files, because the dictionary construction, for example, depends on the file with information on the patterns and also on the list of attributes from the table.

To this end, in order to avoid typing errors and to have a friendly environment to apply the methodology, a new CT was developed. Using this new tool it is possible to build all necessary files to apply the methodology, besides having a joint vision of these files. Furthermore, the information is now stored in a XML file. The XML structure makes possible to store the information in an adequate format for later retrieval.

Figure 11 illustrates the interface used to insert information in the dictionary, while Figure 12 illustrates the XML file that refers to the information mapped by this interface.

Once the input files are built they are used by the search and filling algorithm to fill in the attribute-value table.



fundo
antro
corpo
corpo_baixo
parede_posterior
parede_anterior
incisura_angularis
pequena_curvatura
prePilorica

normal	fundo	normal
anormal	fundo	anormal
edematosa	fundo_edematosa	sim
acentuacao_padrao_vascular	antro_acentuacao_padrao_vascular	sim

areas	fundo_edematosa_distribuicao	areas
difuso	fundo_edematosa_distribuicao	difuso
elevada	antro_erosao_altura	elevada

Figure 11: Interface for the dictionary information registration.

4 Case Study

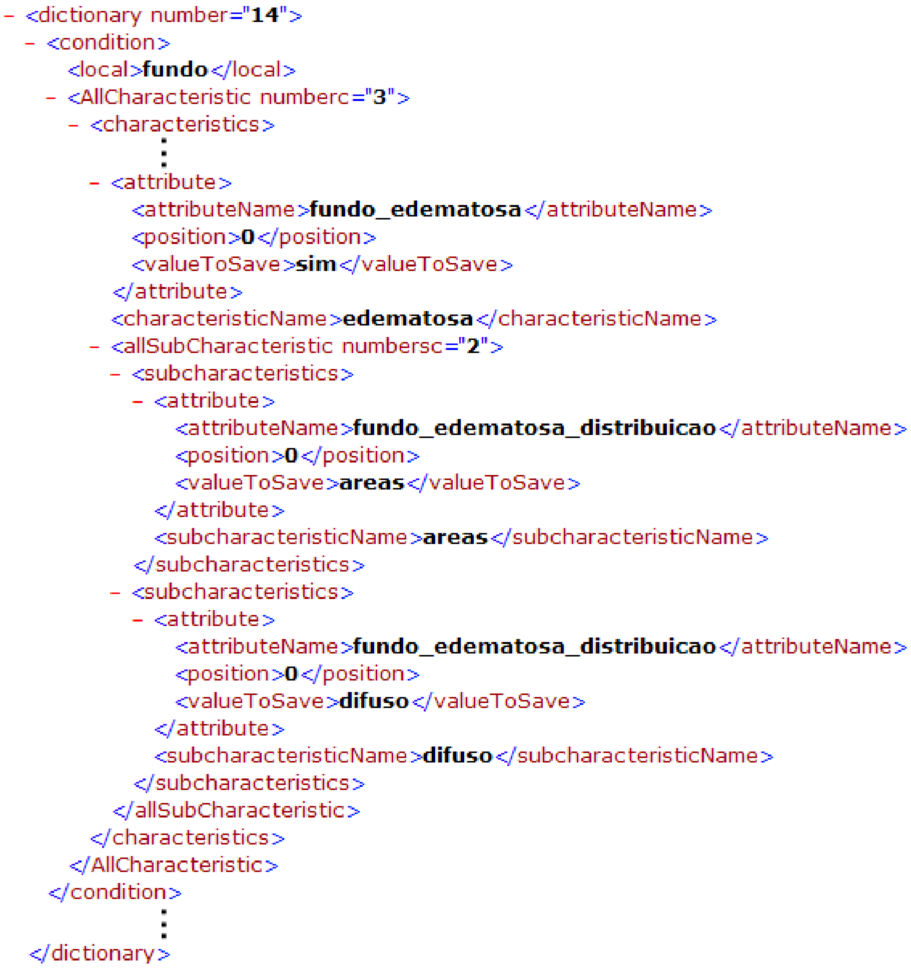
Nowadays, gastroduodenal peptic diseases such as ulcers and gastritis characterize a frequent pathological entity in the population. For this reason, there is an increasing research interest in this area [10]. UDE is an important exam for the diagnosis of such diseases. From this exam, a medical findings report with information about esophagus, stomach, duodenum and results of the exam is generated. In this case study, a collection of 609 findings without patient identifications were used. These findings were collected by the Digestive Endoscopies Service of the Municipal Hospital of Paulinia, São Paulo state, Brazil, from March to November of 2001. The methodology and the new CT presented in the previous section were used to build the attribute-value table from stomach information present in the collection of 609 findings.

4.1 Results and Discussion

Initially, 5213 phrases from stomach information present in the collection of findings were collected. After the identification of unique phrases, 348 phrases were stored in SUP1, representing 6.7% of the 5213 initial phrases. *i.e.* a reduction of 93.3%. Afterwards, the construction of the standardization file from the information contained in the SUP1 file was initiated with the experts support. The removal of *stopwords* and the standardization process were applied, resulting in SUP2 with 259 phrases. SUP2 represents a reduction of 25.6% compared to SUP1. Thus, it is possible to represent all phrases from the stomach findings with only 5% of the initial 5213 phrases. The *stemming* process was applied to SUP2 generating SUP3 with 257 phrases, *i.e.*, two phrases less than SUP2. Figure 13 illustrates some phrases in SUP1 (on the left side) and the respective phrases in SUP3 (on the rigt side) after the removal of *stopwords*, standardization and application of the stemming process.

Afterwards, using the set of unique phrases SUP3, the standardization file, and the assistance of the domain experts, two experiments were performed. The first experiment was carried out with the methodology developed in this work, whereas the second experiment was carried out with the previous methodology proposed in [5]. The results are presented in Table 2.

It can be observed that the attribute-value table built using the new methodology has 130 more attributes than the previous methodology (30 + 130 = 168). As already mentioned, the new methodology considers the relations place × characteristic × sub characteristic in order to construct an attribute, while the previous



```
- <dictionary number="14">
  - <condition>
      <local>fundo</local>
    - <AllCharacteristic numberc="3">
      - <characteristics>
              ⋮
        - <attribute>
            <attributeName>fundo_edematosa</attributeName>
            <position>0</position>
            <valueToSave>sim</valueToSave>
          </attribute>
          <characteristicName>edematosa</characteristicName>
        - <allSubCharacteristic numbersc="2">
          - <subcharacteristics>
            - <attribute>
                <attributeName>fundo_edematosa_distribuicao</attributeName>
                <position>0</position>
                <valueToSave>areas</valueToSave>
              </attribute>
              <subcharacteristicName>areas</subcharacteristicName>
            </subcharacteristics>
          - <subcharacteristics>
            - <attribute>
                <attributeName>fundo_edematosa_distribuicao</attributeName>
                <position>0</position>
                <valueToSave>difuso</valueToSave>
              </attribute>
              <subcharacteristicName>difuso</subcharacteristicName>
            </subcharacteristics>
          </allSubCharacteristic>
        </characteristics>
      </AllCharacteristic>
    </condition>
              ⋮
  </dictionary>
```

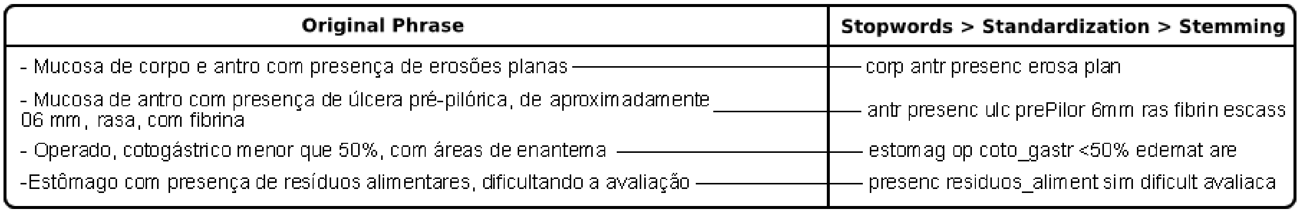
Figure 12: Example of the *XML* code generated using the new computational tool.

methodology only considers the relation place × characteristic for the same process. Therefore, a more detailed mapping of the information present in the findings is performed with the new approach. It can also be observed that in the previous approach the table was composed of 23142 cells (609 × 38) while the search and filling algorithm completed 4934 cells. On the other hand, the attribute-value table constructed using the new approach was composed of 102312 (609 × 168) cells and the search and filling algorithm filled in 5875 cells. Although the previous approach obtained a higher filling percentage in relation to the table size, in the new approach 941 new more cells were filled in. This means that more detailed information from the findings were found and mapped.

Furthermore, an analysis related to the collection of findings and to the resulting attribute-value table was performed for both approaches. This analysis showed that the correct mapping of the table was 100% for the information present in the findings and for the information that was mapped into the corresponding dictionary structure. Thus, the whole amount of information present in the findings and correctly mapped into the dictionary was filled correctly in the attribute-value table. However, we consider that this perfect performance was obtained due to the fact that the mapping of the findings was performed by only one physician, who kept a homogeneous writing style. It is also important to remember that although the constructed table is

Table 2: Results of the two conducted experiments.

	Number of attributes	Total number of cells	Total number of filled cells
previous methodology	38	23142	4934
current methodology	168	102312	5875



Original Phrase	Stopwords > Standardization > Stemming
- Mucosa de corpo e antro com presença de erosões planas	corp antr presenc erosa plan
- Mucosa de antro com presença de úlcera pré-pilórica, de aproximadamente 06 mm, rasa, com fibrina	antr presenc ulc prePilor 6mm ras fibrin escass
- Operado, cotogástrico menor que 50%, com áreas de enantema	estomag op coto_gastr <50% edemat are
-Estômago com presença de resíduos alimentares, dificultando a avaliação	presenc residuos_aliment sim dificult avaliaca

Figure 13: Phrases before and after the standardization.

sparse, as expected in text mining, it is important to map the maximum possible amount of information from the findings. Afterwards, before the application of the pattern/knowledge extraction algorithms using as input the attribute-value table, an attribute selection process can be used in order to find the most relevant attributes to be used by the pattern/knowledge extraction algorithms.

5 Conclusion

Several expansions performed in the methodology proposed in [5] were presented in this work. A case study using the original and the expanded methodologies using medical information presented in UDE, specifically in the portion of the stomach, is also presented. According to the experimental results, the expanded methodology obtained better results for the mapping of semi-structured information because it enables the mapping of more information in the attribute-value table. In addition, it reduces the time cost for the transformation of the information and avoids interpretation subjectivity. Furthermore, new findings can be easily transformed into new registers of the attribute-value table by simply using the information already in the constructed dictionary. As future work, this new methodology will be used to extract information related to duodenum, which is also present in the collection of findings used in this work. The methodology will also be applied to medical findings of other areas, such as andrology and coloproctology, where we are already working with the assistance of experts in these areas. In order to further facilitate the experts´ work, we are investigating hybrid terminology extraction methods [9] which can be applied to findings in order to identify terminological units representing the domains in use.

References

[1] Bekhouche, D., Pollet, Y., Grilheres, B., and Denis, X. Architecture of a medical information extraction system. In *9th International Conference on Applications of Natural Language to Information Systems* (Salford, UK, 2004), pp. 380–387.

[2] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining.* AAAI Press, 1996.

[3] Ferro, M., Lee, H. D., and Esteves, S. C. *Intelligent data analysis: A case study of the diagnostic sperm processing.* In *Proceedings of the ACIS - CSITeA02* (Foz do Iguaçu, PR, Brasil, 2002), pp. 352–356.

[4] Harkema, H., Roberts, I., Gaizauskas, R., and Hepple, M. Information extraction from clinical records. In *Proceedings of the 4th UK e-Science All Hands Meeting* (Nottingham, UK, 2005).

[5] Honorato, D. D. F., Lee, H. D., Monard, M. C., Wu, F. C., Machado, R. B., Neto, A. P., and Ferrero, C. A. Uma metodologia para auxiliar no processo de construção de bases de dados. In *Anais do V Encontro Nacional de Inteligência, XXV Congresso da Sociedade Brasileira de Computação* (Porto Alegre, RS, Brasil, 2005), pp. 593–601.

[6] Lee, H. D. Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado, ICMC-USP, `http://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/`, 2005.

[7] Monard, M. C., and Lee, H. D. *Processamento de Sêmen Diagnóstico*, 1 ed. Editora Manole, Barueri, SP, Brasil, 2003, pp. 461–463.

[8] Muslea, I. Extraction patterns for information extraction tasks: A survey. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction* (Menlo Park, California, USA, 1999).

[9] Pavel, S., and Nolet, D. *Handbook of terminology.* Minister of Public Works and Government Services Canada, Québec, Canadá, 2002.

[10] Pellicano, R., Fagoonee, S., Palestro, G., Rizzetto, M., Figura, N., and Ponzetto, A. The diagnosis of helicobacter pylori infection: guidelines from the maastricht 2-2000 consensus report. *Minerva Gastroenterol Dietol vol. 50(2)* (2004), 125–33.

[11] Pyle, D. *Data Preparation for Data Mining.* Morgan Kaufmann, San Francisco, CA, 1999.

[12] Taira, R. K., Soderland, S. G., and Fakobovits, R. M. Automatic structuring of radiology free-text reports. *Radiographics 21* (2001), 237–245.

[13] Wall, L., Christiansen, T., and Schwartz, R. L. *Programming Perl*, 2 ed. O'Reilly & Associates, 1996.

[14] Zhou, X., Han, H., Chankai, I., Prestrud, A., and Brooks, A. Approaches to text mining for clinical medical records. In *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing* (New York, NY, USA, 2006), ACM Press, pp. 235–239.