Adapting Linguistic Tools for the Analysis of Italian Medical Records

Giuseppe Attardi
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
attardi@di.unipi.it

Vittoria Cozza
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
cozza@di.unipi.it

Daniele Sartiano
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
sartiano@di.unipi.it

Abstract

English. We address the problem of recognition of medical entities in clinical records written in Italian. We report on experiments performed on medical data in English provided in the shared tasks at CLEF-ER 2013 and SemEval 2014. This allowed us to refine Named Entity recognition techniques to deal with the specifics of medical and clinical language in particular. We present two approaches for transferring the techniques to Italian. One solution relies on the creation of an Italian corpus of annotated clinical records and the other on adapting existing linguistic tools to the medical domain.

Italiano. *Questo lavoro affronta il problema del riconoscimento di entità mediche in referti medici in lingua italiana. Riferiamo su degli esperimenti svolti su testi medici in inglese forniti nei task di CLEF-ER 2013 e SemEval 2014. Questi ci hanno consentito di raffinare le tecniche di Named Entity recognition per trattare le specificità del linguaggio medico e in particolare quello dei referti clinici. Presentiamo due approcci al trasferimento di queste tecniche all'italiano. Una soluzione consiste nella creazione di un corpus di referti medici in italiano annotato con entità mediche e l'altro nell'adattare strumenti tradizionali per l'analisi linguistica al dominio medico.*

1 Introduction

One of the objectives of the *RIS* project (RIS 2014) is to develop tools and techniques to help identifying patients at risk of evolving their disease into a chronic condition. The study relies on a sample of patient data consisting of both medical test reports and clinical records. We are interested in verifying whether text analytics, i.e. information extracted from natural language texts, can supplement or improve information extracted from the more structured data available in the medical test records.

Clinical records are expressed as plain text in natural language and contain mentions of diseases or symptoms affecting a patient, whose accurate identification is crucial for any further text mining process.

Our task in the project is to provide a set of NLP tools for extracting automatically information from medical reports in Italian. We are facing the double challenge of adapting NLP tools to the medical domain and of handling documents in a language (Italian) for which there are few available linguistic resources.

Our approach to information extraction exploits both supervised machine-learning tools, which require annotated training corpora, and unsupervised deep learning techniques, in order to leverage unlabeled data.

For dealing with the lack of annotated Italian resources for the bio-medical domain, we attempted to create a silver corpus with a semi-automatic approach that uses both machine translation and dictionary based techniques. The corpus will be validated through crowdsourcing.

2 Medical Training Corpus

Currently Italian corpora annotated with mentions of medical terms are not easily available. Hence we decided to create a corpus of Italian medical reports (IMR), annotated with medical mentions and to make it available on demand.

The corpus consists of 10,000 sentences, extracted from a collection of 23,695 clinical records of various types, including discharge summaries, diagnoses, and medical test reports.

The annotation process consists in two steps: creating a silver corpus using automated tools

and then turning the corpus into a gold one by manually correcting the annotations.

For building the silver corpus we used:

- a NER trained over a silver English resource translated to Italian;
- a dictionary-based entity recognition approach.

For converting the silver corpus into a gold one, validation by medical experts is required. We organized a crowdsourcing campaign, for which we are recruiting volunteers to whose we will assign micro annotation tasks. Special care will be taken to collected answers reliability.

2.1 Translation based approach

The CLEF-ER 2013 challenge (Rebholz-Schuhmann et al., 2010) aimed at the identification of mentions in bio-medical texts in various languages, starting from an annotated resource in English, and at assigning to them a concept unique identifier (CUI) from the UMLS thesaurus (Bodenreider, 2004). UMLS combines several multilingual medical resources, including Italian terminology from MedDRA Italian (MDRITA15_1) and MESH Italian (MSHITA2013), bridged through their CUI's to their English counterparts.

The organizers provided a silver standard corpus (SSC) in English, consisting of 364,005 sentences extracted from the EMEA corpus, which had been automatically annotated by combining the outputs of several Named Entity taggers (Rebholz-Schuhmann et al., 2010).

In (Attardi et al., 2013) we proposed a solution for annotating Spanish bio-medical texts, starting from the SSC English resource. Our approach combined techniques of machine translation and NER and consists of the following steps:

1. phrase-based statistical machine translation is applied to the SSC in order to obtain a corresponding annotated corpus in the target language. A mapping between mentions in the original and the corresponding ones in the translation is preserved, so that the CUIs from the original can be transferred to the translation. This produces a Bronze Standard Corpus (BSC) in the target language. A dictionary of entities is also created, which associates to each pair (entity text, semantic group) the corresponding CUIs that appeared in the SSC.
2. the BSC is used to train a model for a Named Entity tagger, capable of assigning semantic groups to mentions.
3. the model built at step 2) is used for tagging entities in sentences in the target language.
4. the annotated document is enriched by adding CUIs to each entity, looking up the pair (entity, group) in the dictionary of CUIs, of step 1.

For machine translation we trained Moses (Koehn, 2007) using a biomedical parallel corpus consisting of EMEA, Medline and the Spanish Wikipedia for the language model.

In task A of the challenge, on mention identification, our submission achieved the best score for the categories disease, anatomical part, live being and drugs, with scores ranging between 91.5% and 97.4% (Rebholz-Schuhmann et al., 2013). In task B on CUI identification, the scores were however much lower.

As NE tagger, we used the Tanl NER (Attardi et al., 2009), a generic sequence tagger based on a Maximum Entropy Markov Model, that uses a rich feature set, customizable by providing a feature model. Such kinds of taggers perform quite well on newswire documents, where capitalization features are quite helpful in identifying people, organization and locations. With a proper feature model we were able to achieve satisfactory results also for medical domain.

Adapting the CLEF-ER approach to Italian required repeating the translation step with an en-it parallel corpus, consisting of EMEA and UMLS for the medical domain and (Europarl, 2014; JRC-Acquis, 2014) for more general domains.

A NE tagger for Italian was the trained on the translated silver corpus.

Due to a lack of annotated Italian medical texts, we couldn't perform validation on the resulting tagger. Manual inspection confirms the hypothesis that accuracy is similar to the Spanish version, given that the major difference in the process is the translation corpus and that Spanish and Italian are similar languages.

2.2 Dictionary based approach

Since the terminology for entities in medical records is fairly restricted, another method for identifying mentions in the IMR corpus is to use a dictionary. We produced an Italian medical thesaurus by merging:

- over 70,000 definitions of treatments and diagnosis from the ICD-9-CM terminology;
- about 22,000 definitions from the SnoMed semantic group "Symptoms and Signs, Disease and Anatomical part" in the UMLS;
- over 2,600 active ingredients and drugs from the "Lista dei Farmaci" (AIFA, 2014).

We identified mentions in the IMR corpus using two techniques: *n*-gram based and parser based.

The IMR text is preprocessed first with the Tanl pipeline, performing sentence splitting, tokenization, POS tagging and lemma extraction.

To ease matching, text is normalized by lowercasing each word.

The *n*-gram based technique tries matching each *n*-gram (with *n* between 1 and 10) in the corpus with entries in the thesaurus in two ways: with lemma match and with approximate match. Approximate matching involves removing prepositions, punctuations and articles from both *n*-grams and entries and performing an exact match.

The parse based matching enables to deal also with some kinds of discontiguous mentions, i.e. entity mentions interleaved with modifiers, e.g. adjectives, verb or adverbs. The matching process involves parsing each sentence in the IMR with the DeSR parser (Attardi et al., 2009), selecting noun phrases corresponding to subtrees whose root is a noun and consisting of certain patterns of nouns, adjectives and prepositions, and finally searching these noun phrases in the thesaurus.

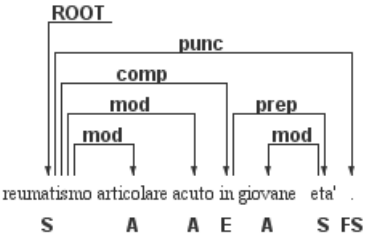

Figure 1: A parsed sentence from the IMR.

Figure 1 shows a sample sentence and its parse tree from which the following noun phrases are identified: `reumatismo acuto`, `reumatismo articolare`, `reumatismo in giovane età`, `giovane età`. Among these, `reumatismo acuto` is recognized as an ICD9 disease.

Overall, we were able to identify over 100,000 entities in the IMR corpus by means of the dictionary approach. The process is expected to guarantee high precision: manual inspection of 4,000 sentences, detected a 96% precision.

Besides recognized entities in the thesaurus, we annotated temporal expressions by a version of HeidelTime (Strotgen and Gertz, 2013) adapted to Italian (Attardi and Baronti, 2014).

3 NE Recognition

We split the analysis of medical records into two steps: recognition of mentions and assignment of unique identifiers (CUI) to those mentions.

The training set consisted of 199 notes with 5,816 annotations, the development set of 99 notes with 5,351 annotations.

For the first step we explored using or adapting traditional NER techniques. We performed experiments in the context of the *SemEval 2014 task 7, Analysis of Clinical Text,* where we could try these techniques using suitable training and test data, even though in English rather than Italian (Attardi et al., 2014). We tested several NER tools: the Tanl NER, the Stanford NER (CRF_NER, 2014) and a Deep Learning NER (NLPNET, 2014), which we developed based on the SENNA architecture (SENNA, 2011). While the first two taggers rely on a rich feature sets and supervised learning, the Deep Learning tagger uses almost no features and relies on word embeddings, learned through an unsupervised process from unannotated texts, along the approach by Collobert et al. 2011.

We created an unannotated corpus (UC) combining 100,000 terms from the English Wikipedia and 30,000 additional terms from a subset of unannotated medical texts from the MIMIC corpus (Moody and Marks, 1996). The word embeddings for the UC are computed by training a deep learning architecture initialized to the values provided by Al-Rfou' et al. (2013) for the English Wikipedia and to random values for the medical terms.

All taggers use dictionary features. We created a dictionary of disease terms (about 22,000 terms from the "Disease or Syndrome" semantic type of UMLS) excluding the most frequent words from Wikipedia.

The Tanl NER could be customized with additional cluster features, extracted from a small window of input tokens. The clusters of UC terms were calculated using the following algorithms:

- DBSCAN (Ester et al., 1996) as implemented in scikit-learn (SCIKIT, 2014). Applied to the word embeddings it produced a set of 572 clusters.
- Continuous Vector Representation of Words (Mikolov et al., 2013), using the word2vec library (WORD2VEC, 2014) with several settings.

We obtained the best accuracy with word2vec using a set of 2,000 clusters.

3.1 Conversion to IOB format

Before applying NE tagging, we had to convert the medical records into the IOB format used by

most NE taggers. The conversion is not straightforward since clinical reports contain discontiguous and overlapping mentions. For example, in:

```
Abdomen is soft, nontender, non-
distended, negative bruits
```

there are two mentions: `Abdomen nontender` and `Abdomen bruits`.

The IOB format does not allow either discontinuity or overlaps. We tested two conversions: one by replicating a sentence, each version having a single mention from a set of overlapping ones. The second approach consisted in using additional tags for disjoint and shared mentions (Tang et al., 2013): DISO for contiguous mentions; DDISO for disjoint entity words that are not shared by multiple mentions; HDISO for the head word that belongs to more than one disjoint mentions.

We tested the accuracy of various NE taggers on the SemEval development set. The results are reported in Table 1. Results marked with *discont* were obtained with the additional tags for discontiguous and overlapping mentions.

NER	Precision	Recall	F- score
Tanl	80.41	65.08	71.94
Tanl+dbscan	**80.43**	64.48	71.58
Tanl+word2vec	79.70	**67.44**	**73.06**
Nlpnet	80.29	62.51	70.29
Stanford	80.30	64.89	71.78
CRFsuite	79.69	61.97	69.72
Tanl discont	78.57	65.35	71.35
Nlpnet discont	77.37	63.76	69.61
Stanford discont	80.21	62.79	70.44

Table 1: Accuracy on the development set.

3.2 Semeval 2014 NER for clinical text

The task 7 of SemEval 2014 allowed us to test NE tagging techniques on medical records and to adapt them to the task. Peculiarly, only one class of entities, namely diseases, is present in the corpus.

We dealt with overlapping mentions by converting the annotations. Discontiguous mentions were dealt in two steps: the first step identifies contiguous portions of a mention with a traditional sequence labeler; then separate portions of mentions are combined into a full mention with guidance from a Maximum Entropy classifier (Berger et al., 1996), trained to recognize which pairs belong to the same mention. The training set consists of all pairs of terms within a document annotated as disorders. A positive instance is created if the terms belong to the same mention, a negative one otherwise.

The classifier was trained using a binned distance feature and dictionary features, extracted for each pair of words in the training set.

For mapping entities to CUIs we applied fuzzy matching (Fraser, 2011) between the extracted mentions and the textual description of entities present in a set of UMLS disorders. The CUI from the match with highest score is chosen.

Our submission reached a comparable accuracy to the best ones based on a single system approach (Pradhan et al., 2014), with an F-score of 0.65 for Task A and 0.83 for Task A relaxed. For Task B and Task B relaxed the accuracies were 0.46 and 0.70 respectively. Better results were achieved by submissions that used an ensemble of taggers.

We also attempted combinations of the outputs from the Tanl NER (with word2vec cluster features), Nlpnet NER and Stanford NER in several ways. The best results were obtained by a simple one voting approach, taking the union of all annotations. The results of the evaluation, for both the multiple copies and discount annotation style, are shown below:

NER	Precision	Recall	F- score
Agreement multiple	73.96	**73.68**	**73.82**
Agreement discont	**81.69**	65.85	72.92

Table 2: Accuracy of NER system combination.

4 Conclusions

We presented a series of experiments on biomedical texts from both medical literature and clinical records, in multiple languages, that helped us to refine the techniques of NE recognition and to adapt them to Italian. We explored supervised techniques as well as unsupervised ones, in the form of word embeddings or word clusters. We also developed a Deep Learning NE tagger that exploits embeddings. The best results were achieved by using a MEMM sequence labeler using clusters as features improved in an ensemble combination with other NE taggers.

As an further contribution of our work, we produced, by exploiting semi-automated techniques, an Italian corpus of medical records, annotated with mentions of medical terms.

Acknowledgements

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

References

AIFA open data. 2014. Retrieved from: http://www.agenziafarmaco.gov.it/it/content/dati-sulle-liste-dei-farmaci-open-data

Rami Al-Rfou', Bryan Perozzi, and Steven 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.

Giuseppe Attardi et al., 2009. Tanl (Text Analytics and Natural Language Processing). SemaWiki project: http://medialab.di.unipi.it/wiki/SemaWiki

Giuseppe 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.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi and Joseph Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In *Proc. of Workshop Evalita'09* - Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, ISBN 978-88-903581-1-1.

Giuseppe Attardi, Andrea. Buzzelli, Daniele Sartiano. 2013. Machine Translation for Entity Recognition across Languages in Biomedical Documents. *Proc. of CLEF-ER 2013 Workshop*, September 23-26, Valencia, Spain.

Giuseppe Attardi, Vittoria Cozza, Daniele 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

Giuseppe Attardi, Luca Baronti. 2014. Experiments in Identification of Temporal Expressions in Evalita 2014. *Proc. of Evalita 2014*.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.* 22, 1 (March 1996), 39-71.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, vol. 32, no. supplement 1, D267–D270.

Ronan Collobert et al. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, pp. 2461–2505.

CRF-NER. 2014. Retrieved from: http://nlp.stanford.edu/software/CRF-NER.shtml

EUROPARL. European Parliament Proceedings Parallel Corpus 1996-2011. 2014. http://www.statmt.org/europarl/

Jenny Rose Finkel, Trond Grenager and Christopher D. Manning 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proc. of the 43nd Annual Meeting of the ACL*, 2005, pp. 363–370.

Neil Fraser. 2011. Diff, Match and Patch libraries for Plain Text.

JRC-Acquis Multilingual Parallel Corpus, Version 2.2. 2014. Retrieved from: http://optima.jrc.it/Acquis/index_2.2.html

Philipp Koehn, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the ACL* on Interactive Poster and Demonstration Sessions. ACL.

MDRITA15_1. 2012. Medical Dictionary for Regulatory Activities Terminology (MedDRA) Version 15.1, Italian Edition; MedDRA MSSO; September, 2012.

MSHITA2013. 2013. Italian translation of Medical Subject Headings. Istituto Superiore di Sanità, Settore Documentazione. Roma, Italy.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. of Workshop at ICLR*.

George B. Moody and Roger G. Mark. 1996. A Database to Support Development and Evaluation of Intelligent Intensive Care Monitoring. *Computers in Cardiology* 23:657–660.

NLPNET. 2014. Retrieved from https://github.com/attardi/nlpnet/

Sameer 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.

Dietrich Rebholz-Schuhmann et al. 2010. The CALBC Silver Standard Corpus for Biomedical Named Entities - A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers. In *Proc. of the Seventh International Conference on Language Resources and Evaluation* (LREC'10). Valletta, Malta.

Dietrich Rebholz-Schuhmann, et al. 2013. Entity Recognition in Parallel Multi-lingual Biomedical Corpora: The CLEF-ER Laboratory Overview. *Lecture Notes in Computer Science*, Vol. 8138, 353-367

RIS: Ricerca e innovazione nella sanità. 2014. POR RIS of the Regione Toscana. homepage: http://progetto-ris.it/

SCIKIT. 2014 Retrieved from http://scikit-learn.org/

SENNA. Semantic/syntactic Extraction using a Neural Network Architecture. 2011. Retrieved from http://ml.nec-labs.com/senna/

Jannik Strotgen and Michael Gertz. Multilingual and Cross-domain Temporal Tagging. 2013. Language Resources and Evaluation, June 2013, Volume 47, Issue 2, pp 269-298.

Buzhou Tang et al. 2013. Recognizing and Encoding Discorder Concepts in Clinical Text using Machine Learning and Vector Space Model. *Workshop of ShARe/CLEF eHealth Evaluation Lab 2013*.

Buzhou Tang, et al. 2014. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *BioMed Research International*, Volume 2014, Article ID 240403.

Jorg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov et al., eds.: *Recent Advances in Natural Language Processing*. Volume V. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, pp. 237–248.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL'03 Shared Task: Language-Independent Named Entity Recognition. In: *Proc. of CoNLL '03*, Edmonton, Canada, 142–147.

WORD2VEC. 2014 Retrieved from http://code.google.com/p/word2vec/