Medical Infobahn for Europe
A. Hasman et al. (Eds.)
IOS Press, 2000

A Toolset for Medical Text Processing

Robert H Baud, Christian Lovis, Patrick Ruch, Anne-Marie Rassinoux
Medical Informatics Division. University Hospital of Geneva, Switzerland

Abstract: *The processing of medical texts is a burden in the absence of a toolset designed for simple operations such as recognizing morphological variants, updating and accessing a word dictionary of the domain and segmenting words with multiple morpho-semantems. The apparent simplicity of these basic operations is an illusion because it soon becomes clear that quality implementation is a long-term task. Coherency between subtasks may be lacking unless strict rules are enforced. In fact, good tools are rarely available or have not been tailored for the medical profession. This paper aims at defining a complete toolset for medical word processing. In addition, it provides relevant examples of the inherent difficulties of this task. It reports on typical results that can be expected from an industry-standard implementation.*

1. Introduction

Textual information is an overwhelming attribute of medical practice, which is always a swing between decision-making and therapeutic action or care dispensation. The decision process is certainly dependent on the availability of adequate information, most of it being in written form. Computer systems store the textual information and retrieve it for display when required. However, such systems are blind regarding the content.

This situation provoked the idea of processing medical texts in order to extract relevant information. The simplest approach is the string pattern matching technique, which has proved to be extremely efficient regarding processing time [1]. A further target beyond this somewhat blind domain independent solution is the processing of several indexes from each paragraph and the implementation of a retrieval algorithm based on them. Any word extraction process, as presented in this paper, is a valuable step in this direction. A recent comparison between database information versus free text information illustrates another application [2]. A more ambitious target aims at full parsing and analysis of texts, towards a knowledge representation of its content.

2. Functional needs

The goal is to process any medical text and to retrieve all the individual words from this text contained in a dictionary of the domain. Basically, failure to reach this goal can be due to: the word being missing from the dictionary, or not being recognized due to a possible morphological variant. The rate of success can be calculated by dividing the number of recognized words by the total number of words. The target should be at least 95% depending on the quality of the text and as high as 99% for a label of excellence. The latter value is rarely reached, however, for numerous practical reasons.

The pursuit of this goal requires a number of distinct functionalities to be described hereafter: 1) Cutting the text into words or units of meaning; 2) Recognizing the basic form from morphological variants; 3) Building a language dictionary; 4) Retrieving a word from the dictionary; 5) Handling morpho-semantems.

The above tasks are not sufficient in order to speak of a so-called tokeniser as advocated in the best practice of linguists. The tasks which are missing are: handling of multi-word expressions (necessary step when expressions are in the dictionary), treatment of contracted forms (currently found in French and German with articles), taking into consideration special characters and other local interferences, provision for correlated distant items (verb and particle in German, 2-parts negation in French), etc. Such tasks, though equally important, are beyond the scope of the present paper.

Finding the words

At first it appears that words are separated by spaces, but this is far from the reality. Multiple accidents are possible, especially in the presence of separators like punctuation marks or parentheses. In addition, other characters may act as a space: a dash is potentially a separator but this is not always true depending on the dictionary available; any "end of line" character is also a separator as well as other characters like the tab. Finally, for some languages, the apostrophe is both a part of the preceding word and a separator.

Recognizing the basic forms

This task is essentially language dependent though similarities may occur at the level of implementation. For several dominant western languages, morphological variants are limited to endings, possibly with a modification of the root (umlaut in German). A rule-based expert system can certainly handle the recognition process of allowable endings and rebuild from the root the basic word as stored in the dictionary. This means that a set of rules acts as the knowledge representation of morphological variations of a specific language: switching the set of rules may be enough to switch to another language.

Morphological variants are generally: plural to singular, feminine to masculine and any case to nominative. However, others are possible like stressed or unstressed form in Romanian [3]. This means that different rules have to be set up for each kind of variant. In addition, a list of all the exceptions to the rules is constructed. An expert rule-based engine must be designed in order to coherently select the rules and to fire them when some initial constraints are met. Each rule will result in a transformation of the word, multiple rules of different kinds being possibly triggered (one rule of each kind). The final resulting word is a candidate to be searched for in the dictionary. The authors' implementation for French has a list of 184 exceptions, 20 rules for singular form and 41 rules for gender form.

Building the dictionary

There are two contradictory targets when building a dictionary: a sufficient size for significant coverage of the domain and a permanent quality when collecting the attributes of words. Pressure to augment the size of the dictionary may easily result in lower quality or rigor for each entry. In a recent paper [4] the authors have shown that a relevant size for a medical dictionary is 40'000 entries: it is certainly not an easy task to build such a dictionary. Different strategies exist in order to develop automatic acquisition of vocabulary [5, 6].

Within the context of the toolset described in this paper, the following attributes must be considered: lexical category, gender, number and correctness, where applicable. Correctness is a Boolean attribute defining a word as correct regarding the scholar's view of the language. Other attributes are necessary for verbs especially with languages like French and German. More about the lexical content of a dictionary can be found elsewhere [7].

One of the real problems, when building a dictionary, is the fact that more or less any noun can be equally represented by a corresponding adjective or prefix: this is a characteristic of the medical domain. The implementation solution is to define an underlying concept, which is made common to all related words. For example, with *lip* comes *cheilo* and *labial*, and with *vertebra* it is a good idea not to forget *spondylo*. This task is intrinsically difficult and is not easily computer-assisted.

Accessing the dictionary

The classical access methods to a word dictionary are based on alphabetical order. This is no longer true with computers and would be an unacceptable constraint when working with enciphered dictionaries. Hopefully, access methods based on powerful index techniques or hashing algorithms are available. They do not have to preserve sequential reading because the only service they need to provide is to indicate whether or not a given word is present in the dictionary.

An efficient access method is certainly a major condition. If the processing of 1000 words typically results in 4000 dictionary accesses and if this operation should only take one second, the time per operation is 250 microseconds. If only half this quantity is really used for dictionary access, 125 microseconds are left. Only an experienced programmer could achieve such a target!

In order to achieve a fast access time, an efficient solution is the letter tree access method, a classical one amongst computer algorithms. This method builds a tree of letter nodes, where words with common left parts share the same branch until they diverge. The benefit of this method is obtained at the cost of a pre-processing of the dictionary and an increase in RAM memory. Accessing a word in the dictionary consists in following a branch not longer than the number of letters in the word. With this approach, the authors have been able to achieve 20 microseconds access time on a 400 Mhz personal computer with the whole dictionary loaded in RAM memory. This means a rate of 50'000 accesses each second!

Analysing the morpho-semantems

The authors have published numerous papers [8, 9, 10] on the importance of morpho-semantems and the need to recognize them in medicine. Other authors share this point of view [11, 12, 13]. At the present time the technology behind this kind of analysis is fully mastered from the authors' point of view. Due to the existing publications on this topic, this will not be developed further.

3. Experiencing with clinical texts

Two sets of textual sources have been chosen for this experiment in French: the 12'317 systematic ICD10 expressions and a set of 20 reports from the digestive surgery Department. The size of the second corpus with relation to the ICD corpus is 6%. This corpus has been entirely anonymized.

The dictionary in use has nearly 30'000 entries, including stop words, proper names, Latin expressions and drug brand names. This dictionary was built from the ICD10 source and for this reason all ICD words are present, resulting in no unknown words. In this experiment, we apply the full morphological resolution and morpho-semantems decomposition, and we search for concepts behind the words.

ICD 10 systematic expressions

The processing of this corpus of text has found 101'017 occurrences of words or other items, resulting in 4612 different words only. Figure 1 gives the distribution by word category. Proper names are mixed with other names.

Nouns and adjectives account for nearly 50% of all words, and verbs are rare: this is the confirmation of the dominance of noun phrases for this classification style. Prepositions are numerous, but nearly 88% of them are the very common *of, to, without* and *in*. The fifth in frequency is *during* accounting for 2.5% of all prepositions. The importance of morphosemantems is clearly visible: the number of prefixes and suffixes is 7837. This value, relative to the total number of nouns and adjectives, gives a percentage of 16%. After the most common modal prefixes like *hyper, anti, extra* and *dys*, the most significant are *ostéo, arthro, néphro, myo, cardio* and *pneumo*. The number of different prefixes is 469. Latin expressions are not as numerous as expected in French, contrary to German.

noun	31829	31,5	short prefix	1821	1,8	ordinal	63	0,1
adjective	18071	17,9	noun suffix	1771	1,8	past part	59	0,1
preposition	16557	16,4	adverb	662	0,7	poss adj	48	0,0
article	12064	11,9	latin expr	416	0,4	dem pron	32	0,0
conjunction	5191	5,1	alphanum	348	0,3	abbrev	11	0,0
punctuation	4694	4,6	verb	199	0,2	int pron	10	0,0
full prefix	4134	4,1	number	117	0,1	pers pron	6	0,0
indef. det	2802	2,8	adj suffix	111	0,1	unknown	1	0,0

Figure 1: Word categories in systematic ICD10 French expressions.

The extension of concepts found is roughly 50% of the significant words. Some 29’000 words have a link to a concept. They are nouns (15’900), adjectives (8’800) or prefix and suffix (3’500). 927 different concepts are used, but only 59 have more than 100 occurrences. The more frequent concepts are: wound (1430), disease (1061), inflammation (955), accident (891), lesion (642), blood circulation (630), trauma (563), affection (541), system (353), infectious (315), syndrome (291), person (287), anomaly (286), acute (282), tumour (268), etc.

Texts from the Electronic Patient Record

A set of 20 randomly selected documents in the sub-domain of digestive surgery has been selected, in order to contrast the ICD classification to clinical documents.

The number of words found is 6132 from which 362 occurrences or 5.9% are unknown, resulting in 214 different words from a total of 1545, or 13.9%. This is not a good performance, but shows that ICD words are not the best for describing surgical procedures.

The groups of unknown words are as follows: missing words (65 words, 43 verbs, 11 proper names, total 119 or 55.6%), local jargon or abbreviation (61 or 28.5%), missing words due to temporary program default or unexpected situation (18 or 8.4%), user error or mistyping (16 or 7.5%).

The rate of known words is slightly above 86%. As a direct consequence, mare than 13% of words are missing in the dictionary and should be added sooner or later. This situation clearly demonstrates the need for a solution of automatic word acquisition from a corpus of text. The rate of user errors or mistyping at 7.5% is not a surprise and computer-assisted input may lower this value to about 1%.

noun	1219	19,9	short prefix	81	1,3	ordinal	14	0,2
adjective	827	13,5	noun suffix	55	0,9	past part	2	0,0
preposition	688	11,2	adverb	135	2,2	poss adj	15	0,2
article	581	9,5	latin expr	23	0,4	dem pron	24	0,4
conjunction	154	2,5	alphanum	258	4,2	abbrev	8	0,1
punctuation	698	11,4	verb	600	9,8	int pron	19	0,3
full prefix	222	3,6	number	20	0,3	pers pron	99	1,6
indef. det	27	0,4	adj suffix	1	0,0	unknown	362	5,9

Figure 2: Word categories in a set of 20 reports from Digestive Surgery and Rx Department.

Figure 2 shows the distribution by word category. It appears that the number of verbs is clearly more important when describing patients in clinical situations, especially surgery reports. As a consequence the proportion of nouns and adjectives is lower. Other categories are not significantly different, apart from punctuation because ICD expressions are stored without a full stop!

The extension of found concepts is 39,4% of the significant words against nearly 50% with ICD. This is caused by the use of verbs, which are rarely linked to concepts in the used dictionary. 320 different concepts are used, and only 51 have more than 6 occurrences. The mean number of concepts found by document is 16. This is a good value for future automatic indexing. The more frequent concepts are different to those expected: patient (37), a meaning absence of (32), right (31), left (29), normal (24), examination (24), lung (23), excision (21), heart (21), spleen (16), treatment (16), abdomen (15), etc.

4. Conclusion

This experience demonstrates the problem of achieving a good coverage of the domain with the dictionary. With the set of 20 documents, the number of "unknown" words is too high for practical application. This aspect need to be improved by further work on word acquisition. Automatic extraction is a must if the target is to lower the rate of unknown words to below 2% (presently 13%). Despite the problem of domain coverage, this experiment is encouraging from a qualitative point of view. Practical word extraction is mastered from the morphological variation (including verbs) and word decomposition points of view. This is achieved from a general viewpoint in the medical domain. Concept assignment is encouraging with a rate of 39 to 50%. The set of concepts may act as an initial version for automatic indexing of medical texts. Conceptual retrieval on this basis is feasible.

References

[1] Lovis C, Baud RH. Fast exact string pattern matching algorithms adapted to the characteristics of the medical language. Accepted for publication in J Am Med Inform Assoc. 2000.

[2] Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the degree of concordance of coded and textual data in answering clinical queries from a classical data repository. J Am Med Inform Assoc. 2000;7(1):42-42.

[3] Filip F, Haras C. CIM Explorer – Intelligent Tool for Exploring the International Classification of Diseases (Romanian Version). Submitted to MIE'2000, Hanover, Germany, Aug 2000.

[4] Baud RH, Ruch Patrick, Lovis C, Rassinoux A-M. Recherche conceptuelle dans les textes médicaux. Journées francophones d'Informatique médicale, Marseille 2000, Springer Verlag.

[5] Hersh WR; Campbell EH; Evans DA; Brownlow ND. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. Proc AMIA Annu Fall Symp (United States), 1996, p159-63

[6] Baud R; Lovis C; Rassinoux AM; Michel PA; Scherrer JR. Automatic extraction of linguistic knowledge from an international classification. Medinfo (Canada), 1998, 9 Pt 1 p581-5.

[7] McCray AT. The nature of lexical knowledge. Methods Inf Med (Germany), Nov 1998, 37(4-5) p353-60

[8] Lovis C; Baud R; Rassinoux AM; Michel PA; Scherrer JR. Medical dictionaries for patient encoding systems: a methodology. Artif Intell Med 1998 Sep-Oct;14(1-2):201-14.

[9] Lovis C; Baud R; Michel PA; Rassinoux AM; Rodrigues JM; Scherrer JR. Full text multilingual automatic morphosemantems for stand-alone or Internet based applications. Medinfo 1998.

[10] Baud RH; Lovis C; Rassinoux AM; Scherrer JR. Morpho-semantic parsing of medical expressions. Proc AMIA Symp 1998;:760-4.

[11] Norton LM, Pacak MG. Morphosemantic Analysis of Compound Word Forms Denoting Surgical Procedures. Meth Inform Med, 22: 29-36, 1983.

[12] Wolff S. The Use of Morphosemantic Regularities in the Medical Vocabulary for Automatic Lexical Coding. Meth Inform Med, 23: 195-203, 1984.

[13] Dujols P, Aubas P, Baylon C, Grémy F. Morphosemantic Analysis and Translation of Medical Compound Terms. Meth Inform Med, 30: 30-35, 1991.