Annotation and Extraction of Relations from Italian Medical Records

Giuseppe Attardi, Vittoria Cozza, Daniele Sartiano

Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
I-56127 Pisa, Italy
{attardi,cozza,sartiano}@di.unipi.it

Abstract. We address the problem of extracting knowledge from large scale clinical records written in Italian by physicians. We perform recognition of relevant entities such as symptoms, diseases, treatments, measurements, drugs and so forth, and then we determine their semantic relations. We developed suitable training corpora in order to apply machine learning techniques to this task. We report on experiments performed on medical data provided in the context of a regional research project on technologies for health care.

Keywords. Information extraction; Natural language processing; Semantic analysis; Medical ontologies.

1 Introduction

Clinical records are a vast potential source of information for healthcare systems, whose analysis may produce valuable data for building systems to support diagnosis, to predict drug risks, to estimate the effectiveness of treatments. An electronic medical record (EMR) provides detailed information on patient history, laboratory tests and findings of a patient consultation, often expressed in a narrative style. Such records abound in mentions of clinical conditions, anatomical sites, medications, and procedures. Many different surface forms are used to represent the same concept and the mentions are interleaved with modifiers, e.g., adjectives, verb or adverbs, or are abbreviated. Sophisticated techniques of language analysis are required for recognizing these mentions. The extracted data, to be amenable to further analysis and data mining, has to be normalized, for example by mapping or linking entities to their definitions in a widely used standard taxonomy, e.g., Snomed-CT[1], ICD9[2] or, more generally, to their key terminology from UMLS metathesaurus [9]. Finally certain information must be contextualized, for example to a temporal duration or within a

[1] Snomed CT: http://www.ihtsdo.org/snomed-ct/
[2] Classification of Diseases, Functioning, and Disability: http://www.cdc.gov/nchs/icd.htm

statement assessing explicitly their validity. All these issues pose relevant challenges for the current techniques of machine reading.

In this paper we report on our approach for dealing with the following tasks: medical entities recognition, mapping entities to a thesaurus, extracting measurements and their associated entity and identifying whether the context of an expression is positive or negative. We exploit both supervised machine-learning techniques, which require annotated training corpora, and unsupervised deep learning techniques, in order to leverage unlabeled data.

For English several medical corpora with syntactic and semantic information are available, manually annotated as the Shared Annotated Resources [16], while in Italian there is a lack of such resources.

Within the RIS project [15] we had access to a relevant number of medical records from the Italian healthcare system that we used for building in a semiautomatic way a training corpus annotated with medical entities [8] and temporal expression [7]. In this paper we extend the corpus annotations, including expressions denoting physiological measures and the entity to which they refer. We also developed a corpus containing annotations about entities present within a negative context. These corpora have been used for training several classifiers, identifying medical entities in clinical records, linking entities to UMLS CUIs (Concept Unique Identifiers), associating them to measurements and identifying negative or speculative expression.

2 Related Work

Named entity recognition, normalization and linking to thesauri are essential preliminary tasks in biomedical record analysis.

Approaches to the extraction of clinical concepts range from early symbolic NLP systems, strongly dependent on domain knowledge, to machine learning systems driven by the increasing availability of annotated clinical corpora.

The 2014 SemEval Task 7 presented a challenge on the analysis of clinical records from the ShARe resource [16]. The task focused on the recognition and normalization of named entity mentions, those classified within the semantic group disorders [9] in UMLS. In [13] a survey of all the systems used for the task is available; the state of the art solutions are those using machine learning approaches and the most applied tools are those using Conditional Random Fields (CRF), Support Vector Machines (SSV) and DNorm.

The best results were obtained by Tang et al. [18] using an ensemble of learning based systems, i.e., a CRF NER and a Structural Support Vector Machine (SSVM) for disorder entity recognition; they developed a Vector Space Model (VSM) based approach to find the most suitable CUI for a given disorder entity: disorder entity was used as query and all the UMLS terms were treated as documents, then they used cosine similarity score to rank the candidate terms. A novelty of their approach was investigating three different types of word representation (WR) features for the NER, including clustering-based representations, distributional representations and word

embedding [1, 10]. They achieved a precision of 83.4%, a recall of 78.6% and F-score of 81.3% [13].

In [12] the authors explore two approaches to medical documents information extraction in Italian: (i) a cascaded, two-stage method based on pipelining two taggers generated via the well-known Linear-Chain Conditional Random Fields (LC-CRFs) learner and (ii) a confidence-weighted ensemble method that combines standard LC-CRFs with the two-stage method above. They experiment on a dataset of 500 radiology reports in Italian annotated with 9 broad topics, by two annotators independently, 190 reports each. They build individual binary classifiers for each tag and evaluate them separately: this assumes independence of tags, which does not hold for all cases dealt in this paper. An average F1 score of 79.3% is obtained by applying the ensemble method to the two test sets annotated by the same annotator (~119 reports).

As for the relation extraction task, the approach presented in this work recalls the approach presented in [14]. The authors proposed a supervised machine learning approach to discover relations among medical problems, treatments and medical tests mentioned in electronic medical records. A rich set of features was developed for the classifier, their experiments showed that lexical and contextual features are very relevant for relation extraction. They validated their techniques in the 2010 i2b2 Challenge and obtained the highest F-score for the relation extraction task of 73.7%.

As for task of detecting negative and speculative information, this is a very common problem for medical report analysis, since these language forms are widely used to express impressions, hypotheses, or explanations of experimental results.

The author in [11] focused on developing a system based on machine-learning techniques that identifies negation and speculation signals and their scope in clinical texts. The proposed system works in two consecutive phases: first, a classifier decides whether each token in a sentence is a negation/speculation signal or not. Then another classifier determines, at sentence level, the tokens affected by the signals previously identified. The system was trained and evaluated on the clinical texts of the BioScope corpus, a freely available resource consisting of medical and biological texts: full-length articles, scientific abstracts, and clinical reports. In the signal detection task, the F-score value was 97.3% in negation and 94.9% in speculation. In the scope-finding task, a token was correctly classified if it had been properly identified as being inside or outside the scope of all the negation signals present in the sentence. They achieved an F-score of 93.2% in negation and 80.9% in speculation.

3 Medical Training Corpus

Our approach to the analysis of clinical records relies on machine leaning techniques which require either annotated corpora for supervised training or a large set of unannotated documents for unsupervised learning.

Since Italian corpora annotated with mentions of medical entities are not easily available, we created a corpus of Italian medical reports (IMR) [8], annotated with mentions of active ingredient, body part, sign or symptom, disease or syndrome, drug and treatment.

As detailed in [8], the distribution of categories in the annotated IMR is listed in Table 1.

Entity Type	Unique entities	Occurrences
Active Ingredient	217	4,115
Body part	282	19,205
Disease or Syndrome	1,382	51,584
Drug	708	20,479
Sign or Symptom	225	6,842
Treatment	419	32,077

Table 1. Medical Entities retrieved into IMR

The IMR also contains mentions of temporal expressions, extracted as in [7].

Unstructured medical texts may also refer to various kinds of physiological measurements. To extract this valuable information we used a NER for measurements. To this aim, the IMR corpus has been annotated with a basic rule-based approach (regular expression).

We started from the list of units in the metric system (see http://en.wikipedia.org/wiki/Metric_system) and filtered a subset of those actually used in the IMR for measuring the following quantities: area, amount of substance, energy, frequency, length, mass, power, pressure, speed, time and volume. To these we added units for: aerobic capacity, concentration, dosage and flow as well as simple numeric quantities and percentages. We applied a regular expression matcher to identify expression consisting of numeric values in combination with these units. The matcher detected 82.240 occurrences of measurements within the IMR distributed among the following measure categories:

Measure	Occurrences in the corpus
Aerobic capacity	400
Amounts of substance	15
Area	303
Concentration	1,182
Dosage	1,890
Energy	79
Frequency	2,283
Flow	1,949
Length	24,609
Mass	17,470
Percentage	14,468
Power	2,248
Pressure	9,817
Quantity	2,656
Speed	8,526
Time	2,593

Table 2. Distribution of measurements annotated in the IMR

Regular expression matching fails short of identifying all possible variants of measurement expressions used by physicians. For example the dosage of a drug or a therapy is written in many variants, like: "`1 cpr/die per 10 giorni`", "`80 ml/ora`", "`0,125 mg/die`", "`5 mg/kg ogni 8 ore per 5-7 giorni`", "`40 mg in 250 cc`", "`1800 Kcal/die`". It is also hard to identify measurements expressed only by numbers without any indication of units (i.e., `classe nyha: III`), by a partitive or when unusual units are used ("`una bustina /die`", "`due fl di Lasix`").

The annotations obtained in this way are to be considered only as a baseline annotated corpus. We are planning to extend the corpus with manual annotations either by experts or by crowd-sourcing as discussed in [8]. As mentioned earlier exploiting a supervised machine learning approach is much more promising than using handcrafted rules. We explain later how we developed a tagger for measurement: while the tagger has been trained on the baseline corpus, it can easily be trained on a corpus annotated with richer or more varies kinds of expressions.

The IMR has been annotated adding in particular a different column for each group of annotations according to the IOB format[3], each additional column being respectively the first one for body part and treatment, the second one for active principles, diseases, drugs and signs, the last one for measurements. In the following example there are three entities with different annotations: "`ecocardiogramma`" as a treatment, "`versamento pericardico`" as a disease or syndrome, "16 mm" as the length.

ID	FORM	A	B	C
1	Inoltre	O	O	O
2	Sia	O	O	O
3	l'	O	O	O
4	ecocardiogramma	B-TREA	O	O
5	Che	O	O	O
6	La	O	O	O
7	TC	O	O	O
8	Cardiac	O	O	O
9	hanno	O	O	O
10	Rilevato	O	O	O
11	Versamento	O	B-DISO	O
12	Pericardico	O	I-DISO	O
13	diffuso	O	O	O
14	,	O	O	O
15	fino	O	O	O
16	a	O	O	O
17	16	O	O	B-LENGTH

[3] IOB annotation format guidelines: http://en.wikipedia.org/wiki/Inside_Outside_Beginning

18	mm	O	O	I-LENGTH
19	.	O	O	O

4 Bio-medical Information Extraction

To deal with the information extraction of clinical records, we performed two step of analysis:

- recognition of bio-medical entity mentions;
- mapping of entities to their unique UMLS CUI (Concept Unique Identifiers), when applicable.

For example, in a sentence containing "ulcere da decubito" we must identify "ulcera da decubito", even if it is expressed in a different number, and then map it to its UMLS CUI, in this case: "C0011127". The UMLS CUI allows obtaining the corresponding ICD9-CM code, in this case "707.0", which is important, since ICD9-CM is the official annotation for diseases and treatments used in Italian healthcare systems.

After mention identification, a further step is to discover semantic relations between entities or the presence of negations.

To identify entities of interest in text we used three classifiers: NER A, for body parts and treatments; NER B for other medical entities; NER C for measurements. NER A and NER B are used on sets of disjoint categories, i.e., each mention belongs to a single category. NER C is applied to the output of the two other classifiers.

The IMR corpus has been annotated with mentions as detailed in the previous section.

5 Experiments

We built three specialized Named Entity recognizers, one for extracting mentions of body parts and treatments, one for extracting other clinical entities, one for recognizing measurements. The first two were built separately since there are occurrences of body parts within diseases or symptoms, e.g., "dolore alla spalla destra" is a symptom and "spalla destra" is a body part.

For the experiments we split the annotated corpus into train, development and test sets, of size 80%, 10% and 10% respectively.

In our previous works [4,6,8] we tested different NE recognizers. In the current experiments we used the Tanl NER [3], a generic, customizable statistical sequence labeler. The tagger implements a Conditional Markov Model and can be configured to use different classification algorithms and to specify templates for extracting features. In our experiments, it has given overall best results in a configuration using a L2-regularized L2-loss support vector classifier.

We experimented with various feature sets, including word shape features, as in [3], dictionary features, prefix and suffix features, bigrams, last words, first words and frequent words, all extracted from the training corpus.

Table 3 reports the results on the test set achieved with the best configuration obtained on the development set.

The accuracy achieved in these tests is to be taken as just indicative, since there is a strong bias due to the fact that the corpus was mostly annotated automatically. In order to be properly representative of its medical content, the corpus will have to be extended with manual annotations of mentions that have escaped the automatic processing.

	NER A (body parts, treatments)	NER B (other mentions)	NER C (measurements)
Accuracy	99.90%	99.67 %	99.81 %
Precision	98.88%	97.53 %	97.66 %
Recall	97.66%	95.85 %	98.17 %
F-measure	98.26 %	96.68 %	97.91 %

Table 3. Results of NER on various types of entities.

6 Relation extraction task

Extracting mentions can be useful for certain statistical analyses of the content of clinical records, for example counting occurrences and computing correlations. However there are aspects of the content that might be missed or interpreted incorrectly. For example certain mentions may appear within a negative (`assenza di febbre`) or speculative context (`probabile trauma`). Accurate analysis of the report requires distinguishing these cases. This analysis requires identifying relations between parts of the text, not just individual components like mentions.

We explored the identification of relations of this kind in two particular cases: negation identification and association of measures to entities. Both of these analysis were based on examining the parse tree of a sentence, which we obtained by using the dependency parser DeSR [4].

The features to be extracted from parse trees in order to perform this analysis should also be learned from a training corpus.

For this reason we manually annotated a small subset of the IMR, about 10%. The corpus is useful for a preliminary analysis and for validating the effectiveness of our approach, but it will have to be extended in the future.

In order to prepare the training corpus for expressing negation, the IMR corpus was extended with a column, according to IOB format, with a negation TAG if the entity is in a negative context. In the following example the entities "`diabete`" and "`ipertensione`" are in a negative context:

ID	FORM	A	B	C	NEGATION
1	Familiarità	O	O	O	O
2	per	O	O	O	O

3	cardiopatia	O	B-DISO	O	O
4	ischemica	O	I-DISO	O	O
5	,	O	O	O	O
6	nega	O	O	O	O
7	diabete	O	B-DISO	O	B-NEGA
8	,	O	O	O	O
9	non	O	O	O	O
10	storia	O	O	O	O
11	di	O	O	O	O
12	ipertensione	O	B-DISO	O	B-NEGA
13	,	O	O	O	O
14	dislipidemia	O	B-DISO	O	O
15	.	O	O	O	O

For representing relations between entities, each annotated entity is assigned a sequence number, uniquely identifying the entity within the sentence. This id is added as an extra attribute to each token, represented as an extra column in the tab separated IOB file format for the NE tagger, '_' means not involved in a relation. In the example below the length measurement is associated to the disease mention "versamento pericardico" refers with:

ID	FORM	B	C	RELATION
…				
11	versamento	B-DISO	O	1
12	pericardico	I-DISO	O	1
13	diffuso	O	O	_
14	,	O	O	_
15	fino	O	O	_
16	a	O	O	_
17	16	O	B-LENGTH	1
18	mm	O	I-LENGTH	1
…				

We trained an extractor on mentions from the output of classifiers NER B and NER C, e.g., associating measurements to medical entities except treatments and body parts. Other cases might be exploited as well, given a suitable training set: for example the outputs of NER A and NER B, to extract relationships between diseases and body parts, NER A with NER C for relationships between body parts and measurements.

6.1 Negative context

For identifying negative contexts in clinical report, we have trained on the above corpus an SVM-based negation tagger that we are currently evaluating.

Given a sentence and a named entity target, the tagger classifies the context of the entity as positive or negative.

The classifier uses as features patterns on dependency parse trees for negative expressions, similar to those in [17] that allow representing the syntactic context. Examples of these patterns are negated verbs (`la patologia non è presente`), negative verbs `(il paziente nega di avere la patologia)`, negative adjective `(Il paziente è privo di patologia)`, negative nouns (`assenza di patologia`).

6.2 Measurement Associations

The measurements extracted in the medical reports are considered as relevant only when it is possible to detect a direct link to the entity they refer to. The task of associating medical entities to measurements is performed by exploiting a binary SVM classifier trained to recognize whether two mentions are related.

The training instances for the pair-wise learner consist of all pairs of mentions within a sentence of either a symptom, disease, active ingredient or drug and measurements (frequency, weight and so forth). A positive instance is created if the terms are associated, negative otherwise.

The classifier was trained using the following features, extracted for each pair as described above:

Distance features
Token distance: quantized distance between the two words;
Tree distance: distance between two words on the parse tree;
NER features
Ner: the entity type of the pair of words
Syntactic features
Pos: the POS of the pair of words

For computing the distance features we preprocessed the corpus by using the DeSR parser [4]. For each pair of words in a parsed sentence that are tagged as mentions, features are extracted and passed to the classifier.

For instance, from the parse tree of the sentence “Ipertrofia totale cuore (500 g) particolarmente evidente”, in Fig. 1, and given that `ipertrofia` is a mention of a disease and 500 g is a mass, for tokens `ipertrofia` and 500 we extract these features:

```
Pos_feat(Ipetrofia, 500) = Sfs-N
Ner_feat(Ipetrofia, 500) = DISOMASS
Sentence_distance(Ipetrofia, 500) = 4
Tree_distance(Ipetrofia, 500) = 2
```

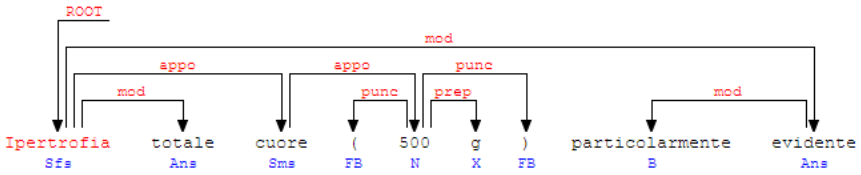

Fig. 1. Sample Parsed Sentence

Sentences are parsed and then for each pair of words that are tagged as mentions, features are extracted and passed to the classifier.

If the classifier assigns a probability greater than a given threshold, the two words are combined into a larger mention. The process is then repeated trying to further extend each relation with additional terms by combining mentions that share a word.

The classifier has been trained with a small corpus, manually annotated. The results can be improved using a richer corpus and trying other algorithms besides SVM.

For example, given the sentence:

```
Ipertrofia totale cuore ( 500 g ) particolarmente evidente a
carico del ventricolo destro ( cuore polmonare , spessore 1 cm
).
```

Applying classifiers A and C we identify the following entities:

- `Ipertrofia` (B-DISO)
- `500` (B-MASS)
- `g` (I-MASS)
- `cuore` (B-DISO)
- `polmonare` (I-DISO)
- `1` (B-LENGTH)
- `cm` (I-LENGTH)

The relation extraction classifier identifies these two relations:

- $\texttt{Ipertrofia}_{\text{DISO}} \leftrightarrow 500\_g_{\text{MASS}}$
- $\texttt{cuore\_polmonare}_{\text{DISO}} \leftrightarrow 1\_cm_{\text{LENGTH}}$

Further examples of retrieved entities from the IMR are:

- $\texttt{rigurgito}_{\text{SIGN}} \leftrightarrow 5\_\%_{\text{PERC}}$
- $\texttt{stenosi}_{\text{DISO}} \leftrightarrow 50\_\%_{\text{PERC}}$
- $\texttt{versamento pericardico}_{\text{DISO}} \leftrightarrow 8\_mm_{\text{LENGTH}}$
- $\texttt{Flumazelin}_{\text{ACTI}} \leftrightarrow 1\_fl_{\text{QUANTITY}}$

7 Applications

Once we are able to extract hidden knowledge in data, several tools can be built to help physicians.

For instance, an interactive tool can be developed for helping physicians when writing clinical records by suggesting a standard code, e.g., those from the ICD9 taxonomy, for the pathologies mentioned in the record.

A visual tool might be provided for correlating the entities on a statistical basis. This tool could be used to graphically visualise the correlations between signs and diseases with different degree of probability, helping doctors in formulating diagnoses as in the example of Fig. 2.

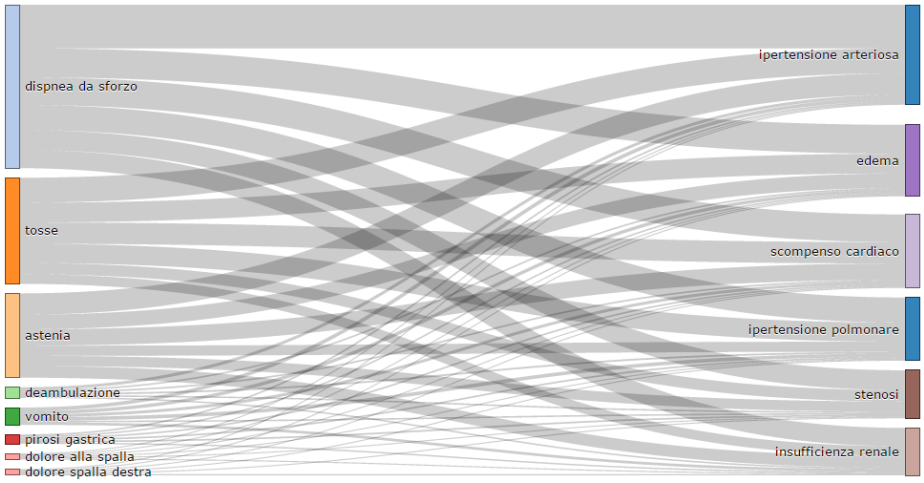

Fig. 2. Entities correlation: symptoms vs diseases.

8 Conclusions

We presented an approach, based on linguistic analysis of biomedical text, for annotating and extracting information from medical records written by Italian clinicians.

Our experiments were carried out within the context of a project on technologies for healthcare, were we had access to a sample of real medical records over a period of 3 year for patients with a pair of major pathologies. The aim of the project was to determine from these data, conditions that might lead to the evolution of these pathologies into a chronic disease. We have extracted a significant amount of data from these records that are being fed to a data mining system for further analysis.

The results obtained are promising, though the corpora we produced need to be further extended.

Acknowledgment. Partial support for this work was provided by project RIS (POR RIS of the Regione Toscana, CUP n° 6408.30122011.026000160).

9 References

1. R. Al-Rfou', B. Perozzi, and S. Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In Proc. of Conference on Computational Natural Language Learning, CoNLL 2013, pp. 183-192, Sofia, Bulgaria.
2. G. Attardi et al., 2009. Tanl (Text Analytics and Natural Language Processing). SemaWiki project: http://medialab.di.unipi.it/wiki/SemaWiki
3. G. Attardi, et al. 2009. The Tanl Named Entity Recognizer at Evalita 2009. In Proc. of Workshop Evalita'09 - Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, ISBN 978-88-903581-1-1.
4. G. Attardi. 2006. Experiments with a Multilanguage Non-Projective Dependency Parser, Proc. of the Tenth Conference on Natural Language Learning, New York, (NY).
5. G. Attardi, A. Buzzelli, D. Sartiano. 2013. Machine Translation for Entity Recognition across Languages in Biomedical Documents. Proc. of CLEF-ER 2013 Workshop, September 23-26, Valencia, Spain.
6. G. Attardi, V. Cozza, D. Sartiano. 2014. UniPi: Recognition of Mentions of Disorders in Clinical Text. Proc. of the 8th International Workshop on Semantic Evaluation. SemEval 2014, pp. 754–760
7. G. Attardi, L. Baronti. 2014. Experiments in Identification of Temporal Expressions in Evalita 2014. Proc. of Evalita 2014.
8. G. Attardi, V. Cozza and D. Sartiano. "Adapting Linguistic Tools for the Analysis of Italian Medical Records". Vol. I: First Italian Conference on Computational Linguistics CLiC-it 2014, 9-10 December 2014, Pisa
9. O. Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, vol. 32, no. supplement 1, D267–D270.
10. R. Collobert et al. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12, pp. 2461–2505.
11. N. P. Cruz Díaz, et al. "A machine-learning approach to negation and speculation detection in clinical texts." Journal of the American society for information science and technology 63.7 (2012): 1398-1410.
12. A. Esuli, D. Marcheggiani, F. Sebastiani, An enhanced CRFs-based system for information extraction from radiology reports, Journal of biomedical informatics 46 (3), 425-435.
13. S. Pradhan, et al. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014), August 2014, Dublin, Ireland, pp. 5462.
14. B. Rink, S. Harabagiu, and K. Roberts. Automatic extraction of relations between medical concepts in clinical texts. Journal of the American Medical Informatics Association, 18(5):594–600, 2011.
15. RIS: Ricerca e innovazione nella sanità. 2014. POR RIS of the Regione Toscana. homepage: http://progetto-ris.it/
16. M. Saeed, C. Lieu, G. Raber, and R.G. Mark. 2002. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol, 29.
17. S. Sohn, S. Wu, C. G. Chute. "Dependency parser-based negation detection in clinical narratives." AMIA Summits on Translational Science Proceedings 2012 (2012): 1.
18. Y. Zhang, J. Wang, B. Tang, Y. Wu, M. Jiang, Y. Chen, H. Xu. "UTH_CCB: A Report for SemEval 2014 – Task 7 Analysis of Clinical Text", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 802–806, Dublin, Ireland, August 23- 24, 2014.