UCM at CLEF eHealth 2013 Shared Task1

Lucía Hervás, Víctor Martínez, Irene Sánchez, Alberto Díaz

NIL Group
Universidad Complutense de Madrid
C/Profesor García Santesmases, Madrid, 28040, Spain,
lhervasmartin@gmail.com
victormartinezsimon@gmail.com
irene.sanchzmartinz@gmail.com
albertodiaz@fdi.ucm.es

Abstract. We are developing a system that analyzes medical reports and extracts a SNOMED-CT based concept representation. The more interesting characteristic of our system is not only that it can detect the concepts. It also takes into account if they appear in an affirmative, negative or speculative context. The system also separates the concept representation according to the structure of the document. Our system takes these steps: automatic orthographic correction, acronyms and abbreviation detection, negation and speculation phrase detection and medical concepts detection. For participating in Task 1 we have adapted our system in order to obtain the mentions that belongs to the Disorder semantic group defined in the guidelines. The approach is based on using MetaMap to detect the concepts and the spans. Our aim was to identify which was the best way to use MetaMap in our system to solve the Task 1.

Keywords: Natural Language Processing, medical report, concept detection, Metamap, UMLS

1 Introduction

The goal of Task 1 is to analyze clinical text documents and find mentions of disorders. There are two subtasks: (1a) discovering the mention boundaries and (1b) mapping each mention to a UMLS CUI. Normalization/mapping is limited to UMLS CUIs of SNOMED codes. Participants are free to use any UMLS resources [2].

For participating in this Task we have adapted a system that analyzes medical reports in order to obtain the mentions that belongs to the Disorder semantic group defined in the guidelines. The approach is based on using MetaMap to detect the concepts and the spans.

Our system extracts a SNOMED-CT based concept representation from a medical report. Before the analysis we have other phases: a language corrector and an acronyms expander. The more interesting characteristic of our system is that not only detect the concepts, it also take into account if they appear in

an affirmative, negative or speculative context. The system also separates the concept representation according to the structure of the document, that is, there is a different representation for each section of the document.

During our research, we discover the ShARe/CLEF eHealth 2013 Shared Tasks [2]. As the these tasks were very close of what we are developing, we decide to participate to increase our knowing and the performance of our system. Our aim was to identify which was the best way to use MetaMap in our system to solve the Task1.

We have submitted runs with no external annotations, two for task 1a and two for task 1b. The difference between the runs is only the DB used. We used the 2012AA USAbase strict model for the first run and the 2011AA USAbase strict model for the second run. Our best results for task 1a show 0.504 *F1* score with strict evaluation, and 0.660 *F1* score with relaxed evaluation. Our best results for task 1b show 0.362 *Accuracy* with strict evaluation and 0.870 *Accuracy* with relaxed evaluation.

2 MetaMap

MetaMap maps biomedical text to concepts in the UMLS Metathesaurus. Several types of lexical/syntatic analysis are performed on the input text to perform this mapping: tokenization, part-of-speech tagging, lexical lookup in the SPECIALIST lexicon and shallow parsing. Per each noun phrase obtained is applied the next processes: variant generation, candidate identification, mapping construction and word sense disambiguation. Final scores are computed per each candidate mapping combining different measures [1].

MetaMap has different parameters that influences its performance: data options, output options and processing options. The data options allow to choose the level of filtering and the UMLS data. The default setting is the Strict model, where all types of filtering are applied. The Relaxed model only includes manual and lexical filterings.

3 Processing

We use the default setting of MetaMap to detect the different concepts and to know their CUI. MetaMap retrieves some concepts, so to reduce the noise, we only take the concepts with the greater score. We also use the MedPost/SKR server included in MetaMap to perform word sense disambiguation. Finally, we configure the system to accept only the next semantic types that corresponds to the Disorder semantic group defined in the guidelines.

- Congenital Abnormality
- Acquired Abnormality
- Injury or Poisoning
- Pathologic Function
- Disease or Syndrome

– Mental or Behavioral Dysfunction
– Cell or Molecular Dysfunction
– Experimental Model of Disease
– Anatomical Abnormality
– Neoplastic Process
– Signs and Symptoms

4 Framework evaluation

Participants will be provided training and test datasets. The evaluation for all tasks will be conducted using the withheld test data. Teams are allowed to use any outside resources in their algorithms.

4.1 Evaluation Measures

In subtask 1a, boundary detection of disorders, the evaluation measures are *F1-score*, *Recall* and *Precision*, where a *TP* is considered when the span obtained is the same that the gold standard span, a *FP* when it is a spurious span, and a *FN* when it is a missing span. There are two variants: Strict and Relaxed, depending if the span is identical to the reference standard span, or if the span overlaps the standard span.

In subtask 1b, identify the boundaries of disorders and map them to a SNOMED-CT code, the evaluation measure is *Accuracy*, where *Correct* is considered as the number of disorder named entities with strictly correct span and correctly generated code and *Total* is considered as the number of disorder named entities. There are also two variants: Strict and Relaxed, depending if *Total* is considered as the number of reference standard named entities, or if *Total* is considered as the number of named entities with strictly correct span generated by the system. In the first case, the system is penalized for incorrect code assignment for annotations that were not detected by the system. In the second case, the system is only evaluated on annotations that were detected by the system.

5 Results

We have submitted runs with no external annotations, two for task 1a and two for task 1b. The difference between the runs is only the DB used. We used the 2012AA USAbase strict model for the first run and the 2011AA USAbase strict model for the second run.

Our best results for task 1a shown a 0.504 *F1* score with strict evaluation, and a 0.660 *F1* score with relaxed evaluation. Our best results for task 1b shown a 0.362 *Accuracy* with strict evaluation, and a 0.871 *Accuracy* with relaxed evaluation.

Table 1. Task 1A. No external annotations. Strict

Team,Country	Precision	Recall	F1-Score
UTHealth_CCB.2, UT, USA	0.800	0.706	0.750
NIL-UCM.2, Spain	0.617	0.426	0.504
NIL-UCM.1, Spain	0.621	0.416	0.498
FAYOLA.1, VW, USA	0.024	0.446	0.046

Table 2. Task 1A. No external annotations. Relaxed

Team,Country	Precision	Recall	F1-Score
UTHealth_CCB.2, UT, USA	0.925	0.827	0.873
NIL-UCM.2, Spain	0.809	0.558	0.660
NIL-UCM.1, Spain	0.812	0.543	0.651
FAYOLA.1, VW, USA	0.504	0.043	0.079

Table 3. Task 1B. No external annotations. Strict

Team,Country	Accuracy(sn2012)	Accuracy(sn2011)
NCBI.2, MD, USA	0.589	0.584
NIL-UCM.2, Spain	0.362	0.362
NIL-UCM.1, Spain	0.362	0.362
NCBI.2, MD, USA	0.006	0.006

Table 4. Task 1B. No external annotations. Relaxed

Team,Country	Accuracy(sn2012)	Accuracy(sn2011)
AEHRC.1, QLD, Australia	0.939	0.939
NIL-UCM.1, Spain	0.871	0.870
NIL-UCM.2, Spain	0.850	0.850
UTHealth_CCB.1, UT, USA	0.728	0.772

6 Discussion

The detection of boundaries of disorders offers bad results mainly due to the limit of MetaMap in the discovering of the spans: the best Recall obtained is around 0.42. Of course, the main problem is related with the discontinuous spans that MetaMap is not able to process. With respect to the difference between our systems, the second version offers slightly better results, as expected, because it uses the 2011AA USAbase database. With respect to the type of evaluation, higher scores are obtained with relaxed evaluation, mainly on Precision, but Recall only increase to 0.558.

With respect to the mapping of CUIs, the results are low in the strict evaluation, but high in relaxed evaluation. That is due to the penalization for incorrect code assignment for annotations not detected by the system.

Our results show the baseline that can be obtained using MetaMap with the strict model configuration. Then, we can conclude that MetaMap is not enough to solve this task.

Acknowledgements

We want to acknowledge the support given by the Shared Annotated Resources (ShARe) project, funded by the United States National Institutes of Health with grant number R01GM090187.

References

1. Aronson , A., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances, Journal of the American Medical Informatics Association, vol. 17, no. 3, pp. 229236, 2010.
2. Suominen, H., Salantera, S., Velupillai, S.: Three Shared Tasks on Clinical Natural Language Processing. Proceedings of CLEF 2013. To appear