Information Extraction from Medical Reports

Liliana Ferreira[1], António Teixeira[1 2] and João Paulo da Silva Cunha[1 2]

[1] Institute of Electronics and Telematics Engineering of Aveiro,
Campus Universitário de Santiago,
3810-193 Aveiro, Portugal
[2] Department of Electronics and Telecomunications,
Campus Universitário de Santiago,
3810-193 Aveiro, Portugal

Abstract. Information extraction technology, as defined and developed through the U.S. DARPA Message Understanding Conferences (MUCs), has proved successful at extracting information primarily from newswire texts and in domains concerned with human activity. This paper presents an Information Extraction (IE) system, intended to extract structured information from medical reports written in Portuguese. A first evaluation is performed and the results are discussed.

1 Introduction

Information Extraction (IE) is a technology dedicated to the extraction of structured information from texts to fill pre-defined templates [1]. IE still suffers a number of limitations that prevent its dissemination through general public presentations. Among these limitations, we can consider the fact that systems are not really portable from one domain to another.

Textual reports of patient are a vast source of clinical information, but information in textual form is not useful for automated clinical applications, and even if electronically available, the information remains locked up within the text. Text is difficult to access because it is extremely diverse and the meanings of words vary depending on the context. IE systems offer potential solutions because they not only extract individual but also represent well-defined relations among words.

In this paper we describe the first steps toward the development of an IE system from medical reports written in Portuguese. The following section describes, briefly, IE technology, section 3 describes the principal processing stages and techniques of our system and section 4 describes the evaluation methodology. The analysis of the results and its discussion ends the paper.

While this system isn't yet complete, indications are that IE can indeed be successfully applied to the task of extracting information from medical reports.

2 Information Extraction Technology

The most recent MUC evaluation (MUC-7)[2] specified five separate component tasks, which illustrate the main functional capabilities of current IE systems:

Ferreira L., Teixeira A. and Paulo da Silva Cunha J. (2006).
Information Extraction from Medical Reports.
In *Proceedings of the 3rd International Workshop on Natural Language Understanding and Cognitive Science*, pages 149-155
DOI: 10.5220/0002499901490155

1. *Named Entity Recognition* (NER) finds and classifies named entities such as organizations, persons, locations, dates and monetary amounts.
2. *Coreference Resolution* (CO) identifies identity relations between entities in texts. These include variant forms of name expression, definitive noun phrases and their antecedents, and pronouns and their antecedents.
3. *Template Element construction* (TE) adds descriptive information to NE results (using CO).
4. *Template Relation filling* (TR) finds relations between TE entities.
5. *Scenario Template filling* identifies relations between template elements as participants in a particular type of event, or scenario, and the construction of an object-oriented structure recording the entities and various details of the relation.

State-of-the-art (MUC-7) results for these five tasks are as follows (in the form recall/precision): named entity - 92/95; coreference - 56/69; template element - 86/87; template relation - 67/86; scenario template - 42/65.

3 A Brief Description of the System

The past few years have witnessed a growing interest in applying NLP techniques to process and understand biological and medical texts. There have been created many resources and processing tools which facilitate access to desired information.

We are currently investigating the use of IE to provide a formalized description of Portuguese neurological reports reported at Hospital Geral de Santo António (HGSA), Porto, Portugal.

The utility of an IE system for health professionals and doctors lies on the ability to obtain sequences of neurological activity that would only be accessible by several searches over the documentation generated after each examination.

The IE system developed to carry out this task is derived from GATE [4]. GATE is an infrastructure for developing and deploying software components that process human language. GATE has been in development at the University of Sheffield since 1995 and has been used in a wide variety of research and development projects [5].

The architecture consists of a pipeline of processing resources which run in series. Many of these processing resources are language and domain-independent (e.g. Tokenizer and Sentence Splitter). However, the main processing, carried out by a gazetteer and by a set of grammar rules, had to be enriched with language and domain-specific parameters. This process is described in the following subsections.

VMP Tagger. We have chosen to substitute the POS tagger available in GATE for one developed by Valentina Munõz and available at *http://sourceforge.net/projects/vmptagger*. The VMP tagger needs as input 4 lists: a lexicon, a lexical rule file, a context-rules file and a bigram list. The lists used in our system are from a POS tagger developed at the University of Minho [6], available at *http://natura.di.uminho.pt/download/sources/EMS/*.

Gazetteer. The original names in the lists were in English and represented no particular domain. Unfortunately, we have no access to a Portuguese electronic medical lexicon,

however, we translated some concepts and included biomedical terms (e.g. names of medical examinations) in order to reuse this processing resource.

Grammar Rules. GATE's IE system is rule-based and requires a developer to manually create rules, so it is not totally dynamic. The grammar rules developed are written in JAPE (Java Annotations Pattern Language)[7]. The rules do not just match instances from the Gazetteer with their occurrences in the text, but also find new instances in the text which do not exist in the Gazetteer, through use of contextual patterns, part-of-speech tags and other indicators.

4 Evaluation Methodology

Our evaluation focused, for now, on the identification and classification of atomic elements in text into predefined categories such as the proper names, names of diseases, time expressions, etc. in the text, that is, on Named Entity Recognition (NER).

The elaboration of the system resources was a constructive process: we first, manually, extracted a set of relevant expressions of the domain (e.g. name of diseases and examinations performed), and later these expressions were described in a grammar that was applied on a larger corpus.

The next sections present the corpus used to evaluate the system and the annotations used to classify the predefined categories.

***Corpus*.** The IE system for Portuguese medical reports was run on a part of the corpus mentioned before. All the text came from the Neurological database from the HGSA, and have been reported between 1992 and 2001. The full corpus is composed from more than 11 000 texts, containing more than 1 104 677 words. About 30 reports have been processed to elaborate the system. The evaluation was made on 200 new texts (about 2575 entities) from the corpus. These texts have not been used during the elaboration of the extraction patterns.

Annotations. The entities to be identified for this task include person names (in this case the names of the doctors responsible for the examination), time expressions, conditions of the patients, substances, numeric expressions and others that do not fit the previous categories.

The set of annotations defined to extract the desired informations is described in table 1. These annotations are based on the ones set for the Evaluation Contest of Named Entity Recognition Systems for Portuguese (HAREM) [8].

5 Results

Evaluation metrics mathematically define how to measure the system's performance against human-annotated gold standard. Traditional IE is evaluated in terms of **Precision** and **Recall** [9]. These are metrics used to measure the system performance in this paper.

Table 1. NER Annotation Set (in brackets the annotations in English).

Category	Type
PESSOA (PERSON)	*INDIVIDUAL* (INDIVIDUAL)
TEMPO (TIME)	*DATA* (DATE) *HORA* (HOUR)
ABSTRACCAO (ABSTRACTION)	*ESTADO* (STATUS)
COISA (THING)	*SUBSTANCIA* (SUBSTANCE)
VALOR (VALUE)	*CLASSIFICACAO* (CLASSIFICATION) *QUANTIDADE* (QUANTITY)
VARIADO (VARIED)	

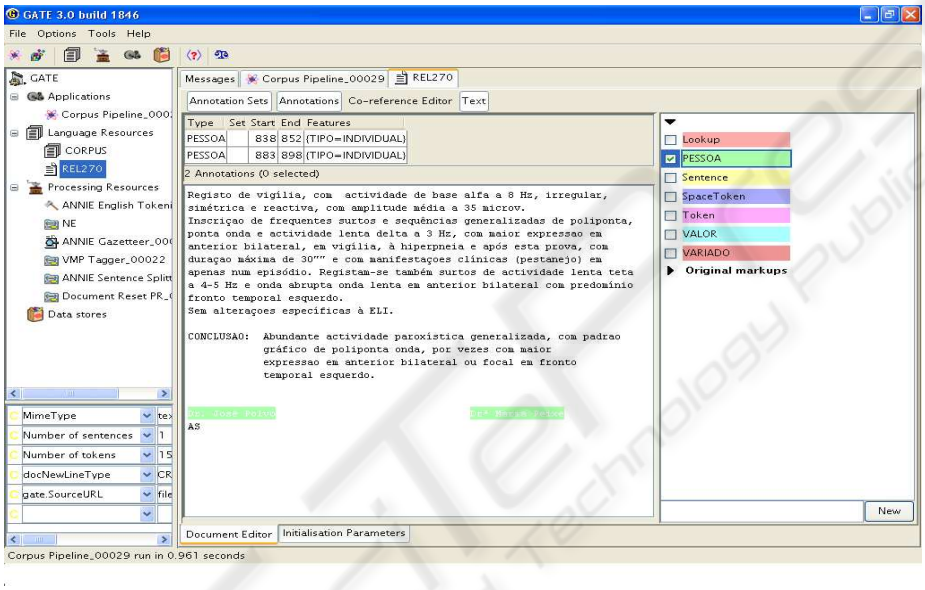

Fig. 1. Example of the result retrieved by the system.

Named Entity Recognition. An example of the result retrieved by the system can be analyzed in Figure 1 where is possible to see the annotation of the entity PESSOA (PERSON).

Named entity recognition results are summarized in Table 2. The lines show the number of entities correctly matched by the system, the ones partially correct, the number of entities the system was not able to identify and the ones that were falsely matched. Results in terms of recall, precision and F-measure are in the bottom lines of the table. Table 2 shows results for each annotation type, while Table 3 presents the overall results.

Template Element filling. Currently, whenever a scientist wants to find a report with a specific type of characteristics he has to perform several searches through large volumes of indexed text. The elaboration of a summary (template) for each report would greatly benefit this kind of searches.

Table 2. Evaluation Results for the several entities.

	ABSTRACCAO (ABSTRACTION)	*COISA* (THING)	*TEMPO* (TIME)	*PESSOA* (PERSON)	*VALOR* (VALUE)	*VARIADO* (VARIED)
Correct matches	332	86	211	278	1046	498
Partially Correct matches	1	0	10	9	12	0
Missing	5	2	14	8	2	3
False Positives	3	0	0	0	55	0
Recall	0,9822	0,9773	0,8979	0,9424	0,9868	0,9940
Precision	0,9881	1,0000	0,9548	0,9686	0,9398	1,0000
F - measure	0,9852	0,9885	0,9254	0,9553	0,9627	0,9970

Table 3. Overall Named Entity Recognition Results.

Correct Matches	2451
Partially Correct matches	32
Missing	34
False Positives	58
Recall	0,9738
Precision	0,9646
F - measure	0,9692

In this particular case scientists and health professionals have an ongoing interest in the type of wave and activity revealed by the patient's Electroencephalogram (EEG). Thus, and to demonstrate the interest of this type of approach, we have designed a template to capture the main information from the results produced by the system described above. The template filling was done through the manipulation of the XML document retrieved by the system with the help of a XSL stylesheet. The stylesheet determines how the information existing in the XML document, returned by GATE, should be presented.

The template definitions for this experiment include three Template Elements: activity (*ACTIVIDADE*), type of wave (*Tipo de ONDA*) and the doctor responsible for the examination (*Médico(a) Responsável*). An example of a result can be analyzed in Figure 2.

6 Information Extraction Used in Information Retrieval

Another experiment that demonstrates the interest of this type of approach is the use of the results obtained with IE techniques to perform Information Retrieval (IR).

Figure 3 presents a table with some of the results from this experiment. This table lists the reports in which is described some type of activity and resumes the type of activity described for each one of these (in terms of Type and Value). About 30 reports were used to perform this experience.

This type of approach allows the accomplishment of more complex searches, such as, for example, find the report that contains an activity with Type1 = '*LENTA*', Type2 = '*DELTA*' and Value >= '2 Hz'. In this case the result is the retrieval of report 270.

No formal evaluation of the IR results were performed yet.

Resumo Relatorio270

Tipo de ONDA

onda lenta

ponta onda

onda

onda abrupta

ACTIVIDADE

LENTA delta a 3 Hz

BASE alfa a 8 Hz

LENTA teta a 4 - 5 Hz

Médico(a) Responsável: Dr F A **Médico(a) Responsável:** Dr . A M

Summary Relatorio270

Type of WAVE

onda lenta

ponta onda

onda

onda abrupta

ACTIVITY

LENTA delta a 3 Hz

BASE alfa a 8 Hz

LENTA teta a 4 - 5 Hz

Doctor Responsible: Dr F A **Doctor Responsible:** Dr . A M

Fig. 2. Example Template: report 270 summary (original on the left, manually translated at right).

Reports summarized by ACTIVITY

Report	Type 1	Type 2	Value (Hz)
250	lenta	delta	--
250	lenta	deltateta	2
250	rápida	beta	--
251	fundo	alfa	10
253	base	--	--
253	lenta	--	--
253	fundo	alfa	9
254	fundo	alfa	10
254	rápida	beta	22
270	**lenta**	**delta**	**3**
270	lenta	teta	4-5
270	base	alfa	8

Fig. 3. Reports summarized by activity (in bold the result of the search described in the text).

7 Discussion

This paper presents the first steps given to develop an IE system intended to extract structured information from medical reports written in Portuguese. Our first evaluation focused on NER. This evaluation is described, the results are analyzed and some experiences that demonstrate the potential of this kind of techniques are presented. However, it should be noticed that the evaluation results are preliminary and we expect to improve with further development.

The results presented and the progress so far provides convincing grounds for believing that IE techniques will deliver effective ways for the extraction of information from unstructured text sources, in particular, in the medical domain.

Acknowledgments

The author would like to thank the Neurophysiology department of the HGSA, Porto, for the anonymized database access.

References

1. Pazienza M.T.: Information Extraction (a multidisciplinary approach to an emerging information technology). Lecture Notes in Computer Science.
2. Seventh Message Understanding Conference (MUC-7). Morgan Kaufmann Publishers, San Francisco, California, 1998
3. Gaizauskas, R., Humphreys, K., Demetriou, G.: Information Extraction from Biological Science Journal Articles: Enzyme Interactions and Protein Strctures. Chemical Data Analysis in the Large: The Challenge of the Automation Age, Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May, 2000, Bozen , Italy
4. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002
5. D. Maynard, H. Cunningham, K. Bontcheva, R. Catizone, G. Demetriou, R. Gaizauskas, O. Hamza, M. Hepple, P. Herring, B. Mitchell, M. Oakes, W. Peters, A. Setzer, M. Stevenson, V. Tablan, C. Ursu, Y. Wilks: A Survey of Uses of GATE. Technical Report CS–00–06, Department of Computer Science, University of Sheffield, 2000
6. R. Reis, J. Almeida: Etiquetador morfo-sintáctico para o Português. In Actas do XIII Encontro da Associaão Portuguesa de Linguística, Lisboa, Portugal, 1997, vol.2, pp. 209–222, Associacão Portuguesa de Linguística
7. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, and C. Ursu. 2002. The GATE User Guide. http://gate.ac.uk/.
8. URL: http://linguateca.di.fc.ul.pt/harem.php
9. D. Jurafsky, J. H. Martin. An Introduction to Natural Language Processing, Computacional Linguistics, and Speech Recognition, Upper Saddle River, New Jersey, 2000. Prentice Hall.