Relation mining from clinical records

Anita Alicante, Anna Corazza, Francesco Isgrò
Department of Electrical Engineering and Information Technologies (DIETI)
Università di Napoli Federico II
via Claudio 21, 80125 Napoli, Italy
{anita.alicante|anna.corazza|francesco.isgro}@unina.it

Stefano Silvestri
Institute for High Performance Computing and Networking, ICAR-CNR
via P. Castellino, 111, 80131 Napoli, Italy
stefano.silvestri@icar.cnr.it

Abstract

English. We propose a system to extract entities and relations from a set of clinical records in Italian based on two preceding works (Alicante et al., 2016b) and (Alicante et al., 2016a). This approach does not require annotated data and is based on existing domain lexical resources and unsupervised machine learning techniques.

Italiano. *Proponiamo un sistema per estrarre entità e relazioni da un insieme di cartelle cliniche in Italiano basato su due precedenti lavori (Alicante et al., 2016b) e (Alicante et al., 2016a). Questo approccio non richiede dati annotati e si basa su risorse lessicali di dominio già esistenti e tecniche di apprendimento automatico senza supervisione.*

1 Introduction

The digitization of medical documents in hospitals has produced plenty of information which should be adequately organized. While part of the material, mainly including international scientific publications, is in English, increasingly more material is being created in the language of the country of the medical institution. The main part of the local language material is represented by patient records. They contain important information not only for preparing care plans or solve problems for the particular patient, but also to extract statistics useful for research and also for logistics administration.

Automatic processing of such repositories still can not be straightforwardly applied. One of the principal issues to be solved is the automatic extraction of relevant information, usually consisting in entities and relations connecting them (Alicante et al., 2016b). In the cited work, we extensively discuss a domain entity and relation recognition system for Italian. Such step is at the basis of more sophisticated analyses, including semantics-based indexing of documents for improved retrieval, advanced query based information extraction, and the application of ontology-based strategies for privacy protection.

General tools, such as TextPro (Pianta et al., 2008), are not adapted for technical domains such as the medical one, as they are trained on generic documents, rather than domain-specific ones. Furthermore, a lot of tools are available for English and only a few of them have been ported to Italian. Another problem to take into account is the occurrence, in clinical records, of typos and nonstandard abbreviations, in addition to the most usual acronyms. Last but not least, passing from text to knowledge processing raises tricky privacy problems. In fact, especially but not only in small hospitals, obscuring the patient names is not sufficient to hide their identity as the medical information reported in records are often sufficient to reconstruct a precise profiling of the patients.

Therefore, *ad hoc* solutions represent the only way to build effective applications to solve this kind of problems. For example, not only domain entities and relations can help identifying potentially dangerous information, but also ontological information can be exploited to better protect patient privacy (Bonatti and Sauro, 2013). Again, ontologies construction and population are based on entity and relation extraction.

Efforts to port systems to languages different from English require, first of all, the development of lexical resources for the considered language. However, they are not sufficient, because of the intrinsic differences between languages. A widely adopted way to tackle such difficulties is represented by machine learning approaches.

Although supervised approaches are usually more effective, they require large corpora of annotated data, which are quite expensive to obtain, as they require that domain experts invest time in a long and tedious annotation activity. In the medical domain, staff should 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 the training phase. Things would be much easier if domain experts are only required to check an automatically produced annotation. We therefore propose to integrate a knowledge-based and a text mining approaches to develop an application which requires the expert intervention only to check on medical and pharmaceutical labels associated to groups of relations.

More in detail, we propose here to integrate the systems discussed in (Alicante et al., 2016b) and in (Alicante et al., 2016a): the former adopts domain dependent lexical resources to extract entities and unsupervised machine learning approaches to decide where relations occur in the text. The latter clusters and labels the extracted relations with an approach based on lexical semantics.

The paper is organized with Section 2 detailing the approach implementation and Section 3 for conclusions and future works.

2 Proposed approach

The framework proposed is composed by three modules, and its logical structure is depicted in Figure 1. 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, namely *clustering*, to avoid the necessity of annotating data, for the relation extraction. 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 are likely to correspond to different relation types.

Relations are clustered and labeled using the approach proposed in (Alicante et al., 2016a). The decision about how a relation can be labeled is only based on the terms involved in the corresponding entity pair, without considering the context in which it occurs. In fact, this is complementary with respect to the task of deciding whether two entities are related, which should be decided on the basis of the context where the two entities occur, as in (Alicante et al., 2016b). On the other hand, by considering only the two involved entities, we can only decide the *type* of a relation. Then, to decide whether the relation is stated or negated, also the context should be considered in the analysis.

The third module of the framework is based on Word Embeddings (WEs) (Mikolov et al., 2013) to represent the words involved in each entity with a real valued array. WEs most interesting characteristic consists in the fact that the mutual position of words in a metric space strongly depends on their meanings, so that words having similar semantics have large similarity, when this is computed, for example, by cosine similarity. Embeddings can be automatically built from a large collection of unannotated text with a very efficient algorithm. Therefore, they can be easily applied to any language, in our case to Italian, provided that enough texts are available. We used documents extracted from Wikipedia for training. In particular, we considered pages flagged as Medicine, Biology and Pharmacy in Italian. For the extraction, we used CatScan v3.0[1], Wikipedia Export tool[2] and Wikiextractor[3].

For each entity, we then consider the embeddings corresponding to each token. As shown in (Paperno and Baroni, 2016), a good representation for a string of words is given by the sum of the corresponding WEs. However, as we do not want that such representation depends on the string length, we normalize the sum by the number of words involved in the entity, obtaining the average or centroid of the corresponding WEs. Each pair of entities occurring in the same sentence represents a possible candidate for a relation. We therefore build the feature vector for each entity pair by juxtaposing the average vectors for each

[1] https://tools.wmflabs.org/catscan2/catscan2.php

[2] https://en.wikipedia.org/wiki/Special:Export

[3] medialab.di.unipi.it/Project/SemaWiki/Tools/WikiExtractor.py

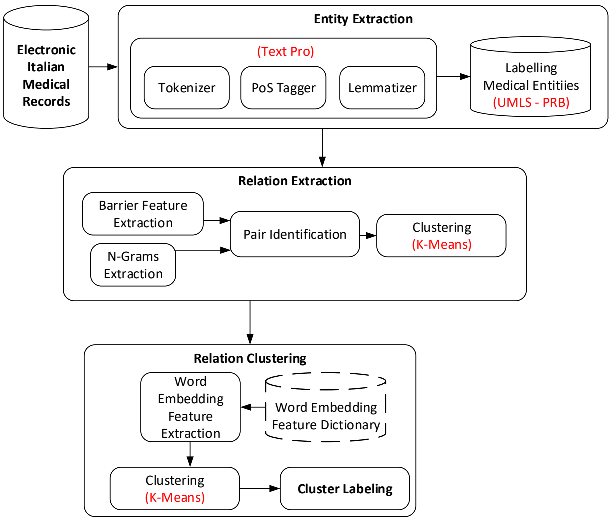

Figure 1: Architecture for Relation mining from clinical records.

entity and input this representation into a k-means clustering (Manning et al., 2008; Shalev-Shwartz and Ben-David, 2014).

2.1 Input Preprocessing

The text, processed by our system, is extracted from anonymized medical records, in the form of plain text encoded in UTF-8. The text includes a small set of special characters, used as delimiters and/or formatters. The largest part of these medical records has been produced by an HL7-compatible information system. At the end of each medical record, there is often an ICD9M (International Standard for Encoding and Classifying Diseases) disease code, which we disregard together with the rest of the structured part of the records.

The text is initially preprocessed for extracting textual parts from the medical records, and to get rid of non-textual characters. The plain text, produced by this preprocessing step, is passed to the natural language processing suite TextPro to perform tokenization, sentence splitting, PoS tagging and lemmatization.

2.2 Entity Extraction

Entity extraction is crucial for our analysis, and a specific module has been implemented with the goal of extracting entities which are relevant for the application domain: biomedical and pharmaceutical entities in our case. The module follows a pattern matching approach by identifying each occurrence of a number of PoS patterns in the input text as a candidate to be further analysed.

Afterwards, for each token occurring in the identified pattern, we search for matches of the corresponding lemma in the dictionaries. In case of multi-word expressions, when several patterns apply to overlapping strings of tokens, we apply a greedy approach by choosing the longest one matching the input.

The output is produced following the TextPro format, that is a line for each token, and a column for each analysis level. In our system these files are enriched by the information about Medical and Pharmaceutical entities obtained from the dictionaries provided by UMLS[4] and PRB[5]. These information are labeled as MED for the medical entities, and FAR for the pharmaceutical ones (the whole entity tag list is shown in the Table 1).

Table 1: List of medical sub-categories

Description	Label
Medical	MED
Pharmaceutical	FAR
Anatomy	ANA
Organisms	ORG
Diseases	MAL
Chemicals and Drugs	CHE
Technical medical equipment	TEC
Psychology and Psychiatric	PSI
Biology	BIO
Natural Sciences	NAT
Anthropology and Social Science	SOC
Technology, Industry and Agriculture	IND
Humanities	UMA
Computer Science	INF
Groups of People	GRU
Health care	ASS
Characteristics of Publication	PUB
Locations	LOC

In addition to a label indicating whether the entity is medical (MED) or pharmaceutical (FAR), we also add to each medical entity annotation the sub-categories included in the UMLS database in correspondence to the dictionary entry. The list of sub-categories labels are summarized in Table 1. A side-effect of such sub-categorization is that the number of potential relations increases while it becomes possible to find more specific relations.

[4] Unified Medical Language System, `http://www.nlm.nih.gov/research/umls`

[5] Pharmaceutical Reference Book, officially mantained by Agenzia Italiana del Farmaco

2.3 Relation Clustering

We apply the *k-means* approach that identifies groups of relations of the same type appearing in the data set. Each pair of entities occurring in the same sentence identifies a potential relation, therefore all possible entity pairs must be considered. We then apply a clustering algorithm to the set of all the potential relations identified. We will disregard all entity pairs belonging to clusters having a size smaller than a given threshold.

We then concentrate on the remaining entity pairs, which are likely to represent actual relations and semantically cluster them. The approach proposed for this is structured in three main modules: *Feature Construction*, *Clustering*, and *Cluster Labeling*. The first module builds a feature vector based on WEs for each relation candidate; for doing this, first it constructs a WE dictionary by using a large collection of unannotated texts, in our case extracted from Wikipedia. This module is based on *word2vec*[6] (Mikolov et al., 2013). For the feature vectors length we chose 500, which is the default choice, and set the minimum word count to 3, to exclude the less frequent words from the dictionary, obtaining a set of $260,680$ vectors.

After that, the k-means clustering is applied to the set of feature vectors obtained by the first module. For every entity pair we then construct a Feature Vector (FV) starting from the WE of each word involved. Each entity can be composed by one or more words, as for example *conati di vomito*: in this case, for each entity, we take the average among the WEs of the words composing the entity associated to the entity pair. Finally, we concatenate the FVs of the two entities, obtaining a FV of $1,000$ entries.

The clustering algorithm is then applied to the FV data set by means of the *C Clustering library* (de Hoon et al., 2004), a fast C implementation of the k-means algorithm. As the k-means is characterized by a random initial choice of the seeds, we repeated each run 10 times, always choosing the best solution. We considered the cosine similarity, choosing a number of clusters equal to 40, which seemed a reasonable choice given the results from the experiments in (Alicante et al., 2016b) and in (Alicante et al., 2016a).

Eventually, to label each cluster we ordered the pairs in each cluster according to its cosine similarity from the cluster centroid: the first four pairs are then chosen to characterize the cluster.

As discussed above, each FV can be partitioned in two parts: the first half corresponds to the first entity in the pair, the second one to the other. Such partition is consistently maintained during the whole processing. Also in the computation of centroids in the k-means clustering algorithm, the former half of each centroid derives from the average of the former half of the involved FVs and then corresponds to the first entity. Correspondingly, the latter half of each centroid vector only depends on the second entity of each involved pair.

The choice of the cluster to which a given item is assigned is based on the cosine similarity. Its computation can be divided in three parts: the dot product of the part of the two FVs corresponding to the first entity, the same for the second entity and eventually the normalization with respect to the whole FV. Therefore, the evaluation of the cosine similarity is based on a trade-off between how similar are the first and the second entities in each pair. In other words, they represent actual entities pairs which are similar to the (abstract) cluster representative, corresponding to the centroid.

3 Conclusions and future work

In this paper we presented a system for the extraction of information from clinical records in Italian. A first part of the system aims to extract domain relevant entities from medical reports by a pattern matching approach. A second part takes the output of the former step and applies a clustering approach to explore possible relations between such entities. A third part is based on WE and aims to give cues about the type of the relations.

Interestingly, the approach does not require annotated data, but only easily available data such as Wikipedia and off-the-shelf tools in addition to the documents to process. Naturally, available tools have been trained on annotated data, but without any adaptation to the specific domain. It would therefore be interesting to port it to a new language, possibly different from English, which represents the most widely studied among all languages.

Acknowledgments

The research presented in this paper was partially supported by the national projects CHIS - Cultural Heritage Information System (PON), and BIG4H

[6] The software is freely available at `https://code.google.com/p/word2vec/`

- Big Data Analytics for E-Health Applications (POR).

References

Anita Alicante, Anna Corazza, Francesco Isgrò, and Stefano Silvestri. 2016a. Semantic cluster labeling for medical relations. In *Proceeding of Innovation in Medicine and Healthcare 2016*, pages 183–193, Puerto de la Cruz, Tenerife, Spain. Springer.

Anita Alicante, Anna Corazza, Francesco Isgrò, and Stefano Silvestri. 2016b. Unsupervised entity and relation extraction from clinical records in Italian. *Computers in Biology and Medicine*, 72:263–275.

Piero A. Bonatti and Luigi Sauro. 2013. A confidentiality model for ontologies. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul T. Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors, *International Semantic Web Conference (1)*, volume 8218 of *Lecture Notes in Computer Science*, pages 17–32. Springer.

Michiel J.L. de Hoon, Seiya Imoto, John Nolan, and Satoru Miyano. 2004. Open source clustering software. *Bioinformatics*, 20(9):1453–1454.

C.D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proc. of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12.

Denis Paperno and Marco Baroni. 2016. When the Whole is Less than the Sum of its Parts: How Composition Affects PMI Values in Distributional Semantic Vectors. *Computational Linguistics*, 42(2):345–350.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 28–30, Marrakech, Morocco. European Language Resources Association (ELRA).

Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA.