UniPi: Recognition of Mentions of Disorders in Clinical Text

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

The paper describes our experiments addressing the SemEval 2014 task on the Analysis of Clinical text. Our approach consists in extending the techniques of NE recognition, based on sequence labelling, to address the special issues of this task, i.e. the presence of overlapping and discontiguous mentions and the requirement to map the mentions to unique identifiers. We explored using supervised methods in combination with word embeddings generated from unannotated data.

1 Introduction

Clinical records provide detailed information on examination and findings of a patient consultation expressed in a narrative style. Such records abound in mentions of clinical conditions, anatomical sites, medications, and procedures, whose accurate identification is crucial for any further activity of text mining. 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 involving implicit terms.

For example, in

```
Abdomen is soft, nontender,
nondistended, negative bruits
```

the mention occurrences are "`Abdomen nontender`" and "`Abdomen bruits`", which refer to the disorders: "`nontender abdomen`" and "`abdomininal bruit`", with only the second having a corresponding UMLS Concept Unique Identifier (CUI). In this case the two mentions overlap and both are interleaved with other terms, not part of the mentions.

Secondly, mentions can be nested, as in this example:

```
left pleural and parenchymal
calcifications
```

where the mention `calcifications` is nested within `pleural calcifications`.

Mentions of this kind are a considerable departure from those dealt in typical Named Entity recognition, which are contiguous and non-overlapping, and therefore they represents a new challenge for text analysis.

The analysis of clinical records poses additional difficulties with respect to other biomedical NER tasks, which use corpora from the medical literature. Clinical records are entered by medical personnel on the fly and so they contain misspellings and inconsistent use of capitalization.

The task 7 at SemEval 2014, Analysis of Clinical Text, addresses the problem of recognition of mentions of disorders and is divided in two parts:

A. *recognition of mentions* of bio-medical concepts that belong to the UMLS semantic group *disorders*;
B. *mapping* of each *disorder* mention to a unique UMLS CUI (Concept Unique Identifiers).

The challenge organizers provided the following resources:

- A training corpus of clinical notes from MIMIC II database manually annotated for disorder mentions and normalized to an UMLS CUI, consisting of 9432 sentences, with 5816 annotations.
- A collection of unannotated notes, consisting of 1,611,080 sentences.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 754–760,
Dublin, Ireland, August 23-24, 2014.

We also had access to the UMLS ontology (Bodenreider, 2004).

Our approach to portion A of the task was to adapt a sequence labeller, which provides good accuracy in Named Entity recognition in the newswire domain, to handle the peculiarities of the clinical domain.

We performed mention recognition in two steps:

1. identifying contiguous portions of a mention;
2. combining separated portions of mentions into a full mention.

In order to use a traditional sequence tagger for the first step, we had to convert the input data into a suitable format, in particular, we dealt with nested mentions by transforming them into non-overlapping sequences, through replication.

For recombining discontiguous mentions, we employed a classifier, trained to recognize whether pairs of mentions belong to the same entity. The classifier was trained using also features extracted from the dependency tree of a sentence, in particular the distance of terms along the tree path. Terms related by a dependency have distance 1 and terms having a common head have distance 2. By limiting the pairs[1] to be considered for combination to those within distance 3, we both ensure that only plausible combinations are performed and reduce the cost of the algorithm.

For dealing with portion B of the task, we apply fuzzy matching (Fraiser, 2011) between the extracted mentions and the textual description of entities present in selected sections of UMLS disorders. The CUI from the match with highest score is chosen.

In the following sections, we describe how we carried out the experiments, starting with the preprocessing of the data, then with the training of several versions of NE recognizer, the training of the classifier for mention combination. We then report on the results and discuss some error analysis on the results.

2 Preprocessing of the annotated data

The training data was pre-processed, in order to obtain corpora in a suitable format for:

1. training a sequence tagger
2. training the classifier for mention combination.

Annotations in the training data adopt a pipe-delimited stand-off character-offset format. The example in the introduction has these annotations:

```
00098-016139-
DISCHARGE_SUMMARY.txt || Dis-
ease_Disorder || C0221755 ||
1141 || 1148 || 1192 || 1198
00098-016139-
DISCHARGE_SUMMARY.txt || Dis-
ease_Disorder || CUI-less ||
1141 || 1148 || 1158 || 1167
```

The first annotation marks `Disease_Disorder` as annotation type, `C0221755` as CUI, while the remaining pairs of numbers represent character offsets within the original text that correspond to spans of texts containing the mention, i.e. `Abdomen nondistended`. The second annotation is similar and refers to `Abdomen bruits`.

In order to prepare the training corpus for a NE tagger, the data had to be transformed and converted into IOB[2] notation. However a standard IOB notation does not convey information about overlapping or discontiguous mentions.

In order to deal with overlapping mentions, as is the case for word "`Abdomen`" in our earlier example, multiple copies of the sentence are produced, each one annotated with disjoint mentions. If two mentions overlap, two versions are generated, one annotated with just the first mention and one with the second. If several overlapping mentions are present in a sentence, copies are generated for all possible combinations of non-overlapping mentions.

For dealing with discontiguous mentions, each annotated entity is assigned an id, uniquely identifying the mention 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.

We processed with the Tanl pipeline (Attardi et al., 2009; Attardi et al., 2010). We first extracted the text from the training corpus in XML format and added the mentions annotations as tags enclosing them, with spans and mentions id as attributes. We then applied sentence splitting, tokenization, PoS tagging and dependency parsing using DeSR (Attardi, 2006).

The tags were converted to IOB format.

Here are two sample tokens in the resulting annotation, with attributes id, form, pos, head, deprel, entity, entity id:

[1] Not implemented in the submitted runs.

[2] http://en.wikipedia.org/wiki/Inside_Outside_Beginning

```
1 Abdomen NNP 2 SBJ B-DISO 1
…
5 nontender NN 10 NMOD B-DISO 1
```

3 Named Entity Tagging

The core of our approach relies on an initial stage of Named Entity recognition. We performed several experiments, using different NE taggers in different configurations and using both features from the training corpus and features obtained from the unannotated data.

3.1 Tanl NER

We performed several experiments using the Tanl NE Tagger (Attardi et al., 2009), a generic, customizable statistical sequence labeller, suitable for many tasks of sequence labelling, such as POS tagging or Named Entity Recognition.

The tagger implements a Conditional Markov Model and can be configured to use different classification algorithms and to specify feature templates for extracting features. In our experiments we used a linear SVM classification algorithm.

We experimented with several configurations, all including a set of word *shape* features, as in (Attardi et al., 2009): (1) the previous word is capitalized; (2) the following word is capitalized; (3) the current word is in upper case; (4) the current word is in mixed case; (5) the current word is a single uppercase character; (6) the current word is a uppercase character and a dot; (7) the current word contains digits; (8) the current word is two digits; (9) the current word is four digits; (10) the current word is made of digits and "/"; (11) the current word contains "$"; (12) the current word contains "%"; (13) the current word contains an apostrophe; (14) the current word is made of digits and dots.

A number of dictionary features were also used, including prefix and suffix dictionaries, bigrams, last words, first word and frequent words, all extracted from the training corpus.

Additionally, a dictionary of disease terms was used, consisting of about 22,000 terms extracted from the preferred terms for CUIs belonging to the UMLS semantic type "Disease or Syndrome".

The first character of the POS tag was also used as feature, extracted from a window of tokens before and after the current token.

Finally *attribute features* are extracted from attributes (Form, PoS, Lemma, NE, Disease) of surrounding tokens, denoted by their relative position to the current token. The best combination of Attribute features obtained with runs on the development set was the following:

Feature	Tokens
POS[0]	w_{i-2} w_{i-1} w_i w_{i+1}
DISEASE	w_i w_{i+1} w_{i+2}

Table 1. *Attribute features* used in the runs.

3.2 Word Embeddings

We explored ways to use the unannotated data in NE recognition by exploiting word embeddings (Collobert et al, 2011). In a paper published after our submission, Tang et al. (2014) show that word embeddings are beneficial to Biomedical NER.

We used the word embeddings for 100,000 terms created through deep learning on the English Wikipedia by Al-Rfou et al. (2013). We then built, with the same procedure, embedding for terms from the supplied unlabelled data. The corpus was split, tokenized and normalized and a vocabulary was created with the most frequent words not already present among the Wikipedia word embeddings. Four versions of the embeddings were created, varying the size of the vocabulary and the size of the context window, as described in Table 1.

	Run1	Run2	Run3	Run4
Vocabulary size	50,000	50,000	30,000	30,000
Context	5	2	5	2
Hidden Layers	32	32	32	32
Learning Rate	0.1	0.1	0.1	0.1
Embedding size	64	64	64	64

Table 2. Word Embedding Parameters.

We developed and trained a Deep Learning NE tagger (nlpnet, 2014) based on the SENNA architecture (SENNA, 2011) using these word embeddings.

As an alternative to using the embeddings directly as features, we created clusters of word embeddings using the Dbscan algorithm (Ester et al., 1996) implemented in the sklearn library. We carried out several experiments, varying the parameters of the algorithm. The configuration that produced the largest number of clusters had 572 clusters. The clusters turned out not to be much significant, since a single cluster had about 29,000 words, another had 5,000 words, and the others had few, unusual words.

We added the clusters as a dictionary feature to our NE tagger. Unfortunately, most of the

terms fell within 4 clusters, so the feature turned out to be little discriminative.

3.3 Stanford NER

We performed experiments also with a tagger based on a different statistical approach: the Stanford Named Entity Recognizer. This tagger is based on the Conditional Random Fields (CRF) statistical model and uses Gibbs sampling instead of other dynamic programming techniques for inference on sequence models (Finkel et al., 2005). This tagger normally works well enough using just the form of tokens as feature and we applied it so.

3.4 NER accuracy

We report the accuracy of the various NE taggers we tested on the development set, using the scorer from the CoNLL Shared Task 2003 (Tjong Kim Sang and De Meulder, 2003).

We include here also the results with CRFsuite, the CRF tagger used in (Tang et al., 2014).

NER	Precision	Recall	F- score
Tanl	80.41	65.08	71.94
Tanl+clusters	80.43	64.48	71.58
nlpnet	80.29	62.51	70.29
Stanford	80.30	64.89	71.78
CRFsuite	79.69	61.97	69.72

Table 3. Accuracy of various NE taggers on the development set.

Based on these results we chose the Tanl tagger and the Stanford NER for our submitted runs.

All these taggers are known to be capable of achieving state of the art performance or close to it (89.57 F1) in the CoNLL 2003 shared task on the WSJ Penn Treebank.

The accuracy on the current benchmark is much lower, despite the fact that there is only one category and the terminology for disorders is drawn from a restricted vocabulary.

It has been noted by Dingare et al. (2005) that NER over biomedical texts achieves lower accuracy compared to other domains, quite within the range of the above results. Indeed, compared with the newswire domain or other domains, the entities in the biomedical domain tend to be more complex, without the distinctive shape features of the newswire categories.

4 Discontiguous mentions

Discontiguous mention detection can be formulated as a problem of deciding whether two contiguous mentions belong to the same mention. As such, it can be cast into a classification problem. A similar approach was used successfully for the coreference resolution task at SemEval 2010 (Attardi, Dei Rossi et al., 2010)

4.1 Mentions detection

We trained a Maximum Entropy classifier (Ratnaparkhi, 1996) to recognize whether two terms belong to the same mention.

The training instances for the pair-wise learner consist of each pair of terms within a sentence annotated as disorders. A positive instance is created if the terms belong to the same mention, negative otherwise.

The classifier was trained using the following features, extracted for each pair of words for diseases.

Distance features

- *Token distance:* quantized distance between the two words;
- *Ancestor distance:* quantized distance between the words in the parse tree if one is the ancestor of the other

Syntax features

- *Head:* whether the two words have the same head;
- *DepPath:* concatenation of the dependency relations of the two words to their common parent

Dictionary features

- UMLS: whether the two words are both present in an UMLS definition

The last feature is motivated by the fact that, according to the task description, most of the disorder mentions correspond to diseases in the SNOMED terminology.

4.2 Merging of mentions

The mentions detected in the first phase are merged using the following process. Sentence are parsed and then for each pair of words that are tagged as disorder, 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 mention

with additional terms by combining mentions that share a word.

5 Mapping entities to CUIs

Task B requires mapping each recognized entity to a concept in the SNOMED-CT terminology, assigning to it a unique UMLS CUI, if possible, or else marking it as `CUI-less`. The CUIs are limited to those corresponding to SNOMED codes and belonging to the following UMLS semantic types: “Acquired Abnormality" or “Congenital Abnormality", “Injury or Poisoning", "Pathologic Function", "Disease or Syndrome", "Mental or Behavioral Dysfunction", "Cell or Molecular Dysfunction", "Experimental Model of Disease" or "Anatomical Abnormality", “Neoplastic Process" or "Sign or Symptom".

In order to speed up search, we created two indices: an inverted index from words in the definition of a CUI to the corresponding CUI and a forward index from a CUI to its definition.

For assigning a CUI to a mention, we search in the dictionary of CUI preferred terms, first for an exact match, then for a normalized mention and finally for a fuzzy match (Fraiser, 2011). Normalization entails dropping punctuation and stop words. Fuzzy matching is sometimes too liberal, for example it matches “`chronic obstructive pulmonary`” with “`chronic obstructive lung disease`”; so we also put a ceiling on the edit distance between the phrases.

The effectiveness of the process is summarized in these results on the development set:

Exact matches	Normalized matches	Fuzzy matches	No matches
1352	868	304	5488

Table 4. CUI identifications on the devel set.

6 Experiments

The training corpus for the submission consisted of the merge of the train and development sets.

We submitted three runs, using different or differently configured NE tagger.

Two runs were submitted using the Tanl tagger using the features listed in Table 5, where DISEASE and CLUSTER meaning is explained earlier.

Feature	UniPI_run0	UniPI_ run1
POS[0]	w_{i-2} w_{i-1} w_i w_{i+1}	w_{i-2} w_{i-1} w_i w_{i+1}
CLUSTER	w_i w_{i+1}	w_i w_{i+1}
DISEASE	w_i w_{i+1} w_{i+2}	

Table 5. *Attribute features* used in the runs.

Since the clustering produced few large clusters, the inclusion of this feature did not affect substantially the results.

A third run (UniPi_run_2) was performed using the Stanford NER with default settings.

7 Results

The results obtained in the three submitted runs, are summarized in Table 6, in terms of accuracy, precision, recall and F-score. For comparison, also the results obtained by the best performing systems are included.

Run	Precision	Recall	F- score
Task A			
Unipi_run0	0.539	0.684	0.602
Unipi_run1	0.659	0.612	0.635
Unipi_run2	0.712	0.601	0.652
SemEval best	**0.843**	**0.786**	**0813**
Task A relaxed			
Unipi_run0	0.778	0.885	0.828
Unipi_run1	0.902	0.775	0.834
Unipi_run2	0.897	0.766	0.826
SemEval best	**0.936**	**0.866**	**0.900**

Table 6. UniPI Task A results, compared to the best submission.

Run	Accuracy
Task B	
Unipi_run0	0.467
Unipi_run1	0.428
Unipi_run2	0.417
SemEval best	**0.741**
Task B relaxed	
Unipi_run0	0.683
Unipi_run1	0.699
Unipi_run2	0.693
SemEval best	**0.873**

Table 7. UniPI Task B results, compared to the best submission.

8 Error analysis

Since the core step of our approach is the NE recognition, we tried to analyze possible causes of its errors.

Some errors might be due to mistakes by the POS tagger. For example, often some words occur in full upper case, leading to classify adjectives like `ABDOMINAL` as `NNP` instead of `JJ`. Training our POS tagger on the GENIA corpus or using the GENIA POS tagger might have helped a little. Spelling errors like `abdominla`

instead of `abdominal` could also have been corrected.

Another choice that might have affected the NER accuracy was our decision to duplicate the sentences in order to remove mention overlaps. An alternative solution might have been to use two categories in the IOB annotation: one category for full contiguous disorder mentions and another for partial disorder mentions. This might have reduced the confusion in the tagger, since isolated words like `abdomen` get tagged as disorder, having been so annotated in the training set. Distinguishing the two cases, `abdomen` would become a disorder mention in the step of mention merging. Counting the errors in the development set we found that 939 out of the 1757 errors were indeed individual words incorrectly identified as disorders.

8.1 After submission experiments

After the submission, we changed the algorithm for merging mentions, in order to avoid nested spans, retaining only the larger one. Tests on the development set show that this change leads to a small improvement in the strict evaluation:

Run	Precision	Recall	F- score
Task A			
devel_run1	0.596	**0.653**	0.624
devel run1_after	**0.668**	0.637	**0.652**
Task A relaxed			
devel_run1	**0.865**	**0.850**	**0.858**
devel run1_after	0.864	0.831	0.847

Table 8. UniPI Task A post submission results.

9 Conclusions

We reported our participation to SemEval 2014 on the Analysis of Clinical Text. Our approach is based on using a NER, for identifying contiguous mentions and on a Maximum Entropy classifier for merging discontiguous ones.

The training data was transformed into a format suitable for a standard NE tagger, that does not accept discontiguous or nested mentions. Our measurements on the development set showed that different NE tagger reach a similar accuracy.

We explored using word embeddings as features, generated from the unsupervised data provided, but they did not improve the accuracy of the NE tagger.

Acknowledgements

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

References

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of Conference on Computational Natural Language Learning*, CoNLL '13, pages 183-192, Sofia, Bulgaria.

Giuseppe Attardi. 2006. Experiments with a Multilanguage Non-Projective Dependency Parser. In *Proceedings of the Tenth Conference on Natural Language Learning,* CoNLL '06, pages 166-170, New York, NY.

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

Giuseppe Attardi, Stefano Dei Rossi, Felice Dell'Orletta and Eva Maria Vecchi. 2009. The Tanl Named Entity Recognizer at Evalita 2009. In *Proceedings 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 *Proceedings of Workshop Evalita'09 - Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, ISBN 978-88-903581-1-1.

Giuseppe Attardi, Stefano Dei Rossi and Maria Simi. 2010. The Tanl Pipeline. In *Proceedings of LREC Workshop on Web Services and Processing Pipelines in HLT, WSPP*, La Valletta, Malta, pages 14-21

Giuseppe Attardi, Stefano Dei Rossi and Maria Simi. 2010. TANL-1: Coreference Resolution by Parse Analysis and Similarity Clustering. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval 2010*, Uppsala, Sweden, pages 108-111

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

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

Shipra Dingare, Malvina Nissim, Jenny Finkel, Christopher Manning and Claire Grover. 2005. A System for Identifying Named Entities in Biomedical Text: how Results From two Evaluations Reflect on Both the System and the Evaluations. *Comp Funct Genomics*. Feb-Mar; 6(1-2): pages 77–85.

Martin Ester, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mi-ing,* KDD 96, pages 226–231.

Jenny Rose Finkel, Trond Grenager and Christopher Manning 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistic*s, 2005, pages 363–370.

Neil Fraser. 2011. Diff, Match and Patch libraries for Plain Text. (Based on Myer's diff algorithm).

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference, EMNLP '96*, pages 17-18.

Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. 2014. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *BioMed Research International*, Volume 2014, Article ID 240403.

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

SENNA. 2011. http://ml.nec-labs.com/senna/

nlpnet. 2014. https://github.com/attardi/nlpnet