Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Unsupervised entity and relation extraction from clinical records in Italian

Anita Alicante*, Anna Corazza, Francesco Isgrò, Stefano Silvestri

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, DIETI, Università degli Studi di Napoli Federico II, Via Claudio, 21 - 80125 Napoli, Italy

ARTICLE INFO

Article history:
Received 21 January 2015
Accepted 16 January 2016

Keywords:
Unsupervised learning
Relation clustering
Entity extraction
Medical information extraction
Entities relation discovery

ABSTRACT

This paper proposes and discusses the use of text mining techniques for the extraction of information from clinical records written in Italian. However, as it is very difficult and expensive to obtain annotated material for languages different from English, we only consider unsupervised approaches, where no annotated training set is necessary. We therefore propose a complete system that is structured in two steps. In the first one domain entities are extracted from the clinical records by means of a metathesaurus and standard natural language processing tools. The second step attempts to discover relations between the entity pairs extracted from the whole set of clinical records. For this last step we investigate the performance of unsupervised methods such as clustering in the space of entity pairs, represented by an *ad hoc* feature vector. The resulting clusters are then automatically labelled by using the most significant features. The system has been tested on a fairly large data set of clinical records in Italian, investigating the variation in the performance adopting different similarity measures in the feature space. The results of our experiments show that the unsupervised approach proposed is promising and well suited for a semi-automatic labelling of the extracted relations.

1. Introduction

Medical information processing systems are of paramount importance for patient care. Indeed, delivering the most appropriate piece of information in every moment and in the correct situation can represent a crucial support for clinical activities. Although the digitization of all medical documents in hospital information systems is now nearly completed, several problems still need to be solved for their effective and reliable automatic processing. A number of them, spanning from semantics-based indexing of documents for improved retrieval to more advanced query based information extraction, and to the application of ontology-based strategies for privacy protection, would take great advantage from improved approaches to extract relevant information from technical texts. Such information usually consists of relevant entities and relations connecting them [1]. Let us, for example, consider the following sentence:

Found subcapital fracture and dislocation
of left shoulder andcontusion of right
hip caused by accidental fall at home.

We can identify the following five medical entities:

Found subcapital < ent1 > fracture < /ent1 >
and < ent2 > dislocation < /ent2 > of left
< ent3 > shoulder < /ent3 > and < ent4 >
contusion < /ent4 > of right < ent5 > hip
< /ent5 > caused by accidental fall at home.

There should be also three relations respectively connecting entities 4 and 5, 1 and 3, and 2 and 3; their label should refer to an illness of a body part.

Progresses in the field mainly regard general domains and produced reliable entity recognition systems including, for example, TextPro [2]. However, the availability in restricted domains of specific resources, such as dictionaries and ontologies, is known to represent an important opportunity for a burst in performance. This is surely the case for scientific documentation, including not only scientific papers, but also medical reports. In fact, the medical decision process is founded on both the availability of up-to-date results from the scientific community, and on the complete information about the current case, now available in digital form, then representing a not to be missed opportunity.

In this paper we propose a system to solve this problem, and describe it in detail, together with the adopted knowledge sources, including lexical resources and natural language processing tools. The system improves the one previously reported in [3], as here

* Corresponding author. Tel.: +39 0 81679267.
E-mail address: anita.alicante@unina.it (A. Alicante).

http://dx.doi.org/10.1016/j.compbiomed.2016.01.014
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

the general domain entity recognition step has been eliminated, while a final cluster labelling has been added. The final conclusion drawn in the preceding publication, that is the choice of the cosine similarity as the more effective similarity measure has been confirmed by the more extended experiments reported here. As a consequence, we choose a slightly different clustering algorithm, the *spherical K-means*, which is based on such measure. Further experiments have then been performed with this refined system on two larger data sets.

The paper is structured as follows. The next section is devoted to a deeper discussion of motivation of this work, while the following one considers the state of the art and Section 4 describes the information sources adopted by the system. Section 5 describes the system architecture focusing on entity recognition and relation clustering. The experimental assessment is discussed in Section 6, while Section 7 analyses the solution obtained both from a quantitative point of view and by considering the resulting cluster labelling. A final discussion concludes the paper in Section 8.

2. Problem statement

Given the availability of digitalized documents, it would be valuable to physicians and researchers being able to retrieve the clinical records for similar past cases, or the immediate availability of a different statistical analysis (e.g., correlation of breast cancer cases with the age). Rather than with the traditional keyword-based search, these tasks would be more effectively attained if a more structured search was possible, as for example looking for all cases involving two given entities involved in a certain relation. To achieve this goal, it is necessary to annotate the input text with entities and relations.

Such situation introduces a number of technical challenges. While scientific publications are written in English to favour international opinion exchanges, patient records are usually written in the hospital country language. In our case, we consider Italian, for the processing of which, as for most languages different from English, no rich literature is available. Another problem to take into account is the occurrence, in clinical records, of typos and not standard abbreviations, in addition to the most usual acronyms. Last but not least, moving from text to knowledge processing raises tricky privacy problems. In fact, especially but not only in small hospitals, obscuring the patient name is not sufficient to hide his/her identity as a precise profiling of the patient can often be obtained from a few of the medical characteristics reported in a record.

Ad hoc solutions are needed to tackle such problems. Beginning from the last one, more effective privacy protection can be based on ontological information [4]. However, the construction and population of the necessary ontologies require the identification of relevant information from medical reports. Furthermore, more in general, the identification of potentially dangerous information is again based on the extraction of domain entities and relations.

A great deal of effort to ease the porting of systems to languages different from English has been put in the development of lexical resources, which are now available also for Italian. However, even if such resources are valuable tools for the porting, they are not enough, because of the intrinsic linguistic differences between languages. It is therefore necessary to take into account the characteristics of the specific language in both entity and relation recognition. However, in this second step they are even more important, as relations strongly depend on the sentence structure, obviously language dependent.

Such problems are usually tackled adopting machine learning approaches, which try to extract information directly from data. Whenever *annotated* data are available, supervised strategies outperform unsupervised ones among machine learning approaches, at the price of an expensive annotation phase. In fact, domain experts are required to invest their time in a long and tedious annotation activity. In our case, this would imply persuading medical staff to invest part of their precious time to annotate data with information about the presence and the type of domain relevant entities and relations in records to be used for training. While this would be really difficult, things are much different if their competence is only required to check on an annotation which has been automatically produced. By following this idea, we propose a framework which integrates a knowledge-based and a text mining approaches: the expert intervention will be finally required to check on natural language labels associated to groups of relations.

The framework is composed by three phases. The first one is devoted to domain entity (i.e., medical and pharmaceutical entities) identification and classification, and exploits domain related lexical resources and standard natural language tools. The second one is based on an unsupervised machine learning approach, *clustering*, to avoid the necessity of annotating data. A potential relation is hypothesized among all pairs of the entities identified in the preceding phase. Clustering is then applied to group similar entity pairs. Small clusters indicate the lack of repetitive patterns and will therefore be considered as entity pairs which are not in relation to each other, while larger clusters will correspond to different relation types.

In order to compensate for the lack of annotation, we try and put as much knowledge as possible in the input representation. In particular, together with *n*-grams of words, we also include *barrier features*, an innovative type of features recently introduced for information extraction [5]. While such features are only based on standard natural language processing and do not require any manual annotation, they capture the context in which entity tokens appear in an efficient and effective way. The third and final phase of the proposed framework aims at associating a label to each cluster. It is based on the analysis of the *n*-grams of words which represent the only lexicalized features adopted in the clustering, and associates to each cluster the terms which best describe it.

3. Related work

Given the growing digitization of medical information processing around the world, several systems for information retrieval and extraction have been proposed, including [1,3,6–10]. Many of them are based on the extraction of entities and relations from free text parts of medical records. In the majority of these systems, entities and relations are considered in two separate steps. Some systems, such as [11,12], only consider general domain entities, while others [6–8,13–16] base the extraction of entities on domain specific external resources. In particular, most use the Unified Medical Language System (UMLS) [17]. Among these, [7] proposes a semantic framework for performing information retrieval in a collection composed by biomedical-chemistry patents and full-text articles. Queries are refined by means of biomedical entities previously extracted and tagged by using the UMLS metathesaurus.

A large part of the proposed approaches are based on machine learning. As an example we mention [8], where Conditional Random Fields (CRFs) are applied for identifying disease mentions; adopted features regard disease-specific contexts, word spelling, general linguistic characteristics, syntactic dependencies and dictionary entries. Experiments are based on the (English) Arizona Disease Corpus (AZDC) [18], which contains detailed annotations of diseases, including information taken from UMLS.

BANNER is another machine-learning system, also based on CRFs, described in [19], which also contains a wide survey of the features adopted for named entity recognition in biomedical texts.