Information Extraction from Portuguese Hospital Discharge Letters

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

[1]Institute of Electronics and Telematics Engineering of Aveiro
Department of Electronics, Telecommunications and Informatics
University of Aveiro, Portugal
{lsferreira, ajst, jcunha}@ua.pt

Abstract

In this paper we describe MedAlert, a system which automatically extracts information from free-text discharge summaries written in Portuguese. We introduce a corpus of 915 hypertension related discharge letters and the method used to create the discharge letters representation model. MedAlert is based on an open-source framework and its components use natural language processing principles to discover elements of the knowledge model. We evaluate MedAlert precision using a set of 10 discharge letters from the MedAlert corpus from which 339 named entities were recognized. MedAlert achieves an entity recognition precision of 1 for entities such as anatomical sites, evolutions and dates and 0.93-0.99 for conditions, findings and therapeutics. A precision value of 0.69 is reported for examination entities due to to the recognition of active substances, such as insulin, as laboratory examinations.

Index Terms: information extraction, medical language processing, medical knowledge representation

1. Introduction

In order to perform research and to improve standards of health care it is required the access to a variety of data sources. The knowledge contained in unstructured textual documents like clinical notes and discharge summaries is critical to achieve these goals.

To bridge the gap between free-text and structured information, an automatic and highly accurate mapping of free-text reports onto a structured representation is required. Natural Language Processing (NLP) systems can retrieve named entities such as diseases and anatomical sites, and may be able to provide links between them in case of a relationship.

In this paper we describe the method used to create a structured representation of the discharge summaries and we describe our system, MedAlert, which automatically recognizes the entities mentioned in the reports. We also report on the performance of MedAlert, measured on a set of discharge summaries of patients admitted with hypertension related disorders.

This system has great clinical value: it can increase the classification of patients for practice management (*how many patients with cerebral hemorrhage were received*), for research (*how many patients used a specific of drug with success*), quality control (*how many patients with cerebral hemorrhage were received and which were their outcome*). In addiction, it would allow physicians to continue to practice using their current descriptive language in free-text reports without a requirement to enter structured data in a complex, time consuming computer-based system.

This paper is organized as follows: Section 1.1 presents some related work in the area of medical language processing systems and its resources. Section 2 discusses the terminologies used in the system. In particular, Section 2.1 provides details about the corpus and the manual annotation process while Section 2.2 presents the knowledge sources used. The architecture of MedALERT, the Medical Language Processing System, is the focus of Section 3. Section 4 presents the evaluation results of entity recognition from the set of free-text reports. We conclude in Section 5 with some future work directions.

1.1. Related work

The goal of information extraction (IE) is to extract structured and semantically well defined concepts from unstructured data sources to facilitate access and retrieval of information [1]. In the clinical domain, information extraction has the potential to help clinicians rapidly answer questions such as *How many patients were diagnosed in 2007 with cerebrovascular diseases?*, *What percentage of these patients had also hypertension?* There are multiple approaches to building IE systems. In general, such systems have NLP components such as tokenizers, part-of-speech taggers and parsers. Two separate frameworks for building information extraction systems were developed and made available as open-source components. One is the Generalized Architecture for Text Engineering (GATE) [3] and the other is the Unstructured Information Management Architecture (UIMA) [2]. Some of the state-of-the-art IE systems in the biomedical and clinical domain are presented in [4] and [5]. The caTIES system [6] extracts several types of named entities (NE) such as histology, anatomical site, size and grade and is based on the GATE framework. No results have yet been published. MedLEE, another clinical natural language processing IE system, extracts domain knowledge from a variety of unstructured reports, such as discharge summaries, radiology reports and pathology reports [7]. MedLEE focus on extracting named entities but there seems to be no results published for extracting information from pathology reports using MedLEE. MedTAS/P, Medical Text Analysis System/Pathology, is a system for the automatic conversion of unstructured pathology reports into a structured and codified knowledge source according to a Cancer Disease Knowledge Representation Model. It is part of a clinical NLP-based system as described in [1]. MedTAS/P achieves F1-scores of 0.97-1.0 for instantiating classes in the knowledge representation model such as histologies or anatomical sites. To the best of our knowledge there as been no effort to develop a system for the extraction of information from clinical reports written in Portuguese and in Portuguese hospitals.

2. Resources

2.1. MedAlert corpus

The biomedical community hasn't yet, to our knowledge, developed a gold-standard training and test corpus of annotated clinical reports written in English which can be used as a shared standard for evaluating automatic knowledge extraction system. The same is valid when referring to Portuguese reports. Therefore, the development of detailed manually annotated corpus for training, validating and evaluating the system is of extreme importance.

For the training, validation and evaluation of the MedAlert system we created a corpus of 915 free text discharge letters written in Portuguese from the Infante D. Pedro Hospital in Aveiro, Portugal. These discharge letters refer to patients admitted with hypertension related problems.

These discharge summaries contain the information added by the clinician during the period the patient was admitted in the hospital. They are divided into 6 sections, each containing information about the patient admission motive, clinical history, physical examination, evolution, therapeutics applied during admission and destiny recommended to the patient after discharge, respectively. The documents do not have any personal data and therefore it is not possible to identify the patient through the analysis of the discharge letters. Table 1 shows the distribution of information in the corpus, in particular the amount of texts, sentences and tokens for each structure.

Table 1: Corpus MedAlert

	Tokens	Sentences	Texts
Admission motive	1 989	225	162
Clinical history	22 386	1 066	185
Physical Examination	7 865	711	134
Evolution	8 998	506	154
Therapeutic	6 299	219	159
Destiny	4 158	262	120
Total	51 695	2 989	914

Given the expense of human annotation, the gold standard set has to be a relatively small subset of the whole corpus of 915 documents. Therefore, to create the gold standard corpus we used a random sample of approximately 10% of the corpus, making a total of 90 documents.

In order to ensure consistency, critical to the quality of the gold standard, it is important that all documents are annotated to the same standard. This is accomplished by a set of guidelines describing in detail what should and should not be annotated and other important considerations such as how to decide if two entities are related and how to deal with co-reference. This guidelines also provide a sequence of steps, a recipe, which annotators should follow when working on a document. The guidelines were developed through a rigorous iterative process by a small team of computational linguistics and clinicians. They were tested against a significant number of documents before the use on the final gold standard.

The resulting representation model, the MedAlert Discharge Letters Representation Model (MDLRM) was implemented within Knowtator [8], a Protégé [9] plugin. The gold standard set is currently being manually annotated by a linguistic and a clinician. The annotators manually fill in the attributes and relations in the classes of the representation model with information from the reports using the Knowtator tool.

Table 2 presents the entities defined in the MDLRM which were used in the entity recognition task presented in this paper.

Table 2: MedAlert Entities

Classes	Description
Condition	Complications, conditions and other problems manifested by a patient;
Anatomical Site	Anatomical structure or location, normally the locus of a *Condition*;
Evolution	The clinical evolution of the patient or *Condition* after a given *Therapeutics*;
Examination	Interaction between doctor and patient or *Anatomical Site* with the purpose of measuring or studying some aspect of a *Condition*;
Finding	The numeric or qualitative finding of an *Examination*, excluding *Condition*;
Location	Geographically defined location, normally where an *Examination* or *Therapeutic* is performed;
Therapeutic	Action performed by a clinician targeted at a patient, *Anatomical Site* or *Condition* with the purpose of changing or treating a *Condition*;
DateTime	Temporal expressions, including dates and times (absolute or relative), duration and frequencies;
Value	Absolute or relative quantifications or classifications;

2.1.1. Development and evaluation sets

For the purpose of development we used the set of documents not used in the manual annotation task, ie, the 825 documents of the corpus which do not belong to the gold standard set were used to develop the system. The entity recognition task described in this paper is evaluated on regards to precision against a randomly selected set of 10 discharge letters belonging to the set gold standard set.

2.2. MedAlert ontologies

In this project we used two codified terminologies and a drugs ontology as the underlying terminologies for *Conditions*, *Examinations*, *Anatomical Sites* and *Therapeutics* entities recognition. Namely, we used the International Classification of Diseases - Ninth Revision, Clinical Modification (ICD-9-CM) and the Unified Medical Language System (UMLS) [10] as codified terminologies. The drugs ontology was created in a semi-automatic way using the information provided by INFARMED, the National Authority of Medicines and Health Products and by the Hospital regarding the medication used and available in the institution. It has a total of 5 228 instances of Medication and Active Substances (and the corresponding relationships between these).

These terminologies, particularly the UMLS semantic network, allows the use of specified vocabularies in entity recognition avoiding the manual creation of lists and gazetteers. This resource and its use in MedAlert is described in more detail in Section 2.2.1. The ICD9 terminology is intended to be used in a future task of automatically assigning codes to the health conditions described in the discharge summaries.

2.2.1. UMLS

The UMLS [10] is a compendium of many controlled vocabularies in biomedical sciences. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems. It may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts. UMLS main purpose is to facilitate the development of computer systems that behave as if they 'understand' the meaning of the language of biomedicine and health.

In order to develop the system presented in this paper we used the 2009 UMLS version and particularly the Portuguese translation of MeSH (the National Library of Medicine's controlled vocabulary thesaurus, a sets of terms naming descriptors in a hierarchical structure that allows search at various levels of specificity.), the DeCS.

3. System Architecture

MedAlert, the medical language processing system, is based on natural language processing principles and contains both rule-based and machine-learning based components and runs within UIMA framework. An application within such framework consists of a set of programs (annotators), each having a configuration file in XML format being the execution sequence, or pipeline, of annotators also described in a configuration file. Annotators mark up an unstructured textual document, inserting 'annotations' that can be associated with a particular piece of text or which can contain objects for other annotations. A subsequent annotator can read and process all previously created annotations. MedALERT, provides a mechanism to use external resources, such as terminologies and ontologies.

The system pipeline can be broken into several components:

1. Ingestion - a component which reads the patient discharge letters XML files and converts them into plain text while keeping information about the document's structure. Also this component is responsible for reading the knowledge sources and converting them into the MedAlert type system.
2. General natural language processing - component for sentence discovery, tokenization, part-of-speech tagging and shallow parsing. This component contains an abbreviation sensitive sentence splitter.
3. Named entity recognition - this component identifies the concepts defined in the MDLRM based on specified terminology. It also determines negation, lateralization and other modifiers of an entity.

3.1. Ingestion

The document ingestion annotator converts embedded tags of an input document or set of documents into annotations and simultaneously adds information as the sections of the document and header information containing the episode number and the codes assign to the episode. The terminologies used are also read in this component and added to UIMA type systems in order to be available in the annotation as the documents are analyzed.

3.2. General Natural Language processing

This component starts by identifying the abbreviations contained in the document. This step is critical for a correct determination of sentence boundaries as for abbreviations containing a period, the process of identifying them involves solving the sentence boundary problem as well (e.g. Dr. should not be considered the end of sentence). We developed a method for detecting abbreviations in clinical notes which is a heuristic rule-based program developed by observing several discharge letters. It utilizes information concerning word formation, such as capital letters, numeric and alphabetic characters and their combinations, together with the a list of Portuguese words created from a general corpus of Portuguese, the corpus developed to be used in the second evaluation contest for named entity recognition in Portuguese, the HAREM II [13] (with 784 119 words). In particular, if a word contains special characters such as "-" and ".", has less than 6 characters, is lower case and is not in the Portuguese list it is considered an abbreviation.

Part of this component are also a tokenizer and a part-of-speech tagger based on the popular Tree Tagger [11]. The component also contains a context tokenizer which is a regular expression annotator that in conjunction with short lists discovers textual mentions describing dimensions and sizes, dates, number, etc.

3.3. Named entity recognition

The named entity recognition (NER) component is one of the most important components of the system. This component makes use of the terminologies to recognize the entities belonging to the knowledge model. As exact string match is not sufficient to recognize entities in unstructured text, the NER annotators have several characteristics that help the specification of the tokens used for lookup. The possibility of ignoring case, skipping terms for lookup if they appear in a stop word list, use the context of the lookup string (sentence, paragraph) and word order independence are the main features of the component.

Part of the NER component are also the negation and lateralization annotators. The first is a generalized algorithm, based on the popular NegEx [12]. Negation trigger words such as *'sem' (without), 'nunca' (never), 'não' (not))* are specified in a user modifiable dictionary. The trigger words become the anchors around which negated sentences are discovered. When such word is found, the following semantic entities in the sentence are marked as negated. The lateralization annotator uses also trigger words like *'esquerda' (left), 'superior' (superior)* to characterize entities of anatomical site.

4. Results

In this section we report on the precision of MedAlert annotations. As the gold standard set is not yet fully annotated and verified we evaluate the system by analysing its output and determining if it corresponds to true or false positives. This allowed us to evaluate MedAlert's precision, which we consider to be the most important characteristic of systems intended to be used in clinical domains.

Table 3 shows the precision for the recognized entities. The system achieved a precision of 1 for the entities of anatomical site, evolution, location, datetime and value. The first three of these classes are also the ones with less representation on the discharge letters. The dateTime and value entities are mainly recognized with help of the context tokenizer which is based on regular expressions.

The entities of therapeutic were recognized with 0.99 precision. This result is a consequence of the use of the drugs ontology which contains all the active substances and medication used in Portugal. The system did not perform with 100% precision due to the use of the same acronym when refering to *Ácido Clavulânico (Clavulanic Acid)*(AC) and *Auscultação Cardíaca (Cardiac Auscultation)*(AC) in one of the discharge letters.

The entities of finding and condition are recognized with a precision of 0.93. Most of the false positives are originated because of the presence of anatomical site entities near keywords such as *mal* (this word can be interpreted as 'sickness' but is mostly meant as 'poorly' as, for example, in *'poorly controlled Diabetes'*).

The precision for the recognition of examination entities is 0.69, which is significantly lower than for the other entities. The recognition of some active substances present in the discharge letters as laboratory examinations is the main reason for this lower precision value. One example of is Insulin, recognized by the system as an hormone (and consequently as a laboratory examination) and as a therapeutic.

Table 3: MedAlert Results

Entities	True Positives	False Positives	Precision
Condition	75	6	0.93
Anatomical Site	4	0	1
Evolution	7	0	1
Examination	24	11	0.69
Finding	28	2	0.93
Location	1	0	1
Therapeutic	146	1	0.99
DateTime	55	0	1
Value	59	0	1
TOTAL	**399**	**20**	**0.95**

5. Conclusions

In this paper we describe an information extraction system MedAlert which automatically extracts information from free text discharge summaries written in Portuguese. We describe the MedAlert corpus of 915 hypertension related discharge letters and the method used to create the discharge letters representation model. Based on this model, detailed annotation guidelines were created and a gold standard set of approximately 10% of the corpus is currently being manually annotated by a linguistic and a clinician. The system runs within the UIMA framework and contains a set of components capable of using external resources, as ontologies and terminologies, to extract information. The precision of the system was evaluated with values ranging from 0.93-1 except for the entities of the class examination. This is mainly due to the recognition of active substances such such as insulin as laboratory examinations.

5.1. Future work

In order to improve the entity recognition results it is essential the use more context information, namely the information concerning the section in which the entity is referred. Another task of great interest in the medical domain is the automatic code assignment to the discharge letters, which is currently being developed. The main objective of the MedAlert system is to act as a medical decision support system capable of inferring doubts/irregularities in the decisions made by the health professionals. The development of components capable of interact between the structured representation of the discharge letters and the information extraction task are essential to reach this objective.

6. Acknowledgements

Our thanks to César Telmo Oliveira from Hospital Infante D. Pedro, Aveiro and Catarina Oliveira, University of Aveiro for helping developing the knowledge model and gold standard set. Our thanks also to Science and Technology Foundation (PhD grant SFRH/BD/27301/2006).

7. References

[1] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, P. C. de Groen, "Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model", Journal of Biomedical Inform, 42(5): 937-949 (2009).

[2] D. Ferrucci, A. Lally, "UIMA an architectural approach to unstructured information processing in the corporate research environment", in Natural Language Engineering (2004), 10:327-348.

[3] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications", in 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Lisbon, July 2002.

[4] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research", in Yearbook of Medical Informatics, 128-144, 2008.

[5] S. Ananiadou, J. Mcnaught, "Text Mining for Biology And Biomedicine", Artech House, Inc., 2005.

[6] <http://caties.cabig.upmc.edu/>

[7] C. Friedman, S. B. Johnson, B. Forman, J. Starren, "Architectural requirements for a multipurpose natural language processor in the clinical environment", in Proceedings Annual Symposium on Computer Applications in Medical Care, 347-351, 1995.

[8] Philip Ogren, "knowtator: A plug-in for creating training and evaluation data sets for Biomedical Natural Language systems", in Proceedings of the 9th International Protégé Conference, 73–76, Stanford, California, 2006.

[9] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubzy, H. Eriksson, N. F. Noy, Samson W. Tu, "The Evolution of Protégé: An Environment for Knowledge-Based Systems Development, in International Journal of Human-Computer Studies, 58:89-123, 2002.

[10] UMLS Knowledge Sources, United Stated National Library of Medicine, 2008

[11] H. Schmid, "Tree Tagger, a language independent part-of-speech tagger", Institut fur Maschinelle Sprachverarbeitung, Stuttgart University, 1995.

[12] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries", in Journal of Biomedical Informatics, 34:301-310, 2001.

[13] "Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM", Cristina Mota and Diana Santos (eds), Linguateca, 2008 http://www.linguateca.pt/LivroSegundoHAREM/, ISBN: 978-989-20-1656-6.