Medical Informatics Europe 2002
G Surján et al. (eds)
IOS Press 2002

Combining voice recognition and automatic indexing of medical reports

André HAPPE [1], Bruno POULIQUEN [2], Anita BURGUN [3],
Marc CUGGIA [3], Pierre LE BEUX [3]

[1] *Intermède La Basse Revachais 35580 GUIGNEN – France (happe@intermede.net)*
[2] *European Commission , IPSC, Joint Research Center – 21020 ISPRA – Italy*
[3] *Laboratoire d'Informatique Médicale – Faculté de Médecine –*
35033 RENNES Cédex - France

Abstract. Medical records have been evolving from the traditional paper-based records to digital ones, from the method of dictating reports and transcription to voice recognition systems. The transition to digital operations will not be complete until we have the ability to combine voice recognition with automated indexing of texts. This paper introduces the methods we used to evaluate existing voice recognition software programs and presents NOMINDEX, a system that turns a medical text into MeSH codes, using the French ADM lexical database. Those systems were applied to 28 patient discharge summaries in French, produced after a coronarography, and extracted from the MENELAS corpus of texts. Using the best configuration for voice recognition, the rate of accurate recognition exceeds 98 percent. Among the indexing concepts assigned by NOMINDEX, 25 percent were not pertinent and 12 percent of the relevant concepts were missing. Most errors were related to confusion between common language and medical language, and to the coverage of the ADM lexical database. Best results would be expected with a more comprehensive lexical resource In addition, only 3 percent of the errors generated by inadequate voice recognition that remained in the configuration that performed better, impacted on automatic indexing by NOMINDEX.

1. Introduction

Several studies evaluate speech recognition software as an alternative to medical transcription. Accuracy rates as high as 98 percent are reported [1]. Computer voice recognition has a shorter turnaround time [2], and is less expensive than traditional transcription [3].

The transition to digital operations will not be complete until we have the ability to combine voice recognition with automated indexing of texts. In several projects, methods are developed, whereby automated indexing methods substitute for manual indexing practices, e.g., [4], [5], [6].

Our objective was to combine voice recognition and automated indexing of medical reports. This study was conducted to evaluate the performance of such a system. We used commercially available voice recognition software packages, and NOMINDEX, a program that we developed and that identifies concepts in medical texts and creates a list of MeSH indexing terms.

2. Material & methods

2.1. Corpus of texts

A set of 28 hospital medical summaries produced after a coronarography constitute the corpus of texts used for testing voice recognition and indexation. They were extracted from the corpus of hospital summaries in French established in the frame of the European project MENELAS [7]. The 28 texts were analysed by two physicians separately to determine for every word whether it belongs to the common vocabulary or to the medical vocabulary. The characteristics of the texts are given in Table 1.

Table 1: Characteristics of the texts

	Words	Words of general French vocabulary	Words of medical vocabulary	Marks of punctuation	Numbers
Average/ text	140.2	119.7	20.3	25	3.8
Standard deviation/text	44.4	38.3	7.1	9.3	2.1
Total of 28 texts	3925	3354	571	734	108

2.2. Tools

2.2.1. Voice recognition

The two voice recognition software packages that were tested were:

- Naturally Speaking 5 professional version (Lernhout and Hauspie™) in standard configuration and with the medical dictionary specialized in Cardiology
- IBM®Via Voice 8 Professional version in standard configuration and with the medical dictionary specialized in Cardiology.

A procedure of digital recording on CD-ROM of the various texts was used in order to standardize their dictations during the various phases of the experiment. Recordings are then proposed to the computer via a CD-ROM drive using the audio input. A single operator recorded all the texts. The phases of initialisation recommended by the editors of two software packages were executed from specific digital recordings. All the treatments were realized on a Pentium III desktop computer (1.4 Ghz with 512 Mb RAM running on Microsoft Windows 2000).

2.2.2. Automatic indexing.

The NOMINDEX program has been developed in order to index all kinds of medical texts [8] . It extracts MeSH®concepts from texts in natural language. NOMINDEX has been written in Perl, using a relational database (Oracle) and the user interface runs under an Apache web server. It runs on a Sun Unix system (Solaris).

NOMINDEX uses the French ADM lexical database [9]. The ADM lexical database contains about 50000 words related to diseases, signs and symptoms, occupation and others items that are necessary to describe medical conditions. The originality of this lexicon is that it includes multiword units, including compound words (e.g., "yellow fever") but also associated words (e.g.,"head pain") which are recognized in sentences like "my patient has a head pain" or "a big pain at head's level". It groups the words into sets (one set contains the flexional words, the synonyms and some derivations). For example, the words "headache", "headaches", "head pain", "cephalgia" belong to the same set.

The MeSH thesaurus is extracted from the UMLS Metathesaurus. More precisely, we focus on the French translation of the MeSH terms [10]. Each MeSH term is indexed by the words it contains. In addition, we use the taxonomy of concepts from the UMLS® [11] (i.e., the "is-a" semantic relation between concepts).

The initial step of indexing consists of segmenting the texts into sentences. For each sentence, we first extract words. Then, for each MeSH term, we test whether all the words it contains are present in the sentence or not. Knowing the terms we then extract the concepts. An additional process generates the hypernyms of the produced concepts (looking at the UMLS taxonomy). The concepts that are found in a sentence are weighted, using the TFIDF formula [12], which allows to put a higher score on concepts that are more specific.

2.3. Methodology

2.3.1. Voice recognition

The objective was to measure the relative performance of voice recognition packages according to their ability to recognize the words of the texts. Two cycles of treatments were set up and will be called later on as follows:

- " at blank ": during this cycle, none of the correction proposed by tools was recorded.
- " with learning ": during this cycle, a set of 6 randomly selected texts of the corpus are given to the tools for recognition and the corrections are recorded before the test is performed on the 22 other texts.

For each tool, the two cycles were executed first using the standard configuration and then using the available medical vocabularies.

The errors of recognition are pointed out as soon as a difference exists between a word of the reference text and what was recognized and suggested by the tested tool. Voice scoring categories, adapted from [13] were:

- Recognition of general French vocabulary
- Recognition of medical vocabulary
- Recognition of numbers
- Recognition of punctuation marks

Concerning the numbers, it was admitted that an non ambiguous restitution of numbers expressed numerically did not constitute an error (e.g., twenty instead of 20). Besides, for the composed words it was decided to consider as acceptable the spellings with or without hyphens.

2.3.2. Automatic indexing

For each document, a set of keywords were manually extracted. Only basic keywords were extracted, i.e., no inference based on is-a relationships was performed. Similarly, concepts that were inferred using the UMLS taxonomy but were not present in the document as such, were not taken into account. For example, “Cardiovascular Disease” can be inferred from “Angina Pectoris”, while not present in the text. In this case, the concept “Cardiovascular Disease”, although suggested by NOMINDEX, was ignored.

The evaluation of NOMINDEX consisted of a comparison between the set of concepts extracted by NOMINDEX and the set of keywords manually extracted. Noise was calculated using the following specific criterion: when a concept C exists in the document D, then, even if C is present in a negative form, C is pertinent. For example, from “without evolution towards necrosis”, the extraction of “Necrosis“ was considered correct.

3. Results

3.1. Voice recognition

Results are presented as percent of errors in recognition related to the number of items of each scoring category. We only present the data concerning cycles with the medical dictionaries. Overall results of Via Voice, respectively, "at blank" and "with learning" were significantly better than those of Naturally Speaking.

Table 2: Rates of errors of the different scoring categories

Scoring category	"At blank"			"With learning"		
	Via Voice	Naturally Speaking	p	Via Voice	Naturally Speaking	p
medical vocabulary	3.68 %	8.23 %	0.001*	2.16 %	4.31 %	0.063
general French vocabulary	1.13 %	3.59 %	< 0.001*	1.22 %	2.67 %	< 0.001*
numbers	2.31 %	6.02 %	0.173	1.19 %	7.14 %	0.122
punctuation	0.43 %	0.36 %	0.682	0 %	0.54 %	0.247
Total rate of errors	**1.36 %**	**3.73 %**	**< 0.001***	**1.16 %**	**2.66 %**	**< 0.001***

3.2. Automatic indexing

Indexing the twenty-two coronarography PDS produced by Via Voice "with learning" voice recognition, using NOMINDEX resulted in 560 indexing concepts. The average number of indexing concepts per PDS is 25 (minimum: 10, maximum: 55).

3.2.1. Noise

Among the 560 indexing concepts, 140 (25%) were not pertinent.

Several mechanisms were involved, including:

- Polysemy, e.g., from "récidive angineuse", the concept Amygdalitis was extracted, since the French word "angineuse" can refer either to amydalitis or to angina pectoris, which share the notion of constriction.
- Inappropriate semantic field, e.g., from "sixth month", a concept related to paediatrics was suggested by NOMINDEX. Similarly approximate semantic interpretation may occur, e.g., from "bypass with mammary artery", the concept "nipple" was extracted.
- Since the French translation of MeSH terms results in terms in capital letters, accentuated characters cannot be taken into account. As a result, the French word "serré" (tighten) ends up indexed as "hoof and claw" (C0019909), which corresponds to «SERRE » (claw).

3.2.2. Silence

50 concepts that were manually extracted from the PDS were not found among the concepts extracted by NOMINDEX. Those missing terms mostly correspond to:

- abbreviations, such as IDM, which stands for myocardial infarction in French,
- procedures, (e.g., lobectomy, bypass) which are out of the scope of the French ADM lexicon
- brand names for drugs, which are unknown
- specific anatomical terms, such as specific coronary arteries.

4. Discussion

Besides the classical problems of "word sense disambiguation" (i.e. a same written word belongs to different semantic fields) inherent to text indexing, the process of combining voice recognition and automatic indexing has to overwhelm also the "sound sense confusion"(i.e. a same sound refers to different concepts for example "HTA" which stands for "hyper blood pressure" is recognized as "acheter à" which means "to buy at").

The comparison between the results of (1) indexing of the 22 original PDS that were used for our study (560 concepts, 141 non relevant, 48 missing concepts), and (2) indexing of the same PDS as the output of a voice recognition package (Via Voice + medical vocabulary + learning) shows no difference as far as global performance is considered. However, four errors in voice recognition had an impact on the indexing results. Two concepts were missing just because the voice recognition system failed at recognizing a medical term. The word "chirurgie" (surgery) was not recognized, and "bronchectasie" was translated into "bronche ectasie", from which NOMINDEX could not extract the MeSH concept for bronchectasia. On the other hand, values for blood pressure were represented as "# mm de mercure" instead of "# mm Hg". As a result, the MeSH concept "C0025424 mercury" was extracted, although not relevant in this context. In addition, when medical acronyms are not recognized, the voice recognition system replaces it by something that sounds identically. Thus, "IVA", which is an acronym for a coronary artery was replaced by a sentence including the word "lit", which means "bed"; As a result, the MeSH concept "C0004916 Beds" was extracted, although, of course, not relevant.

The ADM lexical database is the French lexical resource that is used by NOMINDEX in order to map terms. Consequently, a lack of synonymy relationships in this resource as well as its coverage of the biomedical domain may affect the performance of the NOMINDEX system. In our experiment, some limits of the method are related to missing acronyms and missing concepts (especially for procedures, and anatomy). In addition, needs for a lexicon that would take into account accentuated characters are put forward. Best results would be expected with a more comprehensive lexical resource for the biomedical domain.

Acknowledgments

The authors express their thanks to Alain MEYER, Julien BEGUEC and Jean GARREC for their participation in this study as well as to Nicolas LIOUVILLE (Mysoft) who provided Naturally Speaking and the medical dictionaries for this study. This study was partially supported by funding from CRITT Santé.

References

[1] Zafar A, Overhage JM, McDonald CJ. Continuous speech recognition for clinicians. *J Am Med Inform Assoc.* 1999 May-Jun;6(3):195-204.

[2] Zick RG, Olsen J. Voice recognition software versus a traditional transcription service for physician charting in the ED. *Am J Emerg Med.* 2001 Jul;19(4):295-8.

[3] Borowitz SM.Computer-based speech recognition as an alternative to medical transcription. *J Am Med Inform Assoc.* 2001 Jan-Feb;8(1):101-2.

[4] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM Indexing Initiative. *Proc AMIA Symp.* 2000;:17-21.

[5] Franz P, Zaiss A, Schulz S, Hahn U, Klar R. Automated coding of diagnoses--three methods compared. *Proc AMIA Symp.* 2000;:250-4.

[6] Lowe HJ, Antipov I, Hersh W, Smith CA, Mailhot M. Automated semantic indexing of imaging reports to support retrieval of medical images in the multimedia electronic medical record. *Methods Inf Med.* 1999 Dec;38(4-5):303-7.

[7] Zweigenbaum P. MENELAS: an access system for medical records using natural language. *Comput Methods Programs Biomed.* 1994 Oct;45(1-2):117-20.

[8] Pouliquen B., Delamarre D., Le Beux P., Indexation de textes médicaux par extraction de concepts et ses utilisations, 6th International Conference on Statistical Analysis of Textual Data (JADT'2002), (proceedings to be printed), St Malo, France, March 2002.

[9] Lenoir P. , Roger M.J., Frangeul C., Chales G., Creation, development and maintenance of the data-base of a computer-assisted diagnostic system (ADM). *Med. Inform.* (Lond.).1981,Jan-Mar ;6(1) :33-40

[10] National Library of Medicine .*Medical Subject Headings* Bethesda, Maryland

[11] Lindberg DAB, Humphreys BL., McCray AT., . The Unified Medical Language System *Methods Inf Med.* 1993; 4 (32) : 281-91

[12] Salton G., Buckley C., 1988, Term weighting approaches in automatic text retrieval, *Information Processing and Management*, 1988, 24; 5, 513-23.

[13] Devine EG, Gaehde SA, Curtis AC. Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports.*J Am Med Inform Assoc.* 2000 Sep-Oct;7(5): 462-8.